# On the Dynamics of Gender Learning in Speech Translation

**Beatrice Savoldi[1,2], Marco Gaido[1,2], Luisa Bentivogli[2], Matteo Negri[2], Marco Turchi[2]**
[1] University of Trento
[2] Fondazione Bruno Kessler
beatrice.savoldi@unitn.it
{mgaido,bentivo,negri,turchi}@fbk.eu

## Abstract

Due to the complexity of bias and the opaque nature of current neural approaches, there is a rising interest in auditing language technologies. In this work, we contribute to such a line of inquiry by exploring the emergence of gender bias in Speech Translation (ST). As a new perspective, rather than focusing on the final systems only, we examine their evolution over the course of training. In this way, we are able to account for different variables related to the learning dynamics of gender translation, and investigate when and how gender divides emerge in ST. Accordingly, for three language pairs (en –> es, fr, it) we compare how ST systems behave for masculine and feminine translation at several levels of granularity. We find that masculine and feminine curves are dissimilar, with the feminine one being characterized by a more erratic behaviour and late improvements over the course of training. Also, depending on the considered phenomena, their learning trends can be either antiphase or parallel. Overall, we show how such a progressive analysis can inform on the reliability and time-wise acquisition of gender, which is concealed by static evaluations and standard metrics.

## 1 Bias Statement

Hereby, we study how Speech Translation (ST) systems deal with the generation of masculine and feminine forms for human referents. Despite the impossibility of a perfect alignment between linguistic and extra-linguistic gender reality (Ackerman, 2019; Cao and Daumé III, 2020), these forms affect the representation and perception of individuals (Stahlberg et al., 2007; Corbett, 2013; Gygax et al., 2019), and are actively used as a tool to negotiate the social, personal, and political reality of gender (Hellinger and Motschenbacher, 2015). Thus, we consider a model that systematically and disproportionately favors masculine over feminine forms as biased, since it fails to properly recognize women. Following Crawford (2017), Blodgett et al. (2020), and Savoldi et al. (2021), such behavior is regarded as harmful because language technologies misrepresent an already disadvantaged social group by reducing feminine visibility and by offering unequal service quality.

Moreover, we consider another potential cause of discrimination in *end-to-end* speech technology. Namely, by translating directly from the audio input it comes with the risk of relying on speakers' vocal characteristics – including fundamental frequency – to translate gender.[1] By using biometric features as gender cues, ST models may reduce gender to stereotypical expectations about the sound of masculine and feminine voices, thus perpetuating biological essentialist frameworks (Zimman, 2020). This is particularly harmful to transgender individuals, as it can lead to misgendering (Stryker, 2008) and a sense of invalidation.

Accordingly, we investigate the aforementioned concerns by evaluating systems' output throughout their training process, aiming to shed light on the dynamics through which gender bias emerges in translation models. Note that, while our diagnostic work focuses on the technical side of gender bias, we recognize the paramount importance of critical interdisciplinary work that foregrounds the context of development and deployment of language technologies (Criado-Perez, 2019; D'Ignazio and Klein, 2020). Also, in Section 9, we discuss the limits of working on binary language.

## 2 Introduction

Along with the massive deployment of language technologies, concerns regarding their societal impact have been raised (Hovy and Spruit, 2016; Bender et al., 2021), and glaring evidence of biased behavior has been reported by users themselves. Translation technologies are no exception. On-

[1] As in the case of ambiguous first-person references, e.g. en: *I'm tired*, es: *estoy cansado/a*.

*Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94 - 111
July 15, 2022 ©2022 Association for Computational Linguistics

line interactions exhibited that commercial engines reflect controversial gender roles (Olson, 2018), and further evaluations on both Machine (MT) and Speech translation (ST) systems confirmed that models skew towards a masculine default (Cho et al., 2019; Prates et al., 2020; Bentivogli et al., 2020), except for stereotypical representations (e.g., *nurse* or *pretty doctor* as feminine) (Kocmi et al., 2020; Costa-jussà et al., 2020).

The last few years have witnessed a growing effort towards developing preventive measures (Bender and Friedman, 2018) and mitigating strategies (Saunders and Byrne, 2020; Vanmassenhove et al., 2018; Alhafni et al., 2020). Yet, the complex nature of both neural approaches and bias calls for focused inquiries into our ST and MT models. In this regard, dedicated testing procedures have been designed to pinpoint the impact of gender bias on different categories of phenomena (Stanovsky et al., 2019; Troles and Schmid, 2021; Savoldi et al., 2022). Also, algorithmic choices underpinning the construction of current models have been re-evaluated in light of gender disparities (Renduchintala et al., 2021; Roberts et al., 2020). Despite such promising advancements, many questions still stand unanswered. When does this gender gap emerge? How does gender bias relate to progress in terms of generic performance? To what extent is gender learning altered by the chosen components? To the best of our knowledge, current studies have adopted a static approach, which exclusively focuses on systems' biased behaviors once their training is completed.

Rather than treating training as a black box, in this paper we explore the evolution of gender (in)capabilities across systems' training process. In the wake of prior work highlighting how different target segmentations affect gender bias (Gaido et al., 2021), we compare ST systems built with two techniques: character and byte-pair encoding (BPE) (Sennrich et al., 2016). For three language pairs (en→ es,fr,it), we thus examine their gender learning curves for feminine and masculine translation at several levels of granularity.

Overall, our contributions can be summarized as follows: **(1)** We conduct the first study that explores the dynamic emergence of gender bias in translation technologies; **(2)** By considering the trend and stability of the gender evolution, we find that *(i)* unlike overall translation quality, feminine gender translation emerges more prominently in the late training stages, and does not reach a plateau within the iterations required for models to converge in terms of generic performance. Such trend is however concealed by standard evaluation metrics, and unaccounted when stopping the training of the systems. *(ii)* For easily gender-disambiguated phenomena, masculine and feminine show a generally parallel and upwards trend, with the exception of nouns. Characterized by flat trends and a huge gender divide, their learning dynamics suggests that ST systems confidently rely on spurious cues and generalize masculine from the very early stages of training onwards.

## 3 Background

**Gender bias.** Gender bias has emerged as a major area of NLP research (Sun et al., 2019; Stanczak and Augenstein, 2021). A key path forward to address the issue requires moving away from performance as the only desideratum (Birhane et al., 2021), and – quoting Basta and Costa-jussà (2021) – *interpreting and analyzing current data and algorithms*. Accordingly, existing datasets (Hitti et al., 2019), language models (Vig et al., 2020; Silva et al., 2021) and evaluation practices (Goldfarb-Tarrant et al., 2021) have been increasingly put under scrutiny.

Also for automatic translation, inspecting models' inner workings (Bau et al., 2019) can help disclosing potential issues or explaining viable ways to alleviate the problem (Costa-jussà et al., 2022). Concurrently, studies in both MT and ST foregrounded how taken-for-granted algorithmic choices such as speed-optimization practices (Renduchintala et al., 2021), byte-pair encoding (Gaido et al., 2021), or greedy decoding (Roberts et al., 2020) – although they may grant higher efficiency and performance – are actually disfavoring when it comes to gender bias. Finally, fine-grained analyses based on dedicated benchmarks have shown the limits of generic procedures and metrics to detect gender disparities (Vamvas and Sennrich, 2021; Renduchintala and Williams, 2021).

Such contributions are fundamental to shed light on gender bias, by providing guidance for interventions on data, procedures and algorithms. In this work, we contribute to this line of research by analysing direct ST systems (Bérard et al., 2016; Weiss et al., 2017a). As an emerging technology (Ansari et al., 2020; Bentivogli et al., 2021), we believe that prompt investigations have the potential to inform its future development, rather than keeping concerns over gender bias as an afterthought. In the wake of previous studies pointing out that *i)* ST systems may exploit audio cues to translate gender (Bentivogli et al., 2020), and *ii)* state-of-the-art BPE segmentation comes with a higher gender bias (Gaido et al., 2021), we conduct fine grained analyses on these systems, but by means of a new perspective: over the training process.

**Training and learning process.** Observing the learning dynamics of NLP models is not a new approach. It has been adopted for interpretability analysis to probe when and how linguistic capabilities emerge within language models (Saphra and Lopez, 2018, 2019), or inspect which features may be "harder" to learn (Swayamdipta et al., 2020).

With respect to analyses on a single snapshot, a diachronic perspective has the advantage of accounting for the evolution of NLP capabilities, making them more transparent based on trends' observation. Such an understanding can then be turned into actionable improvements. Accordingly, Voita et al. (2021) looked at the time-wise development of different linguistic abilities in MT, so to inform distillation practices and improve the performance of their systems. Additionally, the studies by Voita et al. (2019a,b) on the learning dynamics of extra-sentential phenomena highlighted how stopping criteria based on BLEU (Papineni et al., 2002) are unreliable for context-aware MT. Finally, Stadler et al. (2021) observed the evolution of different linguistic phenomena in system's output, noting how some of them seem to actually worsen across iterations.

Overall, as Stadler et al. (2021) noted, not much effort has been put into investigating how the training process evolves with regards to measurable factors of translation quality, such as linguistic criteria (grammar, syntax, semantics). We aim to fill this gap by evaluating gender translation of different ST systems at all training checkpoints.

## 4 Experimental Setting

### 4.1 Speech translation models

For our experiments, we rely on direct ST models built with two different target segmentation techniques: byte-pair encoding (BPE)[2] (Sennrich et al., 2016) and characters (CHAR). Since we are interested in keeping the effect of different word segmentations as the only variable, all our systems are built in the same fashion, with the same Transformer core technology (Vaswani et al., 2017) and within a controlled environment favouring progress analyses as transparent as possible. For this reason, we avoid additional procedures for boosting performance that could introduce noise, such as joint ST-ASR trainings (Weiss et al., 2017b; Bahar et al., 2019a) or knowledge distillation from MT models (Liu et al., 2019; Gaido et al., 2020a). Thus, our models are only trained on MuST-C (Cattoni et al., 2021), which currently represents the largest multilingual corpus available for ST. For the sake of reproducibility, details on the architecture and settings are provided in Appendix B.

**Training procedure.** As per standard procedure, the encoder of our ST systems is initialized with the weights of an automatic speech recognition (ASR) model (Bahar et al., 2019a) trained on MuST-C *audio-transcript* pairs. In our ST training, we use the MuST-C gender-balanced validation set (Gaido et al., 2020b)[3] to avoid rewarding systems' biased predictions. Each mini-batch consists of 8 samples, we set the update frequency to 8, train on 4 GPUs, so that a batch contains 256 samples. Within each iteration over the whole training set (i.e. epoch), we record 538 updates for en-es, 555 for en-fr, and 512 for en-it. Given the comparable number of updates across languages, as a point of reference we save the epoch checkpoint (herein ckp) that corresponds to a full pass on the whole training set.

All models reach their best ckp within 42 epochs, with a tendency of BPE to converge faster than CHAR. Specifically, they respectively stop improving after 33/42 epochs (en-es), 25/29 epochs (en-fr), and 29/32 epochs (en-it). As a stopping criterion, we finish our trainings when the loss on the validation set does not improve for 5 consecutive epochs. To inspect the stability of the best model results, our analysis also includes these additional 5 ckps.

### 4.2 Evaluation

**Test set and metrics.** To study the evolution of gender translation over the course of training and how it relates to generic perfomance, we employ the gender-sensitive MuST-SHE benchmark (Bentivogli et al., 2020) and its annotated extension

[2] Using SentencePiece (Kudo and Richardson, 2018).

[3] It consists of an equal number of TED talks data from masculine and feminine speakers: https://ict.fbk.eu/must-c-gender-dev-set/.

(Savoldi et al., 2022).[4] Consisting of instances of spoken language extracted from TED talks, MuST-SHE allows for the evaluation of gender translation phenomena[5] under natural conditions and for several informative dimensions:

· GENDER, which allows to distinguish results for Feminine (F) and Masculine (M) forms, thus revealing a potential gender gap.

· CATEGORY, which differentiates between: CAT1 first-person references to be translated according to the speakers' linguistic expression of gender (e.g. en: I am a *teacher*, es: soy *un profesor_* vs. soy *una profesora*); and CAT2 references that shall be translated in concordance with other gender information in the sentence (e.g. en: *she* is a *teacher*, es: es *una profesora*). These categories separate unambiguous from ambiguous cases, where ST may leverage speech information as an unwanted cue to translate gender.

· CLASS & POS, which allow to identify if gendered lexical items belonging to different parts-of-speech (POS) are equally impacted by bias. POS can be grouped into *open class* (verb, noun, descriptive adjective) and *closed* class words (article, pronoun, limiting adjective).

In MuST-SHE reference translations, each target gender-marked word is annotated with the above information.[6]Also, for each annotated gender-marked word, a corresponding wrong form, swapped in the opposite gender, is provided (e.g. en: ***the*** *girl* ***left***; it: ***la<il>*** *ragazza è* ***andata<andato>*** via). This feature enables pinpointed evaluations on gender realization by first computing[7] *i) Coverage*, i.e. the proportion of annotated words that are generated by the system (disregarding their gender), and on which gender realization is hence measurable, e.g. amigo (friend-M) → amig*; and then *ii) Accuracy*, i.e. the proportion of words generated in the correct gender among the measurable ones, e.g. amigo (friend-M) → amigo. Hence, *accuracy* properly measures model tendency to (over)generalize masculine forms over feminine ones: scores below 50% can signal a strong bias, where the wrong form is picked by the systems more often than the correct one.

In our study, we rely on the above metrics to inspect gender translations, and employ SacreBLEU (Post, 2018)[8] to measure overall translation quality.

**Setup.** Since we aim to observe the learning curves of our ST models, we evaluate both overall and gender translation quality after each epoch of their training process. As explained in Sec. 4.1, training includes also the 5 epochs that follow the best system ckp. To investigate systems' behaviour, we are particularly interested in the two following aspects of the learning curves: *i)* **training trend** (is gender accuracy raising across epochs, does it reach a plateau or can it actually worsen across iterations?); *ii)* **training stability** (is gender learning steady or erratic across epochs?)

Depending on the aspect addressed, we present results with different visualizations, reporting either the actual scores obtained at each ckp (more suitable to detect small fluctuations) or aggregated scores calculated with moving average over 3 ckp (more suitable to highlight general trends). Note that, since the total number of epochs differs for each system, to allow for a proper comparison we also plot results at different percentages of the training progress, where each progress point represents a 5% advancement (i.e 5%, 10%, 15% etc.).

With this in mind, we proceed in our analyses comparing overall performance across metrics (Sec.5.1), and inspecting feminine and gender translation (Sec. 5.2) at several levels of granularity (Sec 5.3 and 5.4). For any addressed aspect, we compare CHAR and BPE models across language pairs.

## 5 Results and Discussion

| | | BLEU | All-Cov | All-Acc | F-Acc | M-Acc |
|---|---|---|---|---|---|---|
| **en-es** | BPE | 27.4 | 64.0 | 66.0 | 49.0 | 80.7 |
| | CHAR | 27.2 | 64.0 | **70.5** | **58.9** | 80.5 |
| **en-fr** | BPE | 24.0 | 53.7 | 65.4 | 51.7 | 77.2 |
| | CHAR | 23.5 | 53.1 | **69.7** | **64.0** | 74.9 |
| **en-it** | BPE | 20.4 | 48.7 | 65.6 | 49.9 | 79.0 |
| | CHAR | 19.1 | 51.2 | **71.2** | **52.9** | 86.7 |

Table 1: BLEU, coverage and accuracy (percentage) scores computed on MuST-SHE.

First of all, in Table 1 we provide a snapshot of the results obtained by our ST models on their best ckp. As expected, the accuracy scores clearly exhibit a strong bias favouring masculine forms in translation (M-acc>F-acc), with feminine forms being generated with a probability close to a random guess for most systems. Moreover, these results

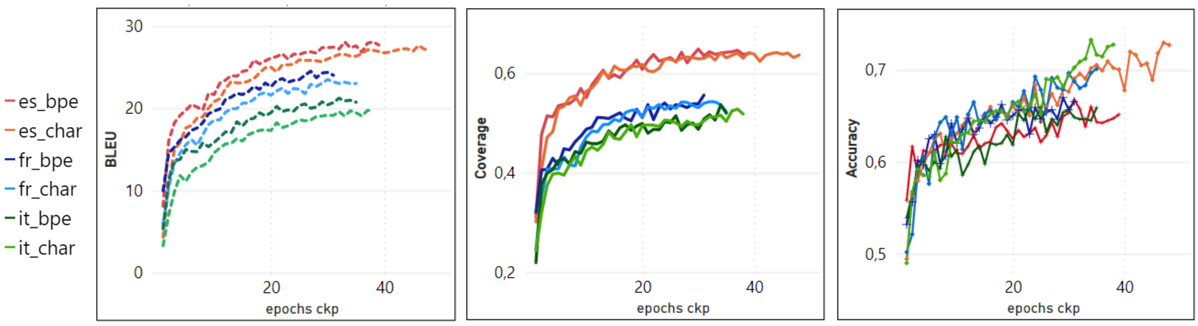
[4] Available at: https://ict.fbk.eu/must-she/

[5] Namely, the translation of a source neutral English word into a gender-marked one in the target languages, e.g. en: ***this*** *girl is* ***a good friend***, es: ***esta*** *chica es* ***una buena amiga***.

[6] Annotation statistics are provided in Appendix A.

[7] We rely on the evaluation script provided with the MuST-SHE extension.

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



(a) BLEU (b) Coverage (c) Accuracy

Figure 1: Results for every ckp of each model: BLEU (a), gender Coverage (b), and gender Accuracy (c).

are in line with the analyses by Gaido et al. (2021) and Savoldi et al. (2022) showing that CHAR has an edge in gender translation (All-Acc),[9] which is largely ascribed to better treatment of feminine gender.[10] Thus, we confirm a previously verified behaviour, which we now further inquiry in terms of its dynamic evolution.

### 5.1 Overall results

Here, we start by looking at the evolution of models' performance assessed in terms of BLEU, coverage, and accuracy (Figure 1) to inspect the time of emergence of the different capabilities captured by such metrics. For a bird's-eye view, we present the actual scores per each ckp.

**The evolution of both overall translation performance and gender translation is positive, but dissimilar in time and quality.** By looking at Figure 1, we observe that the gender accuracy learning curve (1c) immediately stands out. Indeed, the curves for both BLEU (1a) and gender coverage (1b) have a rapid and steady initial increase,[11] which starts to level off around the 20th ckp.[12] Also, the BLEU trends reveal a divide across models (BPE>CHAR) that remains visible over the whole course of training. In terms of coverage, the boundaries between types of models are more blurred, but correlate with BLEU scores for all language pairs. Conversely, by looking at the gender accuracy curves (1c) we asses that, while the overall trends show a general improvement across epochs, **gender learning *i)* proceeds with notable fluctuations, unlike the smoother BLEU and coverage curves; *ii)* emerges especially in the final iterations.** In particular, it is interesting to note that by epoch 30 (roughly 80% of the training process), *all* CHAR models handle gender translation better than *all* the BPE ones, regardless of the lower overall quality of the former group. Notably, the en-it CHAR system - with the lowest BLEU – exhibits the steepest increase in gender capabilities.

*Takeaways.* Generic translation quality improves more prominently in the initial training stages, while gender is learnt later. Thus, standard quality metrics conceal and are inadequate to consider gender refinements in the learning process.

### 5.2 Masculine and feminine gender

Moving onto a deeper level of analysis, we compare the learning dynamics that undergo Feminine (F) and Masculine (M) gender in terms of accuracy. To give better visibility of their *trends* and comparisons across models, in Figure 2 we plot the averaged results. As complementary view into training stability, Figure 3 shows the actual accuracy scores for the en-it models.[13]

**Masculine forms are largely and consistently acquired since the very first iterations.** As shown in Figure 2, masculine gender (M) is basically already learnt at 15% of the training process. Henceforth, its accuracy remains high and stable within 70-80%

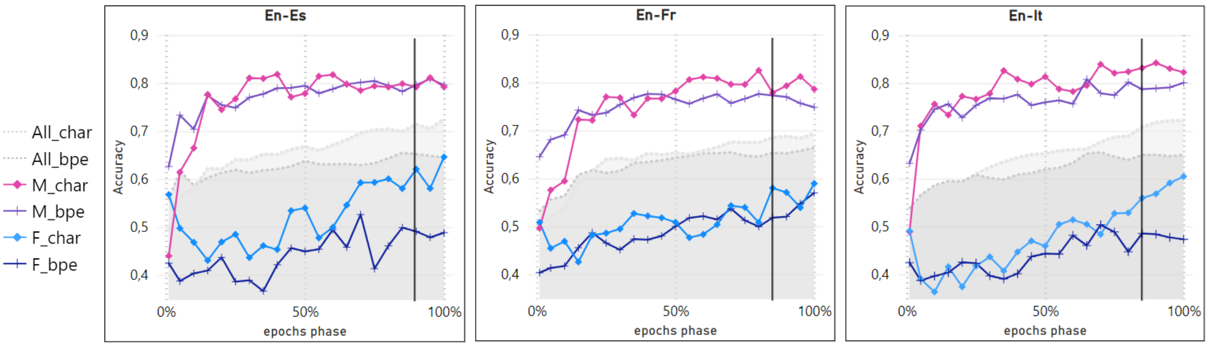
[9] Contemporary to our submission, Libovický et al. (2021) show that en-de MT systems based on character-level segmentation have an edge – with respect to BPE – in terms of gender accuracy on the WinoMT benchmark (Stanovsky et al., 2019). Their results, however, do not distinguish between feminine and masculine translation capabilities.

[10] For the sake of our analysis across epochs, we do not generate our final systems by averaging the 5 models around the best ckp as in Gaido et al. (2021) and Savoldi et al. (2022). For this reason, our systems compare less favourably in terms of BLEU score, also reducing the perfomance gap bewteen *de facto* standard BPE and CHAR.

[11] Computed as a binary task, gender accuracy starts at ∼50-55% in the first ckp. Such scores reflect that correct gender is assigned randomly at the beginning of the training process.

[12] The plateau is particularly visible for en-es CHAR due to its longer training.

[13] Due to space constraints, plots for all language pairs are in Appendix C - Fig. 7, which shows consistent results.



(a) en-es (b) en-fr (c) en-it

Figure 2: (F)eminine vs M(asculine) and over(all) accuracy scores for CHAR and BPE in en-es (2a), en-fr (2b), and en-it (2c). For better comparability across systems and trend visibility, results are shown at different percentages of the training progress (increasing by 5%), and scores at each progress point are calculated with moving average over 3 ckp. The first ckp (0%) is the actual score of the first epoch. The vertical line indicates the average score for the best ckp.

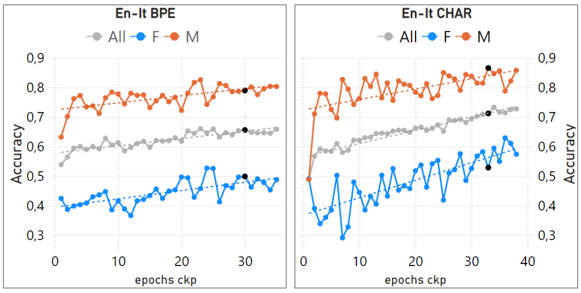


Figure 3: All, F vs Masculine gender accuracy for en-it BPE (left) and CHAR (right) models. Actual scores are reported per each ckp, and black dots indicate best ckp.

average scores for all models. As an exception, we notice a slightly decreasing trend in the iterations that follow the best ckp for en-fr BPE (2b). **Instead, feminine translation exhibits an overall upward trend that emerges later in the training process.** In Table 1, we already attested CHAR's advantage in dealing with feminine translation. Here, we are able to verify how such a capability is developed over the whole course of training. Specifically, CHAR gains a clear advantage over BPE in the last training phases, in particular for en-es (2a) and en-it (2c). Moreover, the overall rising F trend for CHAR models does not seem to dwindle: even after systems have reached their best ckp, feminine translation shows potential for further improvement.

**Unlike CHAR systems, BPE disproportionately favours masculine forms since the first ckp.** In the first ckp of the training, we notice an interesting difference between BPE and CHAR. Namely, the former models are biased since the very beginning of their training with an evident gender divide: ∼65% accuracy for M and only ∼40% for F forms.[14] Conversely, accuracy scores for both F and M forms in CHAR systems present about the same accuracy: both around 50% for en-it and en-fr, whereas the en-es model notably presents lower scores on the M set. From such behaviours, we infer that CHAR systems *i)* are initially less prone towards masculine generalisation, which is instead a by-product of further training; *ii)* promptly acquire the ability to generate both M and F inflections, although they initially assign them randomly. As we further discuss in Sec. 6, they occasionally acquire target morphology even before its lexicon, thus generating English source words inflected as per the morphological rules of the target language, e.g. en: *sister*; es: *sistera* (hermana). We regard this finding as evidence of the already attested capabilities of character-level segmentation to better handle morphology (Belinkov et al., 2020), which by extension may explain the higher capability of CHAR models at generating feminine forms.

**Despite a common upward trend, F and M gender curves progress with antiphase fluctuations.** In Figure 3, we see how this applies to CHAR in particular. Far from being monotonic, the progress of gender translation underlies a great level of instability with notable spikes and dips in antiphase for F and M - although eventually resulting in gains for F. Interestingly, it thus seems that systems become better at enhancing F translation by partially suppressing the representation of the other gender

[14] As outlined in Sec. 4.2, 40% accuracy for F means that in the remaining 60% of the cases systems generate a masculine inflection instead of the expected feminine one.

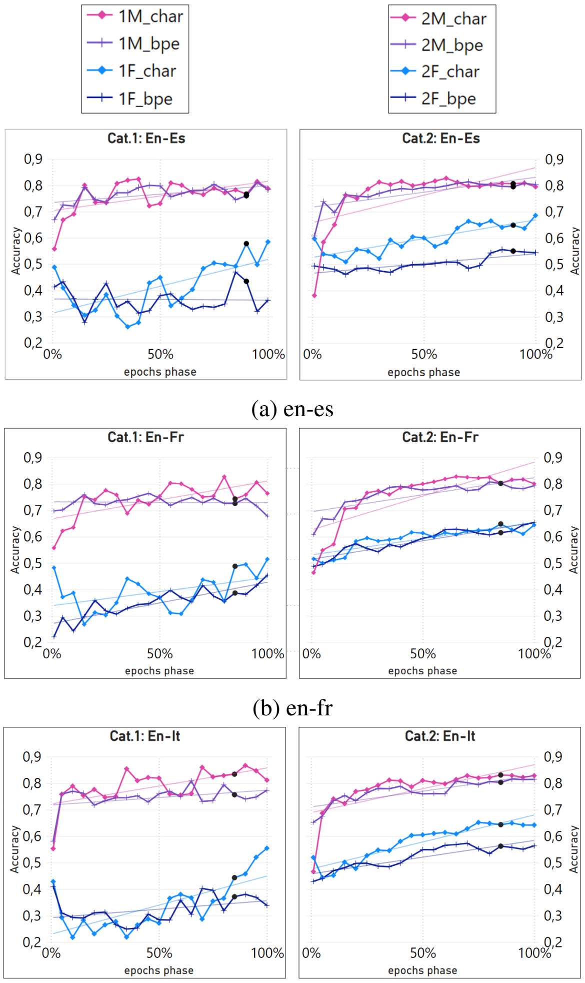
form.

*Takeaways.* The insights are more fine-grained: *i)* F is the actual gender form that is learnt late in the training process; *ii)* the progress of gender translation involves unstable antiphase fluctuations for F and M; *iii)* there is still room for improvements for F gender, especially for CHAR models. Overall, these findings make us question the suitability of standard metrics for diagnosing gender bias (see Sec. 5.1), and of the loss function as a stopping criterion. Along this line, previous work has foregrounded that even when a model has converged in terms of BLEU, it continues to improve for context-aware phenomena (Voita et al., 2019a). Hereby, although we find a good (inverse) correlation between loss and BLEU, we attest that they seem to be unable to properly account for gender bias and the evolution of feminine capabilities. Looking at both Figures 2 and 3, we question whether a longer training would have facilitated an improvement in gender translation and, in light of F and M antiphase relation, if it would lead to a suppression of M by favouring F. If that were the case, such type of diversity could be leveraged to create more representative models. Since more ckps would be needed to investigate this point, we leave it for future work.

### 5.3 Gender category

We now examine the learning curves for the translation of *i)* ambiguous references to the speaker, and *ii)* references disambiguated by a contextual cue (CAT1 and CAT2 introduced in Sec. 4.2). For each category, Figure 4 shows the comparison of feminine (1F, 2F) and masculine (1M, 2M) forms.

**Compared to the extremely unstable learning of CAT1, feminine and masculine curves from the unambiguous CAT2 exhibit a smooth upward parallel trend.** In Figure 4, the differences across categories fully emerge, and are consistent across languages and models. On the one hand, F and M curves from CAT2 show a steady trend which, despite a ∼10-20% accuracy gap across genders, suggests an increasing ability to model gender cues and translate accordingly. On the other hand, CAT1 proves to be largely responsible for the extreme instability and antiphase behaviour discussed for Figure 2, which is so strong to be evident even over the presented averaged scores.[15] Overall, we recognize a moderately increasing trend of 1F curves for all the CHAR models and the en-fr BPE. However, it barely raises above a random prediction, i.e. ∼50-57% accuracy meaning that a wrong masculine form is generated in ∼50-43% of the instances.



(a) en-es

(b) en-fr

(c) en-it

Figure 4: F vs M accuracy for CAT1 and CAT2. Scores are averaged over 3 ckps, and reported for each training phase. Dots indicate averaged scores for best ckp.

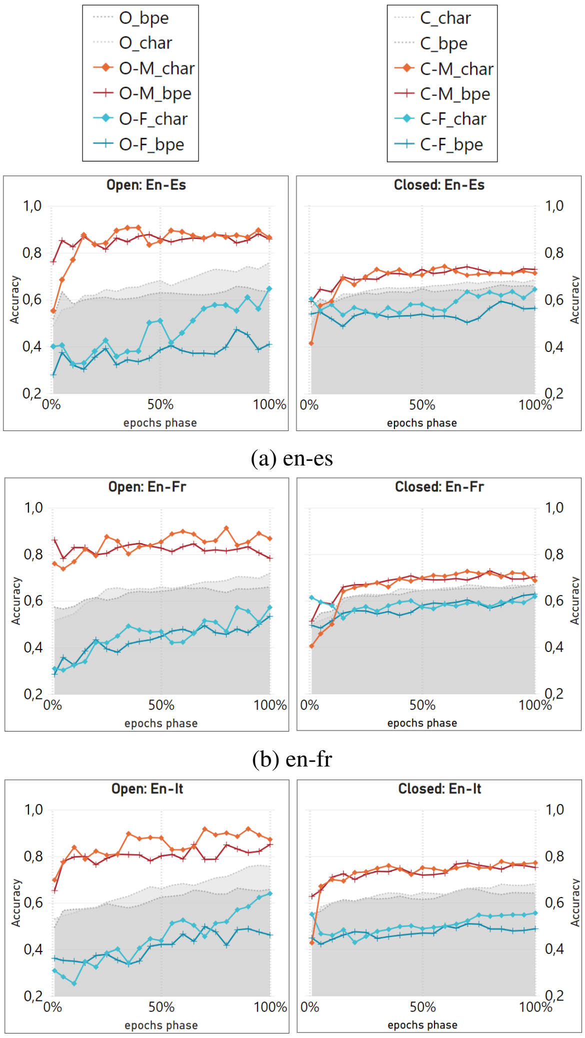
In light of the above, we are brought to reflect upon the hypothesis that direct ST models may use audio information to translate gender.[16] One possible explanation for systems' behaviour on CAT1 is that – although highly undesirable – ST *does* leverage speaker's voice as a gender cue, but finds the association "hard to learn". Another option is that ST *does not* leverage audio information and deals with CAT1 as gender ambiguous input. As a result, more biased BPE models more frequently opt for a masculine output in this scenario. CHAR models,

[15] E.g., the actual scores for 1F accuracy for en-it CHAR plummets as low as 11% at ckp9, and rockets at 60% at ckp36.

[16] This hypothesis was formulated in both (Bentivogli et al., 2020) and (Gaido et al., 2021).

instead, being characterized by a more favourable generation of feminine forms, progressively tend to converge towards a random gender prediction over the 1F set.

Towards the trustworthy development of ST technology, we call for future investigations on this point.



(a) en-es

(b) en-fr

(c) en-it

Figure 5: Open (left) vs Closed (right) classes accuracy scores per F and M gender. Scores are averaged over 3 ckps and reported for training phase.

### 5.4 Class and POS

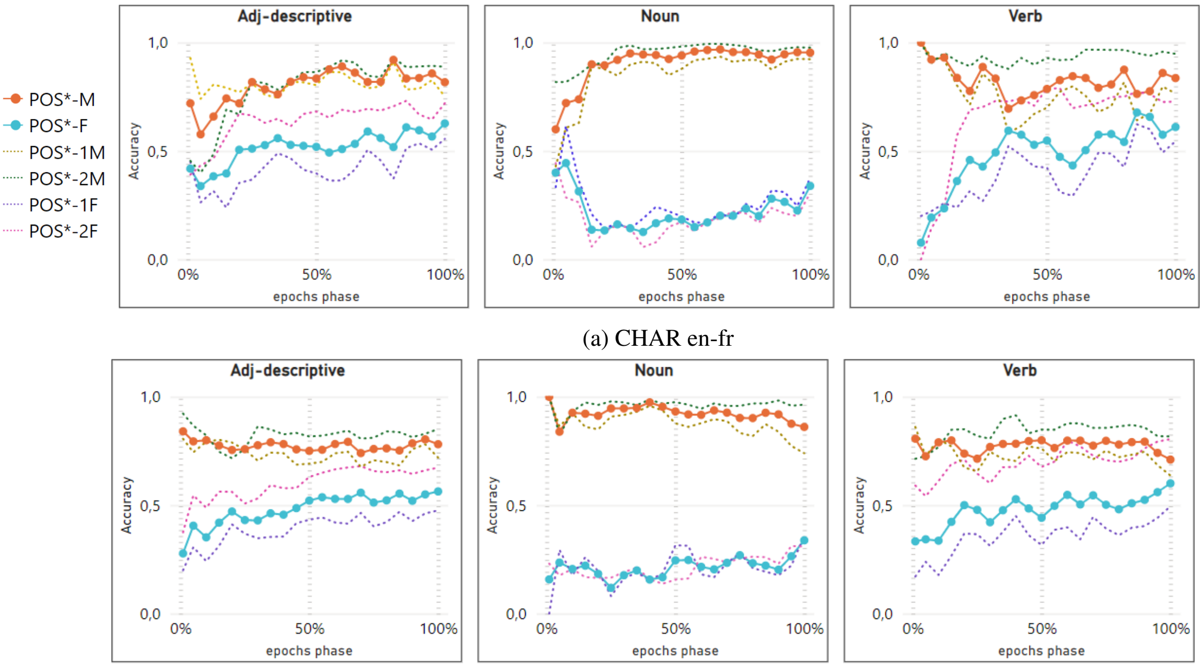
In Figure 5, we compare the gender curves for open and closed class words, which differ substantially in terms of frequency of use, variability and semantic content.

**Both F and M curves of the closed class change very little over the course of training.** In Figure 5, the closed class exhibits a stable trend with minimal increases, and a small F *vs.* M gap compared to the open class. We hypothesise this may be due to simple source constructions involving articles next to a gendered word, which are learnt since the very beginning (e.g., the mum; fr: la mère). **Open class words instead, show an unstable upward trend for F, opposed to the steady and early-learned M translation.** Consistently, the M curve starts off with unprecedented high scores (i.e., ∼80% accuracy within the first 20% of the training process) which further increases to 90% accuracy scores for CHAR. The F curve is progressively improving and – once again – with more significant gains late in training. This also implies that the M/F gap is reduced over the epochs. In light of the evident bias and distinct behaviour of F and M learning progress for the open class, we now turn to examine how each POS in this group evolves over training.

**Nouns are outliers, being the only POS that exhibits low variability in its learning curves, with little to almost no room for improving F translation.** Consistently across languages and models,[17] this claim can be verified in Figure 6 for en-fr. M nouns are basically fluctuation-free and reach almost perfect accuracy since the early ckps. Conversely, the F curve presents extremely low scores throughout the training process, signalling the strongest bias attested so far (i.e., the accuracy for F-nouns is 40% for both CHAR and BPE). Oddly enough, unlike adjectives and verbs, nouns learning dynamics do not even reflect the different trends assessed for CAT1 and the "easier" CAT2 (Sec. 5.3). Namely, despite the presence of a gender cue, the translation of feminine nouns from CAT2 (2F) does not benefit from such a disambiguation information. In fact, the accuracy for 2F nouns is basically on par (or even worse) with the performance of F nouns of CAT1 (1F), whereas for any other POS – and even M-nouns – the subset of CAT2 always exhibits a more positive learning trend.

*Takeaways*. Overall, our remarks are in line with the findings formulated by Savoldi et al. (2022): nouns emerge as the lexical category that is most impacted by gender bias, arguably because systems tend to rely more on stereotypical, spurious cues for the translation of professional nouns (e.g., *scientist*, *professor*). By examining their training progress however, we additionally unveil that *i)* biased associations influence noun translation more than unambiguous and relevant information, which

[17] Due to space constraints, we refer to Appendix C.2.1 for en-es and en-it.



(a) CHAR en-fr

(b) BPE en-fr

Figure 6: Accuracy per each open class POS for the en-fr CHAR and BPE models. The graph shows F vs M scores, also at the level of CAT1/2. Scores are provided for training phases, and calculated as the average over 3 ckps.

is available for CAT2; *ii)* ST models rely on such patterns so confidently that they never really adjust their trend over the epochs.

## 6 Qualitative Insights

We conclude our analysis with a manual inspection of the outputs of our ST systems at two *i)* initial, *ii)* middle, and *ii)* final ckps of their training process. To this aim, we opt for the en-es language pair – for which we observed the highest BLEU and gender coverage scores (Table 1) – to minimize the amount of low-quality translations that could be hard to analyze. Table 2 presents an example sentence from CAT2, translated by both CHAR and BPE, which backs up some of the quantitative observations formulated in Sec. 5. The source sentence contains neutral words (*older*, *a*, *a master*) occurring together with gender-marked words that disambiguate the correct gender (*sister*, *she*, *mother*). Given the presence of these gender cues, the neutral words should be fairly easy to translate.

In the first two ckps of both models, the output has very low quality.[18] It is characterized by extensive repetitions of frequent words, like *mother* (*madre* in A) or *young* (*joven*, *jóvenes* in B-G). Also, whereas functional words[19] are already appropriately employed and inflected with the correct gender (e.g. ***a*** *mother* → ***una*** *madre*, A,G), the noun *master* is not learnt yet and remains out of coverage; notably, BPE generates the word *hombre* (i.e. *man*) instead. Interestingly, if we also look at the gender cue noun *sister*, it maps to another kinship term *daughter* for BPE (G), whereas CHAR generalizes target morphology over English lexicon at these stages (*sistería* in A, *sistera* in B).

Such lexical issues are all refined by the middle ckps, where the systems have acquired both *sister/hermana* (with the feminine inflected adjective *antigua*[20] in C) and *master/maestr**, which in this case undergoes an interesting gender evolution across systems. For CHAR, we assist to an adjustment from masculine inflection (B), to a feminine one (C onwards) that remains stable until the end of training, even after the best ckp. Instead, BPE has a reversed trend. In H, the output *una maestra* reveals that this system can and has learnt to generate feminine forms. However, as the training progresses it switches to M inflections and never rebounds to the F ones. Rather, in the last epochs K-L it produces alternative synonyms, but always

[18] Still, we believe that it is to a certain degree intelligible thanks to the ASR initialization, see Sec. 4.1.

[19] I.e. closed class.

[20] The most fluent choice would be the neutral *mayor*. Here, however, we just focus on gender realizations.

| | | |
|---|---|---|
| **2F** | SRC | My **older** sister Claire, she became **a** young mother, and **a master** at getting things done |
| CHAR$_{Es}$ | (A) INI | Mi madre la *sistería sistencia* , encontró **una** madre joven **una** madre, y **una** madre, y una masa, y una cosa que estaba cosas. |
| | (B) INI | Mi madre *sistera* de clara se convirtió en un joven jóvenes jóvenes jóvenes, y **un maestro** de cosas que hicieron. |
| | (C) MID | Mi hermana más **antigua** clara, se convirtió en **una** madre joven y **una maestra** que hacía cosas. |
| | (D) MID | Mi **antigua** hermana Clare, se convirtió en **una** madre joven y **una maestra** que hice las cosas. |
| | (E) FIN | Mi hermana mayor Clare, se convirtió en **una** madre joven, y **una maestra** de hacer las cosas. |
| | (F) FIN | Mi hermana mayor, Clare, se convirtió en **una** joven madre, y **una maestra** por hacer las cosas. |
| BPE$_{Es}$ | (G) INI | Mi hija de la Tierra se convirtió en **un** joven joven, ella me convertí en **una** madre, y **un** hombre que hizo cosas. |
| | (H) INI | Mi hermana mayor claridad se convirtió en **una** madre joven, y **una maestra**, lo hice. |
| | (I) MID | Mi hermana mayor se volvió a ser **una** joven madre, y una maestría que hice. |
| | (J) MID | Mi hermana mayor declaró, se convirtió en **una** madre joven, y **un maestro** que se está haciendo. |
| | (K) FIN | Mi hermana mayor declaró que se convirtió en **una** madre joven, y **un amo** logrando hacer las cosas. |
| | (L) FIN | Mi hermana mayor Clare se convirtió en **una** madre joven, y **un dueño** de hacer que las cosas se hicieran. |

Table 2: En-es outputs at initial, middle, and final epochs. The source sentence contains **neutral words** to be translated according to the available gender cues. In the outputs, we indicate correct feminine gender translation vs masculine. We also signal repetitions and *copied source lemma*+target morphology combinations.

with the wrong gender (*amo, dueño*).[21] Overall, this case exemplifies how, in spite of *i)* having F morphological capabilities, and *ii)* the presence of a close cue disambiguating gender, the BPE system confidently relies on spurious and irrelevant patterns for gender translation.

## 7 Limitations and Future Work

In this work, we rendered the ST training process less opaque by analyzing the learning process of gender. To do so, we looked into ST outputs. However, a complementary perspective would be to rely on explainability and probing approaches on system's inner mechanisms (Belinkov and Glass, 2019) and verify their compatibility with our findings. Also, a contrastive comparison of the learning curves for gender and other linguistic phenomena implying a one-to-many mapping (e.g. politeness *you* → es: *tu*/*usted*) could pinpoint learning trends which are specific to gender bias. A limit of our analyses is that they include only 5 epochs after their best validation loss. In light of our *a posteriori* finding that F gender – especially for CHAR – does not reach a plateau in the last epochs, future work is needed to confirm whether and to what extent F learning keeps improving. This could inform studies on *i)* how to leverage diversified output to alleviate gender bias in our models, *ii)* gender-sensitive stopping criteria. Finally, we point out that for the most fine-grained level of analyses (Sec. 5.4), our evaluation is based on very specific subsets (e.g. nouns broke down into 1F, 2F, 1M, 2M).[22] This comes with an inherent reduction of the amount of measurable gender-marked words, which could in turn imply noise and additional instability in the visualized results. However, we reduce this risk by presenting them averaged over 3 ckps and, as the noun curves show (Fig. 6), believe in their validity for comparisons within the same dimension and level of granularity.

Note that our study lies on the specificity of three comparable grammatical gender languages. We are thus cautious about generalizing our findings. Experiments on other training sets and language pairs are currently hindered by the lack of an available natural, gender-sensitive ST benchmark that covers alternative gender directions. While bearing this in mind, we however underscore that the conditions of gender translation significantly change depending on the features of the accounted languages and direction (e.g. translating from grammatical gender languages to English and not *vice versa*). Thus, gender phenomena on typologically different gendered languages would not be *directly* comparable and compatible with the presented analyses. Rather than a specific limitation of our setting, we regard this as an intrinsic condition.

## 8 Conclusion

Despite the mounting evidence of biased behaviour in language technologies, its understanding is hindered by the complex and opaque nature of current neural approaches. In this work, we shed light on the emergence of gender bias in ST systems by focusing on their learning dynamics over training. In this way, we adopt a new perspective that accounts for the time-wise appearance of gender capabilities, and examine their stability, reliability and course of development. For three language pairs (en → es, fr, it) we inspect the learning curves of feminine

[21] Although inappropriate in this context, both *amo* and *dueño* are valid mapping to the word *master*).

[22] In the Appendix, we provide MuST-SHE statistics (Table 4) and gender coverage for open-closed class words (Fig. 8).

and masculine gender translation *i)* at several levels of granularity; *ii)* with respect to progress in terms of overall translation quality; *iii)* on the output of ST systems trained on target data segmented as either character or sub-words units (BPE). In our diachronic analysis, we unveil that *i)* feminine gender is learnt late over the course of training, *ii)* it never reaches a plateau within the number of iterations required for model convergence at training time, and *iii)* its refinements are concealed by standard evaluation metrics. Also, by looking at the stability vs. fluctuations of the explored trends, we identify under which circumstances ST models seem to actually progressively acquire feminine and masculine translation, and when instead their erratic, antiphase behavior reflects unreliable choices made by the systems. In this way, we find that *nouns* – the lexical category most impacted by gender bias – present a firm and huge gender divide over the whole training, where ST systems do not rely on relevant information to support feminine translation and never really adjust its generation.

## 9 Impact Statement[23]

In compliance with ACL norms of ethics,[24] we hereby clarify i) the characteristics of the dataset used in our experiments, and ii) our use of gender as a variable (Larson, 2017).

As already discussed, in our experiments we rely on the training data from the TED-based MuST-C corpus[25] (Sec. 4.1), and its derived evaluation benchmark, MuST-SHE v1.2[26] (Sec. 4.2). For both resources, detailed information on the representativeness of TED data is available in their data statements (Bender and Friedman, 2018). As regards gender, it is largely discussed how it is intended and annotated. Thus, we know that MuST-C training data are manually annotated with speakers' gender information[27] based on the personal pronouns found in their publicly available personal TED profile.[28] Overall, MuST-C exhibits a gender imbalance, with 70% vs. 30% of the speakers referred by means of *he/she* pronoun, respectively.[29]

As reported in its release page,[30] 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. Accordingly, when working on the evaluation of speaker-related gender translation for MuST-SHE,[31] we solely focus on the rendering of their reported linguistic gender expressions. 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.

Finally, in our diagnosis of gender bias we only account for feminine and masculine linguistic forms, which are those traditionally in use and the only represented in the used data. However, we stress that – by working on binary forms – we do not imply or impose a binary vision on the extra-linguistic reality of gender, which is rather a spectrum (D'Ignazio and Klein, 2020). Also, we acknowledge the challenges faced for grammatical gender languages like Spanish, French and Italian in fully implementing neutral language, and support rise of neutral language and non-binary neomorphology (Shroy, 2016; Gabriel et al., 2018; Conrod, 2020).

## 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.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Trans-*

[23] Extra page granted as per https://aclrollingreview.org/cfp.

[24] https://www.aclweb.org/portal/content/acl-code-ethics

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

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

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

[28] https://www.ted.com/speakers.

[29] Only one *They* speaker is represented in the corpus.

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

[31] Category 1 (CAT1) in the corpus.

*lation*, pages 1–34, Online. Association for Computational Linguistics.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019a. A Comparative Study on End-to-end Speech to Text Translation. In *Proceedings of International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 792–799, Sentosa, Singapore.

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

Christine Basta and Marta R. Costa-jussà. 2021. Impact of Gender Debiased Word Embeddings in Language Modeling. *CoRR*, abs/2105.00908.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

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.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

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.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*.

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.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

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 Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

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.

Kirby Conrod. 2020. Pronouns and gender in language. *The Oxford Handbook of Language and Sexuality*.

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

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2020. Evaluating gender bias in speech translation. *CoRR*, abs/2010.14465. Accepted at LREC 2022.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Interpreting Gender Bias in Neural Machine Translation: The Multilingual Architecture Matters. *Accepted in 36th AAAI Conference on Artificial Intelligence*.

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

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

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.

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, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2020a. End-to-End Speech-Translation with Knowledge Distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88, Online. Association for Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020b. 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.

GLAAD. 2007. Media Reference Guide - Transgender.

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.

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

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. 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.

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.

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.

Jindřich Libovickỳ, Helmut Schmid, and Alexander Fraser. 2021. Why don't people use character-level machine translation? *arXiv preprint arXiv:2110.08191*.

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.

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.

Parmy Olson. 2018. The Algorithm That Helped Google Translate Become Sexist. Accessed: 2021-02-25.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Tomasz Potapczyk and Pawel Przybysz. 2020. SRPOL'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.

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.

Naomi Saphra and Adam Lopez. 2018. Language Models Learn POS First. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 328–330, Brussels, Belgium. Association for Computational Linguistics.

Naomi Saphra and Adam Lopez. 2019. Understanding Learning Dynamics Of Language Models with SVCCA. 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 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation. *ArXiv e-prints arXiv:2203.09866. Accepted at ACL 2022*.

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.

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

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Patrick Stadler, Vivien Macketanz, and Eleftherios Avramidis. 2021. Observing the Learning Curve of NMT Systems With Regard to Linguistic Phenomena. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 186–196, 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.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing.

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.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

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.

Jonas-Dario Troles and Ute Schmid. 2021. Extending Challenge Sets to Uncover Gender Bias in Machine Translation: Impact of Stereotypical Verbs and Adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in mt: A case study of distilled bias. In *2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265. 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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-Aware Monolingual Repair for Neural Machine Translation. 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 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the Source and Target Contributions to Predictions 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 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017a. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. 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 *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*.

## A MuST-SHE statistics

Table 4 shows the word-level annotation statistics of MuST-SHE v1.2 (Bentivogli et al., 2020) and ist annotated extension (Savoldi et al., 2022). The amount of gender-marked words is balanced across *i)* languages, *ii)* Feminine and Masculine gender forms, *iii)* Categories. The Open/Closed Class and POS distribution vary in light of the gender marking features of the accounted languages.

## B Model Settings

To create the models used in our experiments, we exploited the open source code publicly available at: `https://github.com/`

|  | en-es | en-fr | en-it |
|---|---|---|---|
| BPE | 8,120 | 8,048 | 8,064 |
| Char | 464 | 304 | 256 |

Table 3: Sizes of model dictionaries.

mgaido91/FBK-fairseq-ST. In accordance with (Potapczyk and Przybysz, 2020), our models have 2 3x3 convolutional layers with 64 filters that reduce the input sequence length by a factor of 4, followed by 11 Transformer encoder layers and 4 Transformer decoder layers. We add a logarithmic distance penalty (Di Gangi et al., 2019) to the encoder self-attention layers. As loss function we adopt the label smoothed cross-entropy (Szegedy et al., 2016) with 0.1 as smoothing factor. Our optimizer is Adam using $\beta_1$=0.9, $\beta_2$=0.98, and the learning rate decays with the inverse square root policy, after increasing for the initial 4.000 updates up to $5 \times 10^{-3}$. The dropout is set to 0.2, and to further regularize the training we use as data augmentation technique SpecAugment (Park et al., 2019; Bahar et al., 2019b) with probability 0.5, two bands on the frequency dimension, two on the time dimension, 13 as maximum mask length, and 20 as maximum mask length.

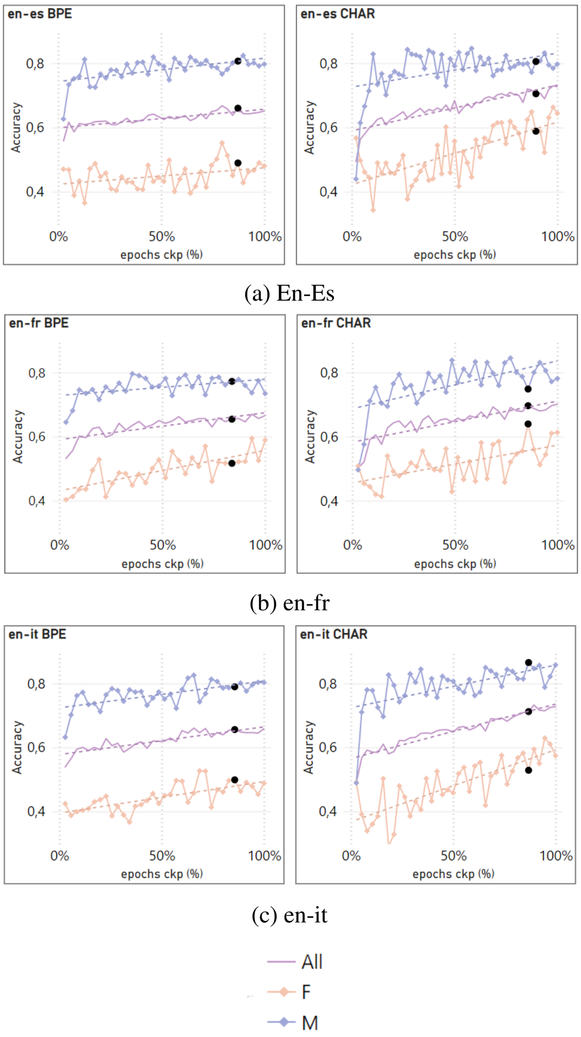
We extract 40 features with 25ms windows and 10ms slides using XNMT[32] (Neubig et al., 2018), after filtering utterances longer than 20s to avoid excessive memory requirements at training time. The resulting features are normalized per-speaker.

We rely on the MuST-C corpus (Cattoni et al., 2021) for training: it contains 504 hours of speech for en-es, 492 for en-fr, and and 465 for en-it, thus offering a comparable amount of data for our three language pairs of interest.

The target text is tokenized with Moses[33] and then segmented. When using BPE, we set the number of merge rules to 8,000, which – following Di Gangi et al. (2020) – results in the most favouring ST performance. The size of the resulting dictionaries is reported in Table 3.

## C Additional visualizations

In this section, we provide additional plots that – due to space constrains – were not inserted in the discussion of the results in Section 5.



(a) En-Es

(b) en-fr

(c) en-it

Figure 7: overAll, Feminine vs Masculine actual accuracy scores per each ckp of BPE and CHAR: en-es (7a), en-fr (7a), en-it (7c). Black dots indicate the best ckp.

### C.1 Feminine and Masculine forms

In Figure 7, we show Feminine vs Masculine gender accuracy actual scores for en-es (7a), en-fr (7b), en-it (7c) for each ckp. As the plots show, gender accuracy scores exhibit a more positive and steeper trend for CHAR, which is however characterized by higher levels of instability. For all models, we can see – to different degrees – the antiphase relation between F and M curves.

### C.2 Open and Closed Class

Figure 8 shows coverage scores over the training progress for words from the open (O) and closed (C) class. As expected, the coverage of functional words is extremely high, firmly maintained over the whole course of training. For the more variable words from the O class, instead, we attest upwards

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

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

| | | En-Es | | | | En-Fr | | | | En-It | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1977 | | | | 1823 | | | | 1942 | | | |
| | | **F** | | **M** | | **F** | | **M** | | **F** | | **M** | |
| | | 950 | | 1027 | | 898 | | 925 | | 898 | | 1044 | |
| | | **1F** | **2F** | **1M** | **2M** | **1F** | **2F** | **1M** | **2M** | **1F** | **2F** | **1M** | **2M** |
| | | 392 | 558 | 419 | 608 | 424 | 474 | 410 | 515 | 401 | 497 | 415 | 629 |
| **Open** | *noun* | 121 | 106 | 151 | 185 | 58 | 62 | 75 | 112 | 48 | 62 | 71 | 138 |
| | *adj-des* | 191 | 190 | 139 | 141 | 177 | 153 | 129 | 107 | 118 | 119 | 92 | 110 |
| | *verb* | 19 | 36 | 12 | 37 | 156 | 90 | 141 | 105 | 178 | 133 | 176 | 129 |
| **Closed** | *article* | 35 | 147 | 75 | 193 | 29 | 89 | 61 | 119 | 41 | 105 | 59 | 177 |
| | *pronoun* | 5 | 33 | 26 | 23 | 1 | 28 | 3 | 25 | 3 | 20 | 6 | 17 |
| | *adj-det* | 21 | 46 | 16 | 29 | 3 | 52 | 1 | 47 | 13 | 58 | 11 | 58 |

Table 4: Word-level statistics for all MuST-SHE dimensions on each language pairs: *i)* Feminine and Masculine gender forms, *ii)* Categories 1 and 2, *iii)* Open/Closed Class and POS.

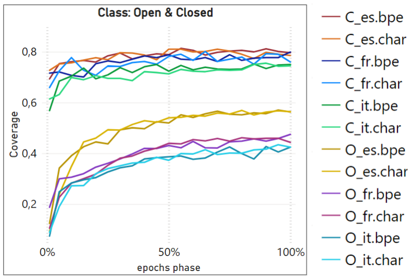
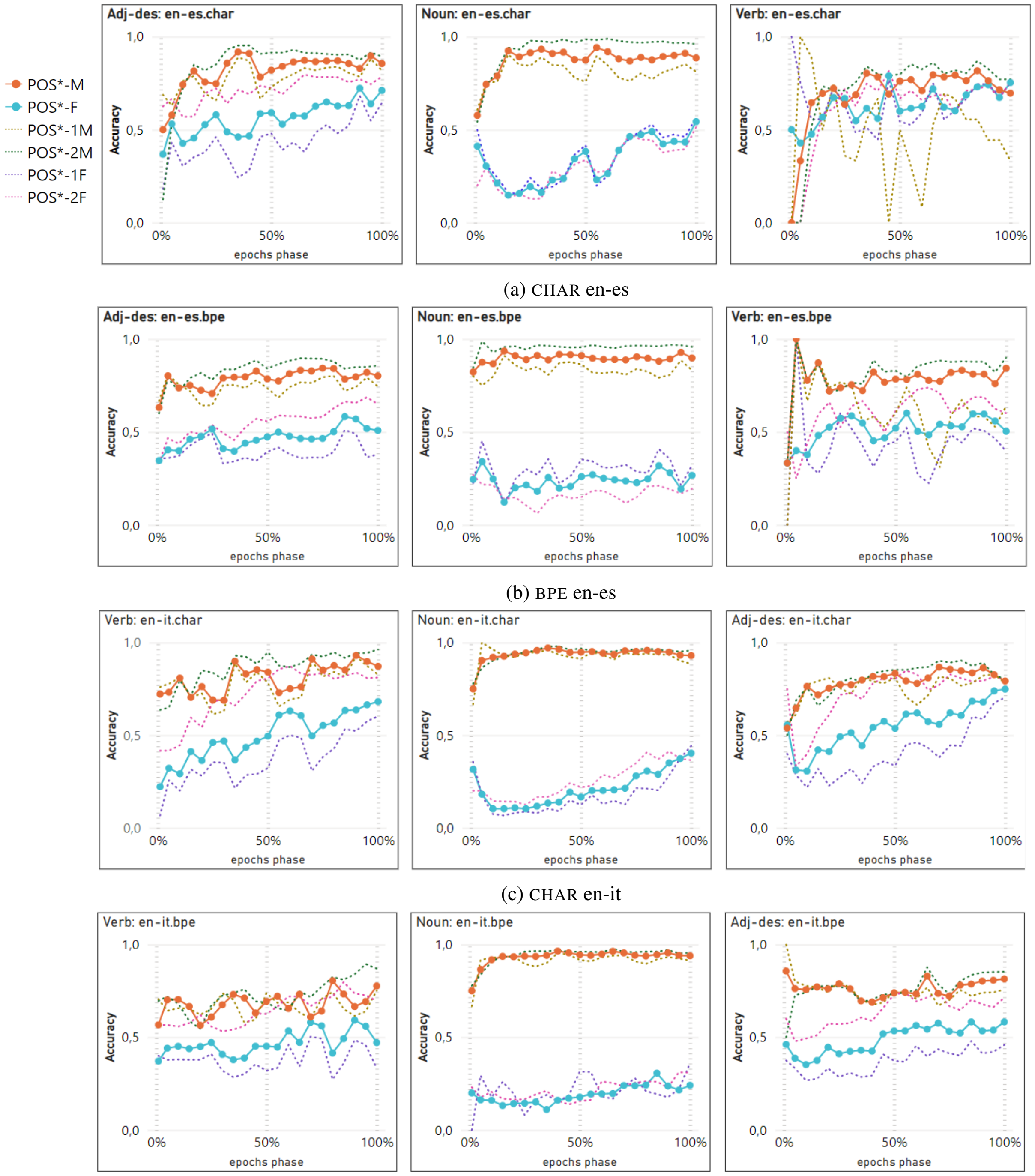


Figure 8: Coverage actual scores for Closed and Open class POS of CHAR and BPE models for all language pairs over percentages of training progress.

trend, which start to reach a plateau in the second half of the training progress. However, it never exceeds ∼58% coverage scores.

### C.2.1 Open Class POS

Figure 9 shows Feminine and Masculine learning curves for en-es and en-it for each of the POS within the Open class: *i)* nouns, *ii)* verbs, and *iii)* descriptive adjectives. Also, we visualized their trend within the subset of CAT1 and CAT2 of each POS. Overall, also for these language pairs we see how *nouns* are outliers: their feminine learning curve exhibits little to no real improvement. The CHAR model for en-es represents a partial exception given that F learning curves shows a steeper upward trend: still, it remains close to only 50% accuracy. Also, the evolution of F nouns from the ambiguous CAT1 and CAT2 (non ambiguous) is basically on par, thus confirming that models do not rely on relevant gender information to adjust the feminine generation of nouns over their training.



(a) CHAR en-es

(b) BPE en-es

(c) CHAR en-it

(d) BPE en-it

Figure 9: Accuracy per each open class POS for en-es (9a. 9b) and en-it (9c, 9d) CHAR and BPE models. The graph shows F vs M scores, also at the level of CAT1/2. Scpres are provided for training phases, and calculated as the average over 3 ckps.