Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation

Anonymous ACL submission

Abstract

Gender bias is largely recognized as a problematic phenomenon affecting language technologies, with recent studies underscoring that it might surface differently across languages. However, most evaluation practices adopt a word-level focus on a narrow set of occupational nouns under synthetic conditions. Such protocols overlook key features of grammatical gender languages, which are characterized by morphosyntactic chains of gender agreement, marked on a variety of lexical items and parts-of-speech (POS). To overcome this limitation, we enrich the natural, gender-sensitive MuST-SHE corpus with two new annotation layers: POS and agreement chains. On this basis, we conduct multifaceted automatic and manual evaluations for three speech translation models, trained on varying amounts of data and different word segmentation techniques. Our work sheds light on model behaviours, gender bias, and its detection at several levels of granularity for English-French/Italian/Spanish.

1 Introduction

As Matasović (2004) posits: "*Gender is perhaps the only grammatical category that ever evoked passion – and not only among linguists.*" That is because, in the case of human entities, masculine or feminine inflections are assigned semantically, i.e. in relation to the extra-linguistic reality of gender (Ackerman, 2019; Corbett, 1991, 2013). Thus, gendered features interact with the – sociocultural and political – perception and representation of individuals (Gygax et al., 2019), by prompting discussions on the appropriate recognition of gender groups and their linguistic visibility (Stahlberg et al., 2007; Hellinger and Motschenbacher, 2015; Hord, 2016). Such concerns also invested language technologies (Sun et al., 2019; Cao and Daumé III, 2020), where it has been shown that automatic translation systems tend to over-represent masculine forms and amplify stereotypes when translating into grammatical gender languages (Savoldi et al., 2021).

Current evaluation practices for assessing gender bias in both Machine (MT) and Speech Translation (ST) commonly inspect such concerning behaviours on synthetic benchmarks and by focusing on a restricted set of occupational nouns only (Stanovsky et al., 2019; Escudé Font and Costa-jussà, 2019; Renduchintala et al., 2021). Also, when relying on lexically richer natural benchmarks, the designed metrics still work at the word level, treating all gender-marked words indiscriminately (Alhafni et al., 2020; Bentivogli et al., 2020). Accordingly, current test sets and protocols: *i)* do not allow us to inspect if and to what extent different word categories participate in gender bias, *ii)* overlook the underlying morphosyntactic nature of grammatical gender on agreement chains, which cannot be monitored on single isolated words (e.g. *en*: a strange friend; *it*: un<u>a/o</u> stran<u>a/o</u> amic<u>a/o</u>).[1]

We believe that fine-grained evaluations including the analysis of gender agreement across different parts of speech (POS) are relevant not only to gain a deeper understanding of bias in grammatical gender languages, but also to inform mitigating strategies and data curation procedures.

Toward these goals, our contributions are as follows. **(1)** We enrich MuST-SHE (Bentivogli et al., 2020) – the only natural gender-sensitive benchmark available for MT and also ST – with two layers of linguistic information: POS and agreement chains.[2] **(2)** In light of recent studies exploring how model design and overall perfomance interplay with gender bias (Roberts et al., 2020; Gaido et al., 2021), we rely on our manually curated resource to compare three ST models, which are trained on varying amounts of data, and built with differ-

[1] To be grammatically correct, each word in the chain has to be inflected with the same (masculine or feminine) gender form, similar to number agreement (see "*a dogs barks").

[2] The resource will be released upon paper acceptance.

ent segmentation techniques: character and byte-pair-encoding (BPE) (Sennrich et al., 2016). We carry out a multifaceted evaluation that includes automatic and extensive manual analyses on three language pairs (en-es, en-fr, en-it) and we consistently find that: *i)* not all POS are equally impacted by gender bias; *ii)* translating words in agreement does not emerge as a systematic issue; *iii)* ST systems produce a considerable amount of neutral rewordings in lieu of gender-marked expressions, which current binary benchmarks fail to recognize. Finally, in line with concurring studies, we find that *iv)* character-based systems favour morphological and lexical diversity when translating gender phenomena.

2 Background

While research in Natural Language Processing (NLP) initially prioritized narrow technical interventions to address the social impact of language technologies, we are recently attesting a shift toward a more comprehensive understanding of bias (Shah et al., 2020; Blodgett et al., 2020). Along this line, focus has been given to bias analysis in models' innards and ouputs (Vig et al., 2020; Costa-jussà et al., 2020b), and to ascertain the validity of bias measurement practices (Blodgett et al., 2021; Antoniak and Mimno, 2021; Goldfarb-Tarrant et al., 2021). Complementary evidence suggests that – rather than striving for generalizations – gender bias detection ought to incorporate contextual and linguistic specificity (González et al., 2020; Ciora et al., 2021; Matthews et al., 2021; Malik et al., 2021; Kurpicz-Briki and Leoni, 2021), which however receives little attention due to a heavy focus on English NLP. Purported agnostic approaches and evaluations (Bender, 2009) can prevent from drawing reliable conclusions and mitigating recommendations, as attested by monolingual studies on grammatical gender languages (Zhou et al., 2019; Gonen et al., 2019) and in automatic translation scenarios (Vanmassenhove et al., 2018; Moryossef et al., 2019). Unlike English, grammatical gender languages exhibit an elaborate morphological and syntactic system, where gender is overtly marked on numerous POS (e.g., verbs, determiners, nouns), and related words have to agree on the same gender features. Still, current corpora and evaluation practices do not fully foreground systems' behaviour on such grammatical constraints.

WinoMT (Stanovsky et al., 2019) represents the standard corpus to evaluate gender bias in MT within an English-to-grammatical gender language scenario. It has been progressively enriched with new features (Saunders et al., 2020; Kocmi et al., 2020), and adapted for ST (Costa-jussà et al., 2020a). While this resource can be useful to diagnose gender stereotyping at scale, it excludes languages' peculiarities since it is built on the concatenation of two corpora designed for English monolingual tasks – WinoGender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) – which consist of synthetic sentences with the same structure and a pre-selected occupational lexicon.[3] To increase variability, Troles and Schmid (2021) extend WinoBias by accompanying occupations with highly gender-stereotypical verbs and adjectives. Their evaluation though, still only considers the translated professions as to verify if the co-occuring words might skew the models' assumptions. However, gender-marking involves also several other, so far less accounted POS categories, but whether they are just as problematic is not clear yet.

