Universitat Pompeu Fabra Barcelona

Treball de fi de màster de Recerca

The impact of tokenization on gender bias in Machine Translation

Nom i Cognoms Audrey Mash

Màster: Lingüística Teòrica i Aplicada

Edició: 2022-2023

Directors: Dra. Maite Melero Nogues Dra.

Any de defensa: 2023

Col·lecció: Treballs de fi de màster

Departament de Traducció i Ciències del Llenguatge

Acknowledgements

I would like to express my sincere gratitude to several individuals and organizations who have contributed significantly to the completion of my master's thesis. First and foremost, I would like to thank my supervisor, