Unmasking Biases: Exploring Gender Bias in English-Catalan Machine Translation through Tokenization Analysis and Novel Dataset

Audrey Mash, Carlos Escolano, Aleix Sant, Francesca De Luca Fornaciari, Maite Melero
Barcelona Supercomputing Centre
Barcelona, Spain
{audrey.mash, carlos.escolano, aleix.santsavall, fdelucaf, maite.melero}@bsc.es

Abstract

This paper presents a comprehensive evaluation of gender bias in English-Catalan machine translation, encompassing the creation of a novel language resource and an analysis of translation quality across four different tokenization models. The study introduces a new dataset derived from the MuST-SHE corpus, focusing on gender-neutral terms that necessitate gendered translations in Catalan. The results reveal noteworthy gender bias across all translation models, with a consistent preference for masculine forms. Notably, the study finds that when context is available, BPE and Sentencepiece Unigram tokenization methods outperform others, achieving higher accuracy in gender translation. However, when no context is provided, Morfessor outputs more feminine forms than other tokenization methods, albeit still a small percentage. The study also reflects that stereotypes present in the data are amplified in the translation output. Ultimately, this work serves as a valuable resource for addressing and mitigating gender bias in machine translation, emphasizing the need for improved awareness and sensitivity to gender issues in natural language processing applications.

Keywords: gender bias, machine translation, subword tokenisation

1. Introduction

As with other rapidly developing language technologies, the growing integration of machine translation tools into our daily lives is prompting an increasing awareness of the unexpected impacts they may bring. Biases present in data and learned by Machine Translation (MT) models can perpetuate stereotypes, reinforce existing inequalities, and lead to the invisibility or under-representation of certain human groups. The most visible form of bias in translation is gender bias, and work has shown that masculine forms are over represented in the output of machine translation systems, with the exception of stereotypically feminine roles and activities.

While much existing work takes the input data as the starting point for dealing with gender bias, e.g. trying to make it more balanced, this paper delves into the impact which tokenization methods can have on gender generation by MT models. We train four MT systems using different tokenization techniques – Character-Based tokenizByte Pair Encoding (BPE), Sentencepiece Unigram and Morphological tokenization – and assess the effect of tokenization on translation quality and gender output.

Additionally, our study contributes to ongoing research into gender bias in translation by presenting a novel dataset for evaluating gendered output in Catalan. We derived this dataset from the English-Spanish MuST-SHE, curating an English-Catalan corpus with gender-neutral English terms necessitating gendered translations in Catalan. This dataset allows for an extensive examination of gender bias in machine translation. Our research aims to enrich the field of computational linguistics by advancing gender-inclusive translation techniques.

2. Background

Accurately translating gender while avoiding bias is a complex task, primarily due to the disconnection between social and linguistic gender categories (Stanczak and Augenstein, 2021) and the diverse ways languages mark gender. In linguistics, gender refers to a noun class governing agreement in a noun phrase (McConnell-Ginet, 2013). Notional gender languages such as English exhibit limited agreement, often explicitly tied to sex and only present on a small subset of word classes such as pronouns, while grammatical gender languages, such as Catalan, feature morphosyntactic gender across multiple word classes with a less direct link to sex (McConnell-Ginet, 2013).

These distinctions pose challenges when translating from a notional to a grammatical gender language. Human translators are able to make use of a broad context window and possible external knowledge, but even the increased context available to recent NMT systems is necessarily limited. When information is insufficient, MT systems tend to produce the statistically most likely gender inflection, potentially reinforcing stereotypes and biases

LREC-COLING 2024, pages 17144–17153
20-25 May, 2024.

(Vanmassenhove et al., 2018). Even where sufficient information is available to the model, studies have shown a tendency to over-ascribe stereotypical gender markers to nouns and pronouns, for example by assigning masculinity to doctors and engineers and femininity to nurses, and running the risk of not only replicating but also amplifying the biases found in the training data (Bolukbasi et al., 2016; Zhao et al., 2017).

Savoldi (2021) categorize the effects of the presence of gender bias in the output of machine translation into representational and allocational harms. The former diminishes the representation of social groups, perpetuating stereotypes. Allocational harms relate to resource allocation, affecting the quality of services provided to less visible groups. Such harms manifest in education, healthcare access, legal outcomes, and social inclusion (Savoldi, 2021). Machine translation tools which produce predominantly masculine forms can lead to women and non-binary people needing to invest more time and energy into revising the output.

Much work which has been done to date on mitigating gender bias in MT has begun from considering pre-existing bias in the data on which the model is trained as the primary source of the imbalance. Gender-tagging is one approach which may yield positive results (Elaraby et al., 2018; Stafanovičs et al., 2020; Vanmassenhove et al., 2019). In this method the training data is automatically annotated with gender information, either at sentence level (Elaraby et al., 2018; Vanmassenhove et al., 2019) or at word level (Stafanovičs et al., 2020).

Vanmassenhove et al. (2019) used the speaker information provided with the Europarl corpus to add tags to sentences containing 1st person singular references. They tested on translation from English to 10 languages, five of which have morphological gender agreement and five of which do not. They saw improvements in BLEU scores compared to their baseline models for translation in 4 of the 5 languages which have morphological gender agreement, with the exception of Spanish, and only 1 of the 5 (Danish) which does not have morphological gender agreement. However, they did not use any more gender-specific metric than BLEU scores and there is a lack of manual analysis to determine the causes of the increase.

Similarly, in the field of Speech Translation, Elaraby et al. (2018) used POS tagging and language specific rules to gender tag both speakers and listeners in a subset of the Open Subtitles English-Arabic corpus and saw improvements in both gender accuracy and BLEU score. At word level, Stafanovičs et al. (2020) extract gender information from the target side of a parallel corpus and use statistical alignments to project this back to the source side as tags for the training data. They saw an improvement in BLEU scores across all language pairs, as well as better performance on the WinoMT evaluation set.

Gender tagging approaches seem to improve both gender accuracy and overall translation accuracy in most cases of translation from a language without morphological gender to one with it. However, it requires a substantial effort to produce suitable training data, as well as necessitating an increased computational cost.

Escudé Font and Costa-jussà (2019) attempted to debias the word embeddings learned from the data using the hard debiasing algorithm developed by Bolukbasi et al. (2016) and a gender neutral update from Zhao et al. (2017). These methods aim to enforce neutrality in specific dimensions of the embeddings which capture the gender direction. This method shows a very slight increase in BLEU score using the gender neutral approach, but it is too small to be significant. They report improved performance in gender accuracy, but at a high computational cost and with limited improvement in overall translation quality.

The works previously discussed all involve training models from scratch. Costa-jussà and de Jorge (2020) fine-tuned a pre-trained model on a smaller gender balanced dataset filtered from Wikipedia and found that fine-tuning with a mix of balanced and original training data was able to both reduce gender bias and improve the BLEU score.

While considering ways to augment or balance the data is common, a less explored approach involves architectural choices in model design and the technical bias which ensues from them. Costa-jussà et al. (2020) trained multilingual MT models with both Shared and Language-Specific Encoder-Decoders and found that the Language-Specific encoder-decoders exhibit less gender bias than the Shared encoder-decoder architecture while also achieving better BLEU scores.

Finally, Gaido et al. (2020) looked at the impact of various types of sub-word tokenisation on gender bias in end-to-end Speech Translation (ST), conducting a study to examine the effects of BPE, Dynamic Programming Encoding (DPE), Character Segmentation, Morfessor, and Linguistically Motivated Vocabulary Reduction (LMVR), on gender bias. The study involved training models with each tokenization method and evaluating the results in terms of overall translation quality (BLEU scores) and the correct generation of gender forms. The findings indicated that BPE, DPE, and LMVR performed similarly in terms of BLEU scores, but due to the computational cost, BPE was considered the best segmentation strategy. However, Character Segmentation had the lowest BLEU scores while performing the best in terms of gender accuracy.

As Gaido et al. (2020) study focused on end-to-

end speech translation, the inherent differences between ST and MT mean that it cannot be assumed that similar results would be obtained in the field of MT. Audio data in ST contains clues about speaker characteristics, such as pitch, intonation, and speech patterns, which may provide indirect indicators of the speaker's gender and emotions.

Considering the alignment between characters and phonemes, character-based tokenization is expected to perform better in ST compared to text based MT. However, for MT, character segmentation is limited due to the lack of semantic information conveyed by characters in different contexts. Therefore, it is unlikely that the best tokenization method for addressing gender bias in MT aligns with the findings of Gaido et al. (2020).

3. Experiments

3.1. Tokenization Methods

This section introduces the four tokenization approaches considered in our experiments: Character-Based, Byte Pair Encoding (BPE), Sentencepiece Unigram, and Morphological (Morfessor). BPE and Sentencepiece Unigram methods were selected due to their status as current State of the Art methods in NMT, while Character Based and Morphological methods were selected so as to compare their performance here with their performance in ST (Gaido et al., 2021).

3.1.1. Character Based

Character-based tokenization offers an effective solution to the limitations of word and sub-word tokenization, which often lead to out-of-vocabulary tokens, especially for rare words (Libovický et al., 2022). In character-based tokenization, the vocabulary size is determined by the number of characters, ensuring complete coverage without out-of-vocabulary tokens. It also preserves orthographic information, capturing character-level relations. This is particularly advantageous for morphologically rich languages, as it may preserve grammatical features like gender and tense. Additionally, character-based systems handle source-side noise well, making them robust against spelling variations and typos.

However, character-based tokenization comes with higher computational costs, demanding more training time, memory, and computational resources. Each character contains less semantic information than larger sub-word tokens. While character-level representations are valuable in decoder-based models, they have been less successful in encoder-based models (Libovický et al., 2022).

3.1.2. Byte Pair Encoding

Byte Pair Encoding (BPE) resolves vocabulary limitations by starting with a character-level vocabulary and iteratively merging frequent character pairs or sequences. This approach combines character and sequence-level representations, making it more efficient and less computationally demanding than character-based tokenization. BPE can handle any input text, even previously unseen words, by breaking them down into sub-word units. However, BPE has limitations. It may tokenize the same sequence in different ways, treating them as distinct inputs. Additionally, BPE-generated sub-word units do not always correspond to linguistically meaningful units, potentially resulting in the loss of semantic information embedded in morphology.

3.1.3. Sentencepiece Unigram

Sentencepiece Unigram tokenization is used in combination with subword regularisation to enhance robustness against noise and segmentation errors, aiming to address BPE's limitations (Kudo and Richardson, 2018). Sentencepiece Unigram considers the frequency of individual characters and character sequences, determining their probabilities in the training data. It begins with a large seed vocabulary, composed of potential segmentations of the input text, and iteratively reduces vocabulary size to match a desired parameter. However, the match between the final and desired vocabulary size is not as precise as in BPE.

Despite its advantages, Sentencepiece Unigram may perform worse than other tokenization methods with limited training data. It may also struggle with rare or unseen words, compared to BPE, which starts from character-level representations, allowing it to handle almost any unseen token with constituent characters present in the training corpus.

3.1.4. Morphological (Morfessor)

Morphological tokenization approaches aim to preserve morphological information in languages with rich morphology. We chose to implement the Morfessor package. Based on the Minimum Defined Length (MDL) model, it is a widely used generative probabilistic model. It can be applied either semi-supervised, using annotated data, or unsupervised to segment the corpus. Although the generated 'morphs' may not precisely align with linguistically recognized morphemes, they aim to provide a closer approximation to true morphological segmentation.

However, there is limited evidence that morphological-based tokenization methods outperform data-driven approaches. In specific cases,

Tokenizer	Tokenized sentence
Char	I _ e s t a v a , _ c o m _ q u a n _ e r a p e t i t , _ c o n s t a n t m e n t _ d i b u i x a n t _ c ò m i c s _ i _ c o s e s _ a i x í .
BPE	I esta@@ va, com quan era peti@@ t, const@@ antment dibuix@@ ant c@@ am@@ ics i coses aix@@ i .
Uni	_I _estava , _com _quan _era _petit , _constantment _dibuix ant _còmic s _i _coses _així .
Morf	I _estava , _com _quan _era _petit , _constant ment _dibuixa nt _còmic s _i _coses _així .

Table 1: Sample sentence tokenized in each of the four chosen methods

such as low-resource or highly-agglutinative languages, they may offer benefits, especially when combined with statistical methods (Mielke et al., 2021; Park et al., 2021; Vania and Lopez, 2017).

3.2. Training Corpus

The models used in these experiments are trained from scratch on a parallel dataset of Catalan-English sentences, which was originally compiled for the creation of the Projecte Aina ca-en MT model (Projecte-Aina/Mt-Aina-ca-En · Hugging Face, n.d.). The dataset encompasses a wide range of domains to ensure the models' adaptability to different text types and contexts. The corpus is an amalgamation of several publicly available datasets, carefully curated to ensure high translation quality.

The combined corpus initially consisted of 11.5 million sentence pairs. To further enhance the translation quality, the dataset underwent a filtering process using a model trained on human-annotated data (de Gibert Bonet et al., 2022). This filtering step resulted in a refined dataset of 8,218,519 sentence pairs, which served as the primary training data.

Subsequently, the filtered dataset of 8.2 million sentence pairs was processed using the join-single-file.py script from SoftCatalà. This script was employed to normalize punctuation across the sentences, ensuring consistency and improving the overall quality of the training data. Furthermore, as a form of data augmentation, each sentence was duplicated into its uppercase counterpart. This augmentation technique increased the dataset size and provided additional variations for training. The final training database consisted of 16,437,038 sentence pairs.

3.3. Preprocessing and Training

Tokenization is the first step in the preprocessing of our data. For the BPE tokenization, we utilized the subword-nmt package, setting the number of merges to 32,000. This resulted in a Fairseq dictionary of 36,452 tokens. The Sentencepiece Unigram tokenization was implemented using the SentencePiece library, with a vocabulary size of 32,000. The resulting Fairseq dictionary consisted of 70,636 tokens. Character tokenization involved manually splitting the input strings into individual characters, resulting in no fixed vocabulary size.

	Char	BPE	Uni	Morf
Vocab size	8,164	36,542	70,636	149,996

Table 2: Fairseq dictionary size by tokenization method

The Fairseq dictionary for character tokenization contained 8,164 tokens. Lastly, we employed the Morfessor package to train a Morfessor-based tokenization model. Morfessor does not have a vocabulary parameter, but to ensure a manageable size for the NMT models, we limited the Fairseq dictionary to 150,000 tokens.

To facilitate fair comparison among the models, we ensured that the NMT models shared a common vocabulary. The Fairseq preprocessing pipeline (Mitchell et al., 2021) was applied consistently to the training data for all four tokenization methods, allowing for a systematic evaluation of their impact on NMT performance.

The training process involved feeding the preprocessed data into the Transformer base model using the fairseq-train script. We trained each model separately, adjusting only the path to the binarized data for each tokenization method. During training, we employed the Adam optimizer with a learning rate of 5e-4 and a weight decay of 0.0001. The models were trained for 250,000 updates with an update frequency of 8, and we utilised a batch size of 3,072 tokens.

3.4. Architecture

All of the machine translation (MT) systems trained for our experiments are based on the Transformer base model proposed by Vaswani et al. (2017). Our models have an embedding table with 512 dimensions, increased to 2048 in the 6 feed-forward layers. The models employ 8 attention heads.

4. Dataset Creation

The test set which we used to evaluate gender bias in Catalan is derived from the English-Spanish MuST-SHE (Bentivogli et al., 2020) [1], a test set for the investigation of gender bias taken from the larger MuST-C (Di Gangi et al., 2019). Both MuST-C and MuST-SHE are multi-modal and designed for

[1]An updated link to the original MuST-SHE corpus can be found here.

Form		
	Category 1: No gender info in text	
	SRC	But if you ask me today, I'm not so sure.
	C-Ref	Si m'ho pregunten avui, no n'estic tan **segura**.
	W-Ref	Si m'ho pregunten avui, no n'estic tan **segur**.
	Gender Terms	segura segur
	Category 2: Gender info in text	
Fem	SRC	**She** was tough, **she** was strong, **she** was powerful.
	C-Ref	Era **dura**, era **forta**, era **poderosa**.
	W-Ref	Era **dur**, era **fort**, era **poderós**.
	Gender Terms	dura dur;forta fort, poderosa poderós
Masc	SRC	But **he** was a political refugee from Angola.
	C-Ref	Però era un **refugiat polític** d'Angola.
	W-Ref	Però era una **refugiada política** d'Angola.
	Gender Terms	un una;refugiat refugiada;polític política

Table 3: Segments from the created English-Catalan corpus organized by category. In the C-Ref (Correct Reference Translation) we see the target gender-marked terms, while in the W-Ref they have been swapped to their opposite gender form. In Category 1 information as to the correct form is present in the audio data but 'correct' and 'wrong' are not relevant for text.

Speech Translation systems, consisting of audio, transcript and translation triplets, with English as the source language. As we are only interested in text-based translation for this project, we have discarded the audio and worked solely with the transcript and translation pairs.

MuST-C is a multilingual corpus compiled from TED talk data. The source language of the corpus is English and both the English language transcriptions and target language translations are generated for TED by volunteers (who may or may not be professional translators). MuST-SHE is a manually curated sub-set of MuST-C, with 1164 En-Es segments, in which each English sentence contains at least one gender-neutral word which requires a gendered translation in Spanish, where the gender of the translation corresponds with the sex of the referent (referred to henceforth as gender terms). In addition to the correct Spanish translation (C-Ref) of the English sentence, a gender-swapped reference (W-Ref) was also manually created by professional linguists (Bentivogli et al., 2020).

For the evaluation of our English-Catalan models, we created a synthetic English-Catalan version of MuST-SHE [2]. In order to do so both the C-Ref and W-Ref Spanish translations were automatically translated into Catalan using the PlanTL Project's Spanish - Catalan model (PlanTL-GOB-ES/Mt-Plantl-Es-ca · Hugging Face, n.d.). This model is based on the Transformer-XLarge architecture (Subramanian et al., 2021) and was trained on an aggregated dataset of approximately 92 million sentences. [3] It was evaluated across various domains and received an average BLEU score of 47.6. All translations were subsequently revised by a native speaker of Catalan with a Masters level education and a background in computational linguistics who was properly compensated for the work.

As in the original MuST-SHE dataset, for evaluation purposes the gender terms were extracted from both the C-Ref and W-Ref sentences and stored in a list of tuples. This was first done automatically using a script based on the Catalan morphologizer available from spaCy (Catalan · SpaCy Models Documentation, n.d.). Our script iterated through tuples of C-Ref and W-Ref sentences to identify tokens which shared an index, were not identical, and both posessed the feature 'Gender'. All such pairs of tokens were extracted and appended to the gender terms list for that sentence pair.

This heuristic for identifying gender terms relied on the C-Ref and W-Ref sentence pairs being identical except for the gender-swapped terms. However, as the translations were generated independently, some pairs had small differences in phrasing which threw out the alignment and led to incorrect gender terms being extracted. There were also several sentences which had contained gender terms in Spanish but did not in Catalan as the Catalan equivalent is gender neutral (i.e. tonto/tonta –> ximple). The extracted gender terms were therefore manually revised. Sentences which did not contain gender terms were discarded, and incorrectly extracted gender terms were manually corrected.

Sentences were also discarded when the only gender term present was a determiner (un/una, el/la) and other non-target determiners were also present in the translations, as it was impossible to

[2]The English Catalan MuST-SHE is available here

[3]Model details can be found here.

distinguish between them at the evaluation stage.

In the original dataset all triplets were assigned a category from 1-4 based on the type of gender reference present in the sentence. Categories 1 and 3 both make reference to the preferred gender of the speaker (singularly in Category 1, in conjunction with others of the same gender in category 3) while Category 4 was for sentences which do not contain any gender-disambiguating information. In the context of Speech Translation these distinctions make sense, but when working purely with text they do not and so we recategorized all segmements into either Category 1: No gender information present in text or Category 2: Gender information present in text.

The remaining dataset after all editing had been completed consisted of 1046 sentence triplets (English reference, C-Ref, W-Ref) with their corresponding gender terms. 480 (45.9%) triplets have gender information present in the source text (Category 2), with 242 (22.8%) containing feminine gender terms and 242 (23.1%) containing masculine gender terms. The remaining 566 (54.1%) segments do not contain gender information in the source text and are therefore classified as Category 1.

5. Evaluation of Gender Accuracy

The MuST-SHE corpus includes a script to evaluate the accuracy of the gender terms generated by the model. This script checks for the presence of the correct gender terms in each sentence of the output. If it is unable to find a correct gender term, it checks for the presence of the incorrect gender term. A count is then stored at the sentence level of the total number of gender terms expected in the sentence, the number found (combining correct and incorrect), the number of correct terms found and the number of incorrect terms found as well as the number of terms not found. It outputs a global score as well as a score broken down by category (see below), with the option to output sentence level scores.

In order to extract more information and customise the results for our experiment we have made slight modifications to this original script. For additional insight into the translation output, it has been amended to store the POS tag for terms in each category, allowing for more fine-grained analysis. This was done using the Catalan morphologizer available from spaCy (Catalan · SpaCy Models Documentation, n.d.). As the Catalan dataset has different categories for ST (CATEGORY) and MT (TEXT-CATEGORY), the option to work purely with text-based translation has been added to the script.

We also amended the script to improve the accuracy of the results. In initial testing we discovered that the script was producing inaccurate results due to multiple instances of the same term occurring with different genders inside one sentence. This particularly affected common function words such as determiners. This was partly remedied by curating the dataset to minimize such occurrences, but we also amended the script to only search for the wrong gender term if the correct term had not been found.

Finally, we made small changes to the script to ensure that word-final punctuation was stripped from words before attempting to match them with the gender terms, thus increasing the percentage of terms which were successfully identified.

	Char	BPE	Uni	Morf
Flores Dev	39.2	41.0	**41.6**	41.2
Flores DevTest	39.1	41.3	**42.0**	41.5
MuST-SHE	51.6	56.4	**57**	56.6
AVG	43.3	46.2	**46.9**	46.4

Table 4: SacreBLEU scores on en-ca test sets with different tokenizers

	Char	BPE	Uni	Morf
Flores Dev	0.8429	**0.8623**	**0.8623**	0.8601
Flores DevTest	0.8428	0.8595	**0.8598**	0.8551
MuST-SHE	0.8194	0.8469	0.8472	0.8484
AVG	0.8350	**0.8562**	**0.8564**	0.8545

Table 5: COMET scores on en-ca test sets with different tokenizers

6. Results

6.1. Overall Translation Quality

To assess tranlation quality we chose to evaluate on the Flores-200 dataset (NLLB Team et al., 2022) in addition to MuST-SHE. Flores is a commonly used evaluation dataset and this gave us a more widely known benchmark.

Table 4 presents the overall translation quality of the MT systems trained with each of the four distinct tokenization methods as measured by BLEU scores. Sentencepiece Unigram perfoms best across all of the test sets. BPE and Morfessor are only 0.2 BLEU apart, on 46.2 and 46.4 respectively, less than 1 BLEU point behind Sentencepiece Unigram. There is a substantial gap (3 BLEU) between these two and Char, which we had predicted would perform more poorly here than in the ST experiments of Gaido et al. (2021) as phonemes are easier to map to individual characters.

While Morfessor achieves a performance level very slightly higher than that of BPE in terms of BLEU scores, it incurs a substantially higher computational cost due to the need to train the tokenisation model. Consequently, BPE emerges as the preferred choice between the two, with greater bal-

	Char	BPE	Uni	Morf
1F	**76.61%**	75.46%	76.14%	73.62%
1M	76.61%	76.86%	**79.43%**	78.15%
2F	76.02%	76.80%	**78.36%**	75.24%
2M	77.62%	81.19%	**83.37%**	80.00%
Global	76.71%	77.58%	**79.33%**	76.75%

Table 6: Term Coverage for the different models

ance between translation quality and computational efficiency. In terms of BLEU scores alone, Sentencepiece Unigram tokenization appears to be the best approach, highlighting its ability to generate accurate and contextually appropriate translations.

In addition to BLEU scores we also calculated COMET scores, as displayed in Table 5. We used the default COMET model, which is trained using a reference-based regression approach and is one of the only MT evaluation metrics to take the source material into consideration. By removing the reliance on n-grams this aims to produce a score more comparable with a human evaluation. COMET also provides a tool to compare multiple systems using Paired T-Test and bootstrap resampling, and the results show no significant difference between the output of the BPE, Sentencepiece Unigram and Morfessor models, but a significant ($p < .001$) difference between the Char model and each of the others.

6.2. Gender Bias

6.2.1. Term Coverage

Only target gender terms which appear in the translated output can be evaluated for accuracy or bias. In cases where the translation and reference text differ in their lexical choices, the expected gender terms may not appear in either their masculine or their feminine form. Thus the number of found terms is a more important metric than the expected number of gender terms and is the one from which the F1 scores are calculated in the following section. No gender terms were found by any model in 109 segments and so these segments were not used in the analysis. The term coverage of the different models can be seen in 6

6.2.2. Gender Accuracy with Context

Gender accuracy, as measured by F1 scores for sentences with available gender information (Category 2), clearly demonstrates the presence of bias in all of the translation models. All models exhibit a substantial disparity in translation accuracy between masculine and feminine forms, with a preference for masculine forms. Despite the availability of contextual information, models struggle to preserve knowledge of feminine gender.

	Char	BPE	Uni	Morf
F	57.95%	**65.99%**	65.92%	60.10%
M	98.47%	**98.78%**	98.34%	98.51%
Overall	78.21%	**88.39%**	82.13%	79.31%

Table 7: F1 scores for gender translation with provided context

BPE achieves the highest gender translation accuracy at 88.39%, followed closely by Sentencepiece Unigram (82.13%). Morfessor does slightly worse with 79.31%, while character-based tokenization lags with an accuracy rate of 78.21%. These results suggest that both BPE and Sentencepiece Unigram are deserving of their current status as the SOTA tokenization methods for NMT.

All models achieve very similar accuracy rates of over 98% for the translation of masculine terms. Inspection of the cases where they were identified as outputting female in place of male shows that these sentences were either ambiguous, often relying on the gender of first names which the models would not have been exposed to in training but which had features that are commonly feminine in this language pair, such as ending with the letter 'a', or that the evaluation script had incorrectly identified the gender terms. The exception to this is Morfessor, which produced feminine outputs in two sentences which had unambiguous masculine features in English.

The errors in generating feminine forms were more nuanced, with the following scenarios occurring across all four models: mis-classification resulting from names that may not have been part of the training data; sentences in which gender ambiguity arises as the conversation transitions from specific to general, prompting the models to revert to masculine terms as a neutral default; sentences in which gender is not explicitly mentioned in direct connection to the gender-terms, but the presence of other female forms suggests a feminine context, a determination that a human translator would be able to make; and correctly gendering a portion of the sentence while leaving the part preceding the introduction of a feminine reference in the masculine form, even though the co-reference is clear.

Collectively, these observations underscore that across the board the models predominantly regard the masculine as the default form. In the absence of unequivocal feminine context, they tend to produce male translations. The differences in the performance of the models was small enough that it is not possible to distinguish patterns in the differences between them, beyond the fact that Char and Morf both produced masculine forms sporadically in sentences that were unambiguously feminine and did not fit into any discernible pattern, and that Char did so with noticeably more frequency.

	Char	BPE	Uni	Morf
F	11.87%	10.35%	10.61%	**15.04%**
M	88.13%	**89.65%**	89.39%	84.96%

Table 8: Generation of Feminine/Masculine terms with no provided context

6.2.3. Gendered Output with No Context

When the models were not provided with any context with which to determine the necessary gender, there is a notable difference between the proportions of feminine forms in the models' output. Although all of the models produce well over 80% masculine forms, Morfessor generates the highest percentage of feminine forms (15.04%), while BPE has the lowest (10.35%), with SentencePiece Unigram slightly better at 10.61% and Character Based at 11.87%.

A trend observed across all models is that, when no context is given, the English word "nurse" is always translated into the feminine form in Catalan, while "doctor" is always translated into the masculine form. This speaks to the influence of stereotypes present in the training data and there may be similar patterns leading to the translation of other terms as female when no disambiguating information is given, but the limited data did not allow us to identify them. Although professions relating to education (educator, professor, teacher) were not always translated as feminine, they did occur as feminine in the output of all models. Other words which occurred in feminine forms but were not consistently translated as feminine include "proud" (orgullosa) and "myself" (mi mateixa)

An investigation of the occasions on which Morfessor had generated female forms while the other models had produced masculine was not able to give any further insights. Although Morfessor produced more feminine forms than any other model, it is limited to a total of 94 feminine gender terms (as opposed to 503 masculine gender terms) and so there is insufficient data to identify more detailed patterns in why this is taking place.