Existing bilingual (Alhafni et al., 2021), and multilingual (Bentivogli et al., 2020) natural benchmarks, instead, are manually curated as to identify a variety of gender phenomena specifically modeled on the accounted languages. As a result, they maximize variability to inspect whether translation models yield feminine under-representation in real-world-like scenarios (Savoldi et al., 2021). However, since this variability is not mapped into fine-grained linguistic information, evaluations on such corpora do not single out which instances may be more responsible for gender bias. Finally, by considering each word in isolation, they neglect the underlying features of gender agreement, which determine the grammatical acceptability of the translation. To the best of our knowledge, only two works have currently interplayed issues of syntactic agreement and gender bias. Renduchintala and Williams (2021) designed a set of English sentences involving a syntactic construction that requires to translate an occupational term according to its unequivocal "gender trigger" (e.g. that *nurse* is a funny man). While they find that MT struggles even in such a simple setting, they only inspect the translation of a single disambiguated word (*nurse*) rather than a whole group of words in agreement. Closer to our intent, Gaido et al. (2020) analyze the

[3] Levy et al. (2021) recently created BUG on natural English data, but still it is limited to the evaluation of occupations.

output of different ST systems and note that their models seem to wrongly pick divergent gender inflections for unrelated words in the same sentence (e.g. *fr*: En tant que chercheuse$_F$, professeur$_M$) but not for dependency-related ones (e.g. *it*: [la classica studentessa asiatica]$_F$). Although limited in scope, their observation is worth being explored systematically. We thus conduct the very first study that intersects POS, agreement, and gender bias.

3 MuST-SHE Enrichment

In light of the above, a fine-grained evaluation of bias focused on POS and gender agreement requires the creation of a new dedicated resource. Rather than building it from scratch, we add two annotation layers to the existing TED-based MuST-SHE benchmark (Bentivogli et al., 2020).[4] Available for en-es/fr/it, it represents the only multilingual MT/ST *GBET*[5] exhibiting a natural variety of gender phenomena. In the reference translations, each gender-marked word – corresponding to a neutral expression in the source – is annotated with its alternative *wrong* gender form (e.g. *en*: ***the** girl **left***; *it*: ***la<il>** ragazza è **andata<andato>** via*). Thus, MuST-SHE allows the identification of numerous and qualitatively different grammatical gender instances under authentic conditions. Furthermore, the target languages covered in MuST-SHE (es, fr, it) are particularly suitable to focus on linguistic specificity. In fact, as Gygax et al. (2019) suggest, accounting for gender in languages with similar typological features allows for proper comparison.

3.1 Phenomena Categorization

Parts-Of-Speech. We annotate each target gender-marked word in MuST-SHE with POS information. As shown in Table 1 (*a-c*), we differentiate among six POS categories:[6] ***i)*** articles, ***ii)*** pronouns, ***iii)*** nouns, ***iv)*** verbs. For adjectives, we further distinguish ***v)*** limiting adjectives with minor semantic import that determine e.g. possession, quantity, space (*my*, *some*, *this*); and ***vi)*** descriptive adjectives that convey attributes and qualities, e.g. *glad*, *exhausted*. This distinction enables to neatly sort our POS categories into the closed class of function words, or into the open one of content words (Schachter and Shopen, 2007). Since words from these two classes differ substantially in terms of variability, frequency, and semantics, we reckon they represent a relevant variable to account for in the evaluation of gender bias.

		PARTS-OF-SPEECH
(a)	SRC	As *one* of the *first* women...
	REF$_{fr}$	En tant que l'**une**$_{Pron}$ des **premières**$_{Adj-det}$ femmes..
(b)	SRC	As a *child growing up* in Nigeria...
	REF$_{it}$	Da **bambino**$_{Noun}$ **cresciuto**$_{Verb}$ in Nigeria.
(c)	SRC	Then *an amazing* colleague...
	REF$_{es}$	Luego **una**$_{Art}$ **asombrosa**$_{Adj-des}$ colega...
		AGREEMENT
(d)	SRC	I was *the first Muslim* homecoming queen, *the first* Somali student *senator*...
	REF$_{es}$	Fui [**la primera** reina **musulmana**] del baile, [**la primera senadora**] somalí estudiantil...
(e)	SRC	She's also *been interested* in research.
	REF$_{it}$	E' [**stata** anche **attratta**] dalla ricerca .
(f)	SRC	I also *became a* high school *teacher*.
	REF$_{fr}$	Je suis aussi [**devenu un professeur**] de lycée.

Table 1: MuST-SHE target **gender-marked words** annotated per $_{POS}$ and [agreement chains].

Agreement. We also enrich MuST-SHE with linguistic information that is relevant to investigate the morphosyntactic nature of grammatical gender agreement. Gender agreement, or *concord* (Corbett, 2006; Comrie, 1999), requires that related words match the same gender form, as in the case of *phrases*, i.e. groups of words that constitute a single linguistic unit.[7] Thus, as shown in Table 1, we identify and annotate as agreement chains gender-marked words that constitute a phrase, such as a noun plus its modifiers (*d*), and verb phrases for compound tenses (*e*). Also, structures that involve a gender-marked (semi-) copula verb and its predicative complement are annotated as chains (*f*), although in such cases the agreement constraint is "weaker".[8] This annotation let us verify whether a model consistently picks the same gender paradigm for all words in the chain, enabling the assessment of its syntagmatic behaviour.

3.2 Manual annotation

POS and agreement annotation was manually carried out by 6 annotators (2 per language pair) undergoing a linguistics/translation studies MA degree, and with native/excellent proficiency in the assigned target language. For each language pair,

[4] Version 1.2: `https://ict.fbk.eu/must-she/`

[5] *Gender Bias Evaluation Testset* (Sun et al., 2019).

[6] Some POS categories (e.g. conjunctions, adverbs) are not considered since they are not subject to gender inflection.

[7] If agreement is not respected, the unit becomes ungrammatical e.g. *es*: *el$_M$ buen$_M$ niña$_F$ (the good kid).

[8] Such structure, due to the semantics of some linking verbs, can enable more flexibility. E.g. in French, *Elle est devenue*$_F$ *un*$_M$ *canard*$_M$ (*She became a duck*) is grammatical, although *un canard* (a duck) is formally masculine.

they annotated the whole corpus independently, based on detailed guidelines (see Appendix §A). For POS, we computed inter-annotator agreement (IAA) on label assignment with the kappa coefficient (in Scott's π formulation) (Scott, 1955). The resulting values of 0.92 (en-es), 0.94 (en-fr) and 0.96 (en-it) correspond to "almost perfect" agreement according to its standard interpretation (Landis and Koch, 1977). For gender agreement, IAA was calculated on the exact match of the complete chains in the two annotations. The resulting Dice coefficients (Dice, 1945) of 89.23% (en-es), 93.0% (en-fr), and 94.34% (en-it), can be considered highly satisfactory given the more complex nature of this latter task. Except for few liminal cases that were excluded from the dataset, all disagreements were reconciled.

We show the final annotation statistics in Table 2. Variations across languages are due to inherently cross-lingual differences.[9] While their discussion is beyond the scope of this work, overall these figures underscore the so far largely unaccounted variability of gender across lexical categories.

	en-es	**en-fr**	**en-it**	**M-SHE All**
POS (tot)	2099	1906	2026	6031
Art	487	325	413	1225
Pronoun	104	61	48	213
Adj-det	118	106	149	373
Adj-des	676	576	448	1700
Noun	607	344	346	1297
Verb	107	494	622	1223
AGR-CHAINS	420	293	421	1080

Table 2: Distribution of POS and agreement chains per each language and in the whole MuST-SHE corpus

4 Experimental Setting

Speech Translation models. Our experiments draw on studies exploring the relation between overall system performance, model size and gender bias. Vig et al. (2020) posit that bias increases with model size as larger systems better emulate biased training data. Working on WinoMT/ST, (Kocmi et al., 2020) correlates higher BLEU scores and gender stereotyping, whereas (Costa-jussà et al., 2020a) shows that systems with lower performance tend to produce fewer feminine translations for occupations, but rely less on stereotypical cues. To account for these findings and inspect the behavior of different models under natural conditions, we experiment with three end-to-end ST solutions, namely: LARGE-BPE, SMALL-BPE, and SMALL-CHAR (see Appendix B for complete details about the models and training setups).

Developed to achieve state-of-the-art performance, **LARGE-BPE** models rely on Transformer (Vaswani et al., 2017) and are trained in rich data conditions (1.25M ASR/ST utterances) by applying BPE segmentation (Sennrich et al., 2016). To achieve high performance, we made use of: *i)* all the available ST training corpora for the languages addressed;[10] *ii)* consolidated data augmentation methods (Nguyen et al., 2020; Park et al., 2019; Jia et al., 2019); and *iii)* knowledge transfer techniques from ASR and MT, namely component pre-training and knowledge distillation (Weiss et al., 2017a; Bansal et al., 2019). In terms of BLEU score – 34.12 on en-es, 40.3 on en-fr, 27.7 on en-it – our LARGE-BPE models compare favorably with recently published results on MuST-C test data (Bentivogli et al. 2021[11] and Le et al. 2021[12]).

Also built with the same (Transformer-based) core technology, the other systems, **SMALL-BPE** and **SMALL-CHAR**, allow for apples-to-apples comparison between the different capabilities of BPE and character-level tokenization, namely: *i)* the syntactic advantage of BPE in managing several agreement phenomena (Sennrich, 2017; Ataman et al., 2019), and *ii)* the higher capability of character-level at generalizing morphology (Belinkov et al., 2020). Given the morphological and syntactic nature of gender, such differences make them enticing candidates for further analysis. So far, Gaido et al. (2021) carried out the only study interplaying the two segmententation methods and gender bias, and found that – in spite of lower overall performance – character tokenization results in higher production of feminine forms for ST. By exploiting our new enriched resource, we intend to further test this finding and extend the analysis to gender agreement. Thus, for the sake of comparison with (Gaido et al., 2021), we train these systems in the same controlled data environment, i.e. on the MuST-C corpus only.

Evaluation method. We employ the enriched MuST-SHE corpus to assess generic performance

[9] Spanish, for instance, relies less than French or Italian on the gender-enforcing to be auxiliary, resulting in less gender-marked verbs (*fr*: est parti/ie; *it*: è partita/o; *es*: se ha ido).

[10] We are aware that MuST-C is characterized by a majority (70%) of masculine speakers (Gaido et al., 2020) . For the other training resources, comprehensive statistics are not available but we can safely consider them as similarly biased.

[11] 32.93 on en-es, 28.56 on en-it.

[12] 28.73 on en-es, 34.98 on en-fr, 24.96 on en-it.

		BLEU	All-Cov	All-Acc	F-Acc	M-Acc
en-es	SMALL-BPE	27.6	65.0	64.1	45.8	79.6
	SMALL-CHAR	26.5	64.2	67.3	**52.8**	79.6
	LARGE-BPE	**34.1**	**72.0**	**69.1**	**52.8**	**83.6**
en-fr	SMALL-BPE	25.9	55.7	64.9	50.3	78.1
	SMALL-CHAR	24.2	55.9	68.5	**57.7**	78.2
	LARGE-BPE	**34.3**	**64.3**	**70.9**	57.1	**83.4**
en-it	SMALL-BPE	21.0	53.1	67.7	52.3	80.3
	SMALL-CHAR	20.7	52.6	**71.6**	**57.2**	83.9
	LARGE-BPE	**27.5**	**59.2**	69.1	52.2	**85.4**

Table 3: BLEU, coverage and gender accuracy scores computed on MuST-SHE.

and gender translation at several levels of granularity. Evaluating gender translation under natural conditions grants the advantage of inspecting diverse informative phenomena. Concurrently, however, the intrinsic variability of natural language can defy automatic approaches based on reference translations: as language generation is an open-ended task, in our specific setting system's outputs may not contain the exact gender-marked words annotated in MuST-SHE. In fact, the released MuST-SHE evaluation script (Gaido et al., 2020) first determines dataset *coverage*, i.e. the proportion of annotated words that are generated by the system, and on which gender translation is hence measurable. Then, it calculates *gender accuracy* as the proportion of words generated in the correct gender among the measurable ones. As a result, all the *out of coverage* words are necessarily left unevaluated.

For all **word-level** gender evaluations (§5.1 and §5.2), we compute accuracy as in the official MuST-SHE script, while for **chain-level** gender agreement evaluation (§6.1) we modify it to process full agreement chains instead of single words.[13]

Finally, since we aim at gaining exhaustive, qualitative insights into systems' behaviour, and at ensuring a sound and thorough multifaceted evaluation, we overcome the described coverage limitation of the automatic evaluation by complementing it with a manual analysis of all the words and agreement chains that remained out of coverage. This extensive manual evaluation was carried out via a systematic annotation of systems' outputs, performed by the same linguists that enriched MuST-SHE, who provided the appropriate knowledge of both resource and evaluation tasks.

5 Word-level Evaluation

5.1 Overall quality and gender translation

Table 3 presents SacreBLEU (Post, 2018),[14] coverage, and gender accuracy scores on the MuST-SHE test sets. All language directions exhibit a consistent trend: LARGE-BPE systems unsurprisingly achieve by far the highest overall translation quality. Also, in line with previous analyses (Di Gangi et al., 2020), SMALL-BPE models outperform the CHAR ones by ∼1 BLEU point. The higher overall translation quality of LARGE-BPE models is also reflected by the coverage scores (All-Cov), where they generate the highest number of MuST-SHE gender-marked words for all language pairs.

By turning to overall gender accuracy (All-Acc) though, the edge previously assessed for the bigger state-of-the-art systems ceases to be clear-cut: for en-es and en-fr LARGE-BPE systems outperform the concurring SMALL-CHAR by ∼2 points only – a slim advantage compared to the huge gap observed on BLEU score –, whereas SMALL-CHAR proves the best at translating gender for en-it.

We further zoom into the comparison of gender translation for feminine (F-Acc) and masculine (M-Acc) forms, where we can immediately assess that all ST models are skewed toward a disproportionate production of masculine forms (on average, 53.1% for F vs. 81.3% for M). However, focusing on LARGE-BPE models, we discover that their higher global gender accuracy (All-Acc) is actually due to the higher generation of masculine forms, while they do not compare favorably when it comes to feminine translation. In fact, in spite of achieving the lowest generic translation quality, SMALL-CHAR prove on par (for en-es) or even better (for en-it and en-fr) than LARGE-BPE models at handling feminine gender translation.

In light of the above, our results reiterate the importance of dedicated evaluations that, unlike holistic metrics, are able to disentangle gender phenomena. As such, we can confirm that higher generic performance does not entail a superior capacity of producing feminine gender. This does not only emerge, as per Gaido et al. (2021), in the comparison of (small) BPE- and char-based ST models. Rather, even for stronger systems, we attest how profiting from a wealth of – uncurated and synthetic (Bender et al., 2021) – data does not grant advantages to address gender bias. This motivates us to continue our multifaceted evaluation by taking into account only small models – henceforth CHAR and BPE – that, being trained on the same MuST-C data, allow for sound and transparent comparison.

[13] The scripts will be released upon paper acceptance.

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

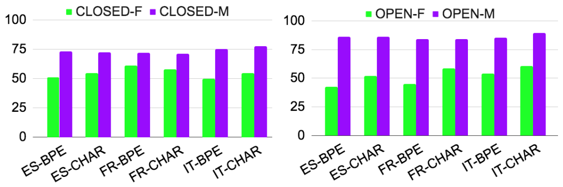

Figure 1: F *vs.* M accuracy for closed and open class words.

		Verbs		Nouns		Adj-des	
		F-Acc	M-Acc	F-Acc	M-Acc	F-Acc	M-Acc
en-es	BPE	44.4	**93.8**	21.1	89.0	57.4	**80.0**
	CHAR	**60.0**	84.2	**37.4**	**89.7**	**61.2**	79.7
en-fr	BPE	51.3	**79.8**	16.4	93.5	50.6	78.6
	CHAR	**68.4**	75.0	**27.4**	**95.3**	**63.0**	**81.4**
en-it	BPE	63.7	83.7	28.6	92.2	62.0	76.7
	CHAR	**66.7**	**89.2**	**33.3**	**94.3**	**70.6**	**84.5**

Table 4: F *vs.* M Accuracy scores per open class POS.

5.2 Word classes and Parts-of-speech

At a finer level of granularity, we use MuST-SHE extension to inspect gender bias across open and closed class words. Their coverage ranges between 74-81% for function words, but it shrinks to 44-59% for content words (see Appendix C.1). This is expected given the limited variability and high frequency of functional items in language. Instead, the coverage of feminine and masculine forms is on par within each class for all systems, thus allowing us to evaluate gender accuracy on a comparable proportion of generated words. A bird's-eye view of Figure 1 attests that, although masculine forms are always disproportionately produced, the gender accuracy gap is amplified on the open class words. The consistency of such a behaviour across languages and systems suggests that content words are involved to a greater extent in gender bias.

We hence analyse this more problematic class by looking into a breakdown of the results per POS, while for function words' gender accuracy we refer to Appendix C.2. Table 4 presents results for *verbs*, *nouns* and *descriptive adjectives*. First, in terms of system capability, CHAR still consistently emerge as the favorite models for feminine translation. What we find notable, though, is that even within the same class we observe evident fluctuations, where nouns come forth as the most biased POS with a huge divide between M and F accuracy (52–77 points). Specifically, scores below 50% indicate that feminine forms are generated with a probability that is below random choice, thus signalling an extremely strong bias.

In light of this finding, we hypothesize that semantic and distributional features might be a factor to interpret words' gender skew. Specifically, occupational lexicon (e.g. lawyer, professor) makes up for most of the nouns represented in MuST-SHE (∼70%). While such a high rate of professions in TED data is not surprising *per se*,[15] it singles out that professions may actually represent a category where systems largely rely on spurious cues to perform gender translation, even within natural conditions that do not ambiguously prompt stereotyping. We exclude basic token frequency by POS as a key factor to interpret our results, as MuST-SHE feminine nouns do not consistently appear as the POS with the lowest number of occurrences, nor do they have the lowest F:M ratio within MuST-C training data. As discussed in §8, we believe that our breakdown per POS is informative inasmuch it prompts qualitative considerations on how to pursue gender bias mitigation in models and corpora (Czarnowska et al., 2021; Doughman et al., 2021).

5.3 Manual analysis

We manually inspect CHAR and BPE system's output on the out-of-coverage (OOC) words that could not be automatically evaluated (see "All-Cov" column in Table 3), which amount to more than 5,000 instances. As shown in Table 5, our analysis discerns between OOC words due to *i) translation errors* (Err),[16] and *ii)* adequate *alternative* translations (i.e. meaning equivalent) for the expected gender-marked word. Such alternatives comprise instances in which word omission is acceptable (Alt-O) (Baker, 1992), and rewordings through synonyms or paraphrases. Since our focus remains on gender translation, we distinguish when such rewordings are generated with correct (Alt-C) or wrong (Alt-W) gender inflections, as well as neutral expressions devoid of gender-marking (Alt-N). Note that – with respect to English (Cao and Daumé III, 2020; Vanmassenhove et al., 2021; Sun et al., 2021) – overcoming the structural pervasiveness of gender specifications in grammatical gender languages is extremely challenging (Gabriel et al., 2018a), but some rewordings can enable indirect neutral language (INL)[17] (López, 2020).

The results of the analysis are shown in Figure 2.

[15] As TED talks are held by field experts, references to education and titles are quite common (MacKrill et al., 2021).

[16] Errors range from misspelling to complete gibberish.

[17] INL relies on generic expressions rather than gender-specific ones (e.g. *service* vs. *waiter/tress*) See §8.

		ERRORS
	SRC	Robert became **fearful** and **withdrawn**.
	REF$_{it}$	Robert divenne **timoroso** e **riservato**.
	OUT$_{it}$	Robert diventò **timore** e **con John**.
		(*Robert became fear and with John*)
		ALTERNATIVES
Alt-O	SRC	He was **an** artist.
	REF$_{fr}$	C'était **un** artiste.
	OUT$_{fr}$	C'était (__) artiste.
Alt-C	SRC	These girls [...], they are so **excited**...
	REF$_{es}$	Estas niñas [...], están **emocionadas**...
	OUT$_{es}$	Estas chicas [...], están **entusiasmadas**...
Alt-W	SRC	Mom [...] **became manager...**
	REF$_{it}$	Mamma [...] venne **messa** a capo di...
	OUT$_{it}$	La madre [...] diventò **capo** di...
Alt-N	SRC	I **felt** really good.
	REF$_{fr}$	Je me suis **senti** vraiment bien
	OUT$_{fr}$	Je me **sentais** vraiment bien .

Table 5: Classification of OOC words.

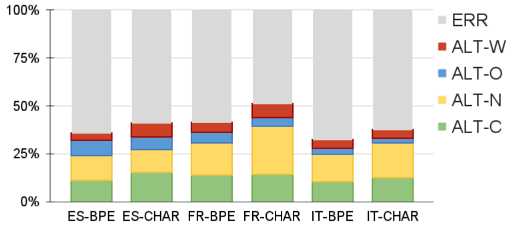

Figure 2: Proportion of OOC words due to translation errors and alternative translations per system.

		All			Feminine			Masculine		
		C	W	NO	C	W	NO	C	W	NO
en-es	*bpe*	74.3	**24.6**	**1.2**	33.9	**64.4**	**1.7**	95.5	**3.6**	0.9
	char	**78.4**	21.0	0.6	**42.4**	57.6	0.0	**96.6**	2.6	0.9
en-fr	*bpe*	67.9	**31.0**	**1.2**	54.1	**45.9**	0.0	78.7	**19.1**	**2.1**
	char	**76.7**	22.3	1.0	**57.5**	40.0	**2.5**	**88.9**	11.1	0.0
en-it	*bpe*	71.7	**27.5**	0.7	47.4	**50.9**	**1.8**	88.9	**11.1**	0.0
	char	**78.5**	20.0	**1.5**	**54.2**	44.1	1.7	**97.4**	1.3	**1.3**

Table 6: Accuracy scores for gender agreement.

Surprisingly, we find that BPE models – in spite of their higher BLEU scores – accumulate more translation errors than their CHAR counterparts.[18] Conversely, CHAR models generate an overall higher proportion of alternatives and, more importantly, alternatives whose gender translation is acceptable (-N, -C). This suggests that CHAR output is characterized by a favourable *adequate* variability that conveys both lexical meaning and gender realization better than BPE.

As a final remark, we find that all systems produce a considerable amount of neutral alternatives in their outputs. To gain insight into such neutralizations, we audit on which POS they are realized. Accordingly, we find that neutralizations of adjectives and nouns are quite limited, and concern the production of epicene synonyms (e.g. *en*: happy; *es-ref*: contento/a; *es-out*: feliz). Verbs, instead, are largely implicated in the phenomenon, since inflectional changes in tense and aspect paradigms (e.g., present, imperfective) that do not convey gender distinctions are feasible (see the -N example in Table 5). Such range of alternatives for verbs is in fact also reflected by its lowest coverage among all POS (as low as ∼32%). Finally, paraphrases based on verbs also represent the most frequent way to neutralize other POS in the output. Since such expressions are suitable, or even preferable, for several scenarios (e.g. to substitute masculine generics, to avoid making unlicensed gender assumptions) our finding encourages the creation of test sets accounting for such a third viable direction, and can shed light on systems' potential to produce INL alternatives.

6 Gender Agreement Evaluation

6.1 Automatic analysis

The final step in our multifaceted analysis goes beyond the word level to inspect agreement chains in translation. To this aim, we define *coverage* as the proportion of generated chains matching with those annotated in MuST-SHE. Then, the *accuracy* of the generated chains accounts for 3 different cases where: *i)* agreement is respected, and with the correct gender (C); *ii)* agreement is respected, but with the wrong gender (W); and *iii)* both feminine and masculine gender inflections occur together, and thus agreement is not respected (NO).

Table 6 shows accuracy scores for all MuST-SHE agreement chains (All), also split into feminine (F) and masculine (M) chains. The overall results are promising: we find very few instances (literally 1 or 2) in which ST systems produce an ungrammatical output that breaks gender agreement (NO). In fact, both systems tend to be consistent with one picked gender for the whole dependency group. Thus, in spite of previous MT studies concluding that character-based segmentation results in poorer syntactic capability (Belinkov et al., 2020), respecting concord does not appear as an issue for any of our small ST models. For the sake of comparability, however, we note that our evaluation involves language pairs that do not widely resort to long-range dependencies; this may contribute to explaining why CHAR better handles

[18] We noticed that CHAR's lower translation quality may have to do more with fluency rather than lexical issues.

correct gender agreement.[19]

Overall, agreement translation was measured on a lower *coverage* (30-50%) – presented in Appendix D.1 – than the world-level one (§3). While this is expected given the strict requirement of generating full chains with several words, we recover such a loss by means of the comprehensive manual evaluation discussed below.

6.2 Manual analysis

Our manual inspection recovers a total of ∼1,200 OOC agreement chains from CHAR and BPE output. Similarly to the approach employed for single words (§5.3), we discern between OOC chains due to: *i)* translation *errors* (Err), and *ii) alternative* translations preserving the source meaning. We distinguish different types of alternatives. First, alternatives that do no exhibit a morphosyntactic agreement phenomenon to be judged, as in the case of neutral paraphrases or rewordings consisting of a single word (NO-chain). Instead, when the generated alternative chain exhibits gender markings, we distinguish if the chosen gender is correct (C), wrong (W), or if the system produces a chain that does not respect gender agreement, because it combines both feminine and masculine gender inflections (NO).

The outcome of such OOC chains categorization is presented in Figure 3. Interestingly, such results are only partially corroborating previous analyses. On the one hand, unlike the OOC words' results discussed in §5.3, we attest that CHAR models produce the highest proportion of translation errors. Thus, it seems that CHAR capability in producing adequate alternatives is confined to the single-word level, whereas it exhibits a higher failure rate on longer sequences. On the other hand, by looking at alternative chains, CHAR still emerges as the best at properly translating gender agreement, with the highest proportion of chains with correct gender (C), and the lowest one with wrong gender (W).

Finally, in line with our automatic evaluation (Table 6), we confirm that respecting agreement is not an issue for our ST models: we identify only 3 cases (2 for en-fr BPE, 1 for en-fr CHAR) where concord is broken (NO). Given the rarity of such instances, we are not able to draw definitive conclusions on the nature of these outliers. Nonetheless, we check the instances in which agreement was not respected (both in and out of coverage). We see that cases of broken concord also concern extremely simple phrases, consisting of a noun and its modifier (e.g. en: *talking to [this inventor],...because he*; fr: *parler à* ***[cette***$_F$ ***inventeur***$_M$*]..., parce qu' il*). However, the most common type among these outliers are constructions with semi-copula verbs (e.g. en: *She... [became a vet]*; it: *...E' [****diventata***$_F$ ***un***$_M$ ***veterinatrio***$_M$*]*), which – as discussed in §3.1 – exhibit a weaker agreement constraint.

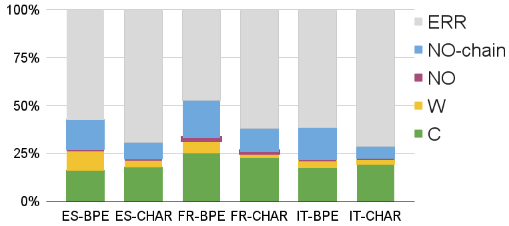

Figure 3: Proportion of OOC chains due to translation errors or alternative agreement translations per system.

7 Conclusion

The complex system of grammatical gender languages entails several morphosyntactic implications for different lexical categories. In this paper, we underscored such implications and explored how different POS and grammatical agreement are involved in gender bias. To this aim, we enriched the MuST-SHE benchmark with new linguistic information, and carried out an extensive evaluation on the behaviour of ST models built with different segmentation techniques and data quantities. On 3 language pairs (en-es/fr/it) our study shows that, while all POS are subject to masculine skews, they are not impacted to the same extent. Respecting gender agreement for the translation of related words, instead, is not an issue for current ST models. We also find that ST generates a considerable amount of neutral expressions, suitable to replace gender-inflected ones, which however current test sets do not recognize. Overall, our work reiterates the importance of dedicated analyses that, unlike holistic metrics, can single out system's behaviour on gender phenomena. Accordingly, our results are in line with previous studies showing that, in spite of lower generic performance, character-based segmentation favours feminine translation at different levels of granularity. As our MuST-SHE extension is available for both ST and MT, we invite MT studies to start from our discoveries and resource.

[19] Due to space constraints we refer to Appendix D.2 for an analysis of longer-range cases of subject-verb agreement.

8 Impact statement

In this paper, we evaluate whether and to what extent ST models exhibit biased behaviors by systematically and disproportionately favoring masculine forms in translation. Such a behavior is problematic inasmuch it leads to under-representational harms by reducing feminine visibility (Blodgett et al., 2020; Savoldi et al., 2021).

Broader impact. While the focus of this work is on the analysis itself, our insights prompt broader considerations. Specifically, our investigation on the relation between model size/segmentation technique and gender bias provides initial cues on which models and components to audit and implement toward the goal of reducing gender bias. This, in particular, may be informative to define the path for emerging direct ST technologies. Also, our results disaggregated by POS invite reflections on how to intend and mitigate bias by means of interventions on the training data. In fact, while it is known that the MuST-C corpus (Cattoni et al., 2020) used for training comprises a majority of masculine speakers,[20] the fact that certain lexical categories are more biased than others suggests that, on top of more coarse-grained quantitative attempts at gender balancing (Costa-jussà and de Jorge, 2020), data curation ought to account for more sensitive, nuanced, and qualitative asymmetries. These also imply *how*, rather than only *how often*, gender groups are represented (Wagner et al., 2015; Devinney et al., 2020). Also, while nouns come forth as the most problematic POS, current practices of data augmentation based on a pre-defined occupational lexicon may address stereotyping (Saunders and Byrne, 2020), but do not increase the production of other nonetheless skewed lexical categories. Overall, our enriched resource[21] can be useful to monitor the validity of different technical interventions.

Ethic statement. The use of gender as a variable (Larson, 2017) warrants some ethical reflections.

Our evaluation on the MuST-SHE benchmark exclusively accounts for linguistic gender expressions. As reported in MuST-SHE data statement (Bender and Friedman, 2018),[22] also for the subset of sentences that contain first-person references[23] (e.g. *I'm a student*), speakers' gender information is manually annotated based on the personal pronouns found in their publicly available personal TED profile, and used to check that the indicated (English) linguistic gender forms are rendered in the gold standard translations.

While our experiments are largely limited to 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.[24] This is also due to the fact that unlike English – which finds itself for several cultural and linguistic reasons as a leader of change toward inclusive forms (Ackerman, 2019) – Direct Non-binary Language based on neomorphemes (Shroy, 2016; Papadopoulos, 2019; Knisely, 2020) is non-trivial to fully implement in grammatical gender languages (Hellinger and Bußman, 2001; Gabriel et al., 2018b) and still largely object of experimentation (Redazione, 2020; Attig and López, 2020). However, our manual evaluation expands to the possibility of INL strategies that could be detected in system's output. We underscore that such strategies are recommended and fruitful to avoid the gendering of referents, but are to be considered as concurring to – rather than replacements of – emerging linguistic innovations (López, 2020).

Lastly, we signal that direct ST models may leverage speakers' vocal characteristics as a gender cue to infer gender translation. Although the potential risks of such condition do not emerge in our setting we endorse the point made by Gaido et al. (2020). Namely, direct ST systems leveraging speaker's vocal biometric features as a gender cue can entail real-world dangers, like the categorization of individuals by means of biological essentialist frameworks (Zimman, 2020). This can reduce gender to stereotypical expectations about how masculine or feminine voices should sound, and can be especially harmful to transgender individuals, as it can lead to misgendering (Stryker, 2008) and invalidation. Note that we experimented with unmodified models for the sake of hypothesis testing without adding variability, but real-world deployment of ST technologies must account for the potential harms arising form the use of direct ST technologies *as is*.

[20] https://ict.fbk.eu/must-speakers/

[21] It will be released under the same CC BY NC ND 4.0 International license as MuST-SHE.

[22] https://ict.fbk.eu/must-she/

[23] Category 1 in the corpus.

[24] Saunders et al. (2020) enriched WinoMT to account for non-binary language. While it is only available for MT, such annotations consist of placeholders for neutrality rather than actual non-binary expressions.

References

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

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online). Association for Computational Linguistics.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2021. The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses. *arXiv preprint arXiv:2110.09216*.

Maria Antoniak and David Mimno. 2021. Bad Seeds: Evaluating Lexical Methods for Bias Measurement. 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 1889–1904, Online. Association for Computational Linguistics.

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.

Remy Attig and Ártemis López. 2020. Queer Community Input in Gender-Inclusive Translations. *Linguistic Society of America [Blog]*.

Mona Baker. 1992. *A Coursebook on Translation*. Routledge.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on High-resource Speech Recognition Improves Low-resource Speech-to-text Translation. 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 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

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. 2009. Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?̀. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? 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 2873–2887, Online. Association for Computational Linguistics.

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.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. 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 1004–1015, 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.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63.

Bernard Comrie. 1999. Grammatical gender systems: a linguist's assessment. *Journal of Psycholinguistic research*, 28(5):457–466.

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

Greville G. Corbett. 2006. *Agreement*. Cambridge University Press.

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

Marta R Costa-jussà, Christine Basta, and Gerard I Gállego. 2020a. Evaluating gender bias in speech translation. *arXiv e-prints*, pages arXiv–2010.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning Neural Machine Translation on Gender-Balanced Datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Marta R. Costa-jussà, 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*.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *ArXiv*, abs/2106.14574.

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, and Marco Turchi. 2019. Adapting Transformer to End-to-end Spoken Language Translation. In *Proceedings of INTERSPEECH*, pages 1133–1137, Graz, Austria.

Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. Gender bias in text: Origin, taxonomy, and implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44, Online. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

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

Ute Gabriel, Pascal M. Gygax, and Elisabeth A. Kuhn. 2018b. 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.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. How to Split: the Effect of Word Segmentation on Gender Bias in Speech Translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online. Association for Computational Linguistics.

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

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages? In *Proceedings of the 2019 Workshop on Widening NLP*, pages 64–67.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B Reflexivization as an Unambiguous Testbed for Multilingual Multi-Task Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648.

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.

Marlis Hellinger and Hadumond Bußman. 2001. *Gender across Languages*. John Benjamins Publishing, Amsterdam, The Netherlands.

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

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation. In *Proceedings of the Speech and Computer - 20th International Conference (SPECOM)*, pages 198–208, Leipzig, Germany. Springer International Publishing.

Levi CR Hord. 2016. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. *Western Papers in Linguistics/Cahiers linguistiques de Western*, 3(1):4.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Giménez. Adrià, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233, Barcelona, Spain.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184, Brighton, UK.

Kris Aric Knisely. 2020. Le français non-binaire: Linguistic forms used by non-binary speakers of French. *Foreign Language Annals*, 53(4):850–876.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender Coreference and Bias Evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364.

Mascha Kurpicz-Briki and Tomaso Leoni. 2021. A World Full of Stereotypes? Further Investigation on Origin and Gender Bias in Multi-Lingual Word Embeddings. *Frontiers in big Data*, 4:20.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

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.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824, Online. Association for Computational Linguistics.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. *arXiv preprint arXiv:2109.03858*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proceedings of Interspeech 2019*, pages 1128–1132, Graz, Austria.

Ártemis López. 2020. Cuando el lenguaje excluye: Consideraciones sobre el lenguaje no binario indirecto. *Cuarenta Naipes*, 3:295–312.

Kate MacKrill, Connor Silvester, James W Pennebaker, and Keith J Petrie. 2021. What makes an idea worth spreading? language markers of popularity in ted talks by academics and other speakers. *Journal of the Association for Information Science and Technology*.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2021. Socially Aware Bias Measurements for Hindi Language Representations. *arXiv preprint arXiv:2110.07871*.

Ranko Matasović. 2004. *Gender in Indo-European*. Universitaetsverlag Winter, Heidelberg.

Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender Bias in Natural Language Processing Across Human Languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54.

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.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2020. Improving Sequence-to-sequence Speech Recognition Training with On-the-fly Data Augmentation. In *Proceedings of the 2020 International Conference on Acoustics, Speech, and Signal Processing – IEEE-ICASSP-2020*, Barcelona, Spain.

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*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR Corpus Based on Public Domain Audio Books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, Brisbane, Australia.

Benjamin Papadopoulos. 2019. Morphological Gender Innovations in Spanish of Gender queer Speakers.

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.

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.

Redazione. 2020. EFFEQU è la prima casa editrice italiana a introdurre la schwa nei suoi libri. Accessed: 2021-11-14.

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

Adithya Renduchintala and Adina Williams. 2021. Investigating failures of automatic translation in the case of unambiguous gender. *arXiv preprint arXiv:2104.07838*.

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.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Proceedings of Visually Grounded Interaction and Language (ViGIL)*, Montréal, Canada. Neural Information Processing Society (NeurIPS).

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, Barcelona, Spain (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*.

Paul Schachter and Timothy Shopen. 2007. Parts-of-speech systems. *Language Typology and Syntactic Description. Vol. 1: Clause Structure*, pages 1–60.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Pubulic Opinion Quarterly*, 19:321–325.

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*.

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.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *arXiv preprint arXiv:2102.06788*.

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.

Jörg Tiedemann. 2016. Opus – parallel corpora for everyone. *Baltic Journal of Modern Computing*, page 384. Special Issue: Proceedings of the 19th Annual Conference of the European Association of Machine Translation (EAMT).

Jonas-Dario Troles and Ute Schmid. 2021. Extending Challenge Sets to Uncover Gender Bias in Machine Translation: Impact of Stereotypical Verbs and Adjectives. *arXiv e-prints*, pages arXiv–2107.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS 2017*, pages 5998–6008, Long Beach, California. NIPS.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9.

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

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284.

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*.

A MuST-SHE Manual Annotation

POS and agreement chains annotations were carried out on MuST-SHE reference translations. To ensure precision, the two layers of linguistic information have been added *i)* in the course of two separate annotation processes; *ii)* following strict and comprehensive guidelines.

A first version of the guidelines was written by an expert linguist based on a preliminary manual analysis of a MuST-SHE sample. Successively, such guidelines have been refined and improved by means of discussions with the annotators, who had carried out a trail annotation round to get acquainted with MuST-SHE language data. Here, we provide the link to the final version of POS and agreement chains **annotation guidelines**:

https://bit.ly/3CdU50s

The 6 annotators were all interning students undergoing a MA degree in Linguistics/Translation Studies, who were selected among other 120 candidates. We ensured that at least one annotator per language was a native speaker. Since the annotations were carried out in the course of this more extensive curricular internship, there was no task-associated compensation.

B ST Models

In this section we describe in detail the ST models created for our study, whose source code will be publicly released upon acceptance of the paper.

B.1 Architecture

The architecture of our ST models is composed of two strided 2D convolutional layers with 64 3x3 kernels, followed by a Transformer (Vaswani et al., 2017) with 11 encoder layers and 4 decoder layers. The two 2D convolutions reduce the length of the input sequence by a factor of 4, allowing the processing by the Transformer encoder layers that have a quadratic memory complexity with respect to the input length (because of the self-attention mechanism). The weights of the encoder self-attention matrices are biased to be close to 0 for elements far from the matrix diagonal (i.e. for elements that are far from the considered vector) with a logarithmic distance penalty (Di Gangi et al., 2019). In both encoder and decoder Transformer layers, we use 8 attention heads, 512 embedding features, and 2048 features for the feed-forward inner-layers. The resulting number of parameters is 60M for BPE models and 52M for character-based models.

B.2 Data

Since the amount of ST data is limited (i.e. MuST-C – Cattoni et al. 2020 – and Europarl-ST – Iranzo-Sánchez et al. 2020), knowledge transfer from the ASR and MT tasks is leveraged by respectively initializing the ST encoder with ASR pretrained weights (Weiss et al., 2017b; Bansal et al., 2019) and by distilling knowledge from a strong MT teacher (Liu et al., 2019). The ASR model used for the pretraining has been trained on Librispeech (Panayotov et al., 2015), Mozilla Common Voice,[25] TEDLIUM-v3 (Hernandez et al., 2018), and the utterance-transcript pairs of the ST corpora and of How2 (Sanabria et al., 2018). The teacher MT models, instead, are trained on a subsample of the Opus (Tiedemann, 2016) repository, filtered using the cleaning pipeline of ModernMT.[26]

SpecAugment is applied to the source audio 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). Time stretch (Nguyen et al., 2020) alters the input utterance with probability of 0.3 and stretching factor sampled uniformly for each utterance between 0.8 and 1.25 is also used to alter the input audio. The target text was tokenized with Moses.[27] We normalized audio per-speaker and extracted 40 features with 25ms windows sliding by 10ms with XNMT[28] (Neubig et al., 2018).

The LARGE-BPE model is trained on all the available (ST and distilled) data for a total of ∼1.25M pairs, while the SMALL-BPE and SMALL-CHAR are trained only on the MuST-C data for a total of 250-275k pairs. The encoder pretraining is used for all the models. The SMALL-* models are initialized with the weights of an ASR trained only on the *(audio, transcript)* pairs of MuST-C, while the LARGE-BPE is initialized with an ASR trained on all the available data.

For the small and large models leveraging BPE, we employed 8k merge rules, while we used a set of 250-400 characters for the SMALL-CHAR model. The resulting vocabulary sizes are reported in Table 7.

[25] https://voice.mozilla.org/

[26] https://github.com/modernmt/modernmt

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

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

	en-es	en-fr	en-it
Large-BPE	11,940	12,136	11,152
Small-Char	464	304	256
Small-BPE	8,120	8,048	8,064

Table 7: Resulting dictionary sizes.

B.3 Training procedure

The models are trained using 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. Each mini-batch consists of 8 samples, we set the update frequency to 8, and we train on 4 GPUs, so a batch contains 256 samples.

The LARGE-BPE training is performed in three consecutive steps. First, we train on synthetic data obtained by automatically translating the ASR corpora transcript with our MT model (Jia et al., 2019). Second, we fine-tune on the ST corpora. In both these training phases the model is optimised to learn the output distributions of the MT teacher (via word-level knowledge distillation). Lastly, the model is fine-tuned on the ST corpora using label-smoothed cross entropy. Trainings are stopped after 5 epochs without improvements on the validation loss and we average 5 checkpoints around the best on the validation set. In the rich-data condition case, as we did not see benefits by the average of the checkpoints, we used the best checkpoint instead. As a validation set we rely on the MuST-C gender-balanced dev set (Gaido et al., 2020).

Our code is built on the Fairseq library (Ott et al., 2019) and trainings are performed on 4 K80 GPUs, lasted 4 days for the MuST-C-only trainings and 12 days for the rich-data models.

C Word-level Evaluation

C.1 Coverage per open and closed class words

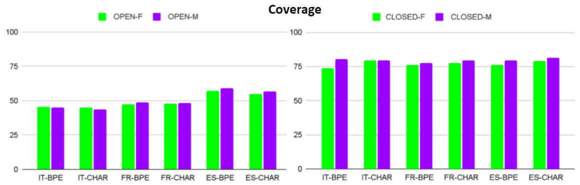

Figure 4: F *vs* M coverage per open and closed class words.

As we can see in Figure 4, function words have a higher coverage than content words. This is expected given the limited variability and high frequency of functional items in language. Instead, the coverage of feminine and masculine forms is on par within each class for all systems, thus allowing us to evaluate gender accuracy on a comparable proportion of generated words.

C.2 Gender accuracy per closed class POS

		Art		Pronoun		Adj-det	
		F-Acc	M-Acc	F-Acc	M-Acc	F-Acc	M-Acc
en-es	*bpe*	51.35	**70.0**	**52.0**	84.9	49.1	86.1
	char	**53.5**	68.4	51.7	**85.7**	**59.3**	**91.2**
en-fr	*bpe*	**52.0**	**69.2**	**65.5**	**78.3**	**82.9**	**79.5**
	char	50.8	68.6	54.2	77.3	79.1	78.6
en-it	*bpe*	47.2	74.6	**75.0**	71.4	50.9	81.8
	char	**52.2**	**76.8**	52.9	**77.8**	**61.8**	**83.3**

Table 8: F *vs.* M accuracy scores per closed class POS.

As we can see in Table 8, CHAR's otherwise attested advantage over BPE is not consistent for function words, where we find variations across POS and languages. Such variations may be due to the fairly restricted amount of MuST-SHE *pronouns* and *limiting adjectives* (Adj-det) on which accuracy can be computed in MuST-SHE (see Table 2), which make very fine-grained evaluations particularly unstable. Additionally – since the present POS evaluation still remains at the word level – we are not able to ponder whether gender translations for modifiers (i.e. articles, determiners) is to some extent constrained by the content words they refer to. We intend to explore such hypothesis in future work by intersecting POS and agreement annotations.

D Agreement Evaluation

D.1 Agreement coverage

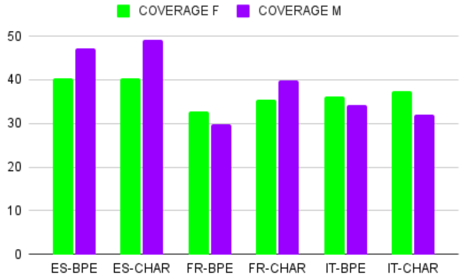

Figure 5: F *vs* M chains coverage

Figure 5 shows coverage of fully generated agreement chains split into feminine (F) and mas-

SRC	**A** young **scientist** that I was working with ..., Rob, **was**..
REF$_{it}$	[**Un** giovane **scienziato**] con cui lavoravo ..., Rob, è **stato**..

Table 9: Example of subject-verb agreement in MuST-SHE.

culine (M) forms. Although we attest notable variations across languages and gender forms, overall masculine and feminine chains are both produced at comparable rate.

D.2 Manual analysis of subject-verb agreement

Considering long-range dependencies that go beyond the phrase level, a gender relevant co-variation is also that of *subject-verb* agreement, as the one shown in Table 9 (see also footnote 1). To account for such longer spans, we considered all MuST-SHE sentences where both *i)* a word (or chain) functioning as a subject, and *ii)* its referring verb or predicative complement are annotated as gender-marked words in the references. We identified 55 sentences for en-es, 54 for en-fr and 41 for en-it, and we manually analyzed all the corresponding systems' outputs.

We found that, even in the case of dependencies within a longer range, systems largely respect agreement in translation and consistently pick the same gender form for all co-related words. In fact, we identified only 4 cases where concord is broken: 1 case each for BPE and CHAR for en-es and en-it, and none for en-fr. Among these 4 cases, we found the above discussed weaker gender-enforcing structures (see the description of (semi-)copula verbs and their predicative complements in §6.2), and we also detected what resembles *agreement attraction errors* (Linzen et al., 2016). Namely, the model does not produce a verb or complement in agreement with its actual (but distant) subject, as other words intervene in the sentence and agreement is conditioned by the verb/complement's preceding word: as a results, subject-verb agreement is not respected. The following (long) sentence is an example of such an attraction error, where the complement *desperate* is inflected in the same masculine and plural form as its just preceding noun rather than the chain functioning as subject (*the nurse*).

(*en-src*) I watched in horror heartbreaking footage of **the head nurse**, Malak, in the aftermath of the bombing, grabbing premature babies out of their incubators, **desperate** to get them to safety, before she broke down in tears.
(*es*-CHAR: Vi una imagen horrible desgarradora de **la enfermera** (F., sing.) mi laguna, en los ratones después del bombardeo, agarrando a los bebés permaturos fuera de sus incubadores (M., pl.) **desesperados**(M., pl.) por hacerlos...

Such kind of agreement issues have more to do with overall syntactic capacity of ST models, rather than being implicated with gender bias. We can thus conclude that, even taking into account longer dependencies, agreement still does not emerge as an issue entrenched with gender bias.