7. Discussion

This research explored a unique dimension of the problem of gender bias in MT by examining the impact of tokenization methods on gender bias in translation. We conducted experiments using four distinct tokenization techniques: Byte Pair Encoding (BPE), Sentencepiece Unigram, Character-Based, and Morphological tokenization. These tokenization methods play a significant role in shaping the output of machine translation systems. Our findings revealed that SentencePiece Unigram and BPE have similar levels of performance, with SentencePiece Unigram having a slight lead on BLEU scores and BPE edging ahead in gender accuracy. It has previously been suggested that BPE may capture some morphological information in its subword segmentation, and it is possible that this, combined with the greater granularity of its tokenization compared to SentencePiece Unigram (which does not begin from character level representations), allows for the improved gender accuracy. However, further work would be necessary to confirm this hypothesis. What is clear is that, as initially proposed, there is a substantial difference between the gender accuracy of Char in ST (Gaido et al., 2021) and its performance in MT.

When it came to context-free sentences, we observed a noteworthy difference between tokenization methods. Morfessor generated the highest percentage of feminine forms, while BPE exhibited the lowest. This divergence underscores the influence of tokenization on the generation of gender information when models are presented with a gender neutral source text. It is clear that all of the models are treating male forms as neutral and thus predominantly generating them unless prompted to output female forms by the presence of stereotypically female terms, although further work is required to identify which characteristics of tokenization with Morfessor lead to the increase in feminine forms. The debate around what should be considered neutral forms in Catalan and other grammatical gender languages is an ongoing one, but in the realm of NLP and MT, this debate is compounded by the challenge of balancing linguistic traditions and societal expectations with the pursuit of fairness and inclusivity. Work has begun on the production of gender-neutral language in English (Sun et al., 2021) and it is our hope that this finding will help to stimulate discussion of how what gender-neutral language should look like for the output of MT models in grammatically gendered languages.

Our results also raised questions about the underlying data-driven bias in machine translation models. The presence of gender bias in the training data has a profound impact on the translation output, as evident from the consistent generation of stereotypically feminine terms, irrespective of contextual clues.

A further contribution of this study is the creation of a new dataset for evaluating gender bias in Catalan. Building upon the English-Spanish MuST-SHE dataset, we synthesized an English-Catalan version, enriching the resources available for examining gender bias in machine translation. Our dataset provides a valuable resource for further research in the field of computational linguistics and gender-inclusive translation techniques.

8. Conclusion

In conclusion, this study advances our understanding of gender bias in machine translation by highlighting the crucial role of tokenization methods. It underscores the complexities of addressing gender bias and the need for comprehensive approaches that consider not only translation quality but also gender accuracy. While tokenization can significantly influence gender bias, it cannot operate in isolation. Future research should focus on holistic strategies that encompass data-driven bias mitigation, architectural design choices, and, of course, tokenization methods. By addressing gender bias in MT, we contribute to the development of more inclusive and equitable language technologies, advancing both the field of computational linguistics and the broader societal goal of promoting diversity and equality in communication.

9. Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335 y 2022/TL22/00215334.

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project.

10. Bibliographical References

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.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. ArXiv:1607.06520 [cs, stat].

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. 2020. Gender Bias in Multilingual Neural Machine Translation: The Architecture Matters. Number: arXiv:2012.13176 arXiv:2012.13176 [cs].

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

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to English-Arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. Technical report. Publication Title: arXiv e-prints ADS Bibcode: 2019arXiv190103116E Type: article.

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, Barcelona, Spain (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. Number: arXiv:2105.13782 arXiv:2105.13782 [cs].

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.

Jindřich Libovický, Helmut Schmid, and Alexander Fraser. 2022. Why don't people use character-level machine translation? ArXiv:2110.08191 [cs].

Sally McConnell-Ginet. 2013. ' Gender and its relation to sex: The myth of 'natural' gender. In '

Gender and its relation to sex: The myth of 'natural' gender, pages 3–38. De Gruyter Mouton.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. ArXiv:2112.10508 [cs].

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology Matters: A Multilingual Language Modeling Analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Beatrice Savoldi. 2021. Data Statement for MuST-SHE.

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

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. ArXiv:2112.14168 [cs].

Sandeep Subramanian, Oleksii Hrinchuk, Virginia Adams, and Oleksii Kuchaiev. 2021. NVIDIA NeMo's neural machine translation systems for English-German and English-Russian news and biomedical tasks at WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 197–204, Online. 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:2102.06788 [cs].

Clara Vania and Adam Lopez. 2017. From Characters to Words to in Between: Do We Capture Morphology? ArXiv:1704.08352 [cs].

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. ArXiv:1909.05088 [cs].

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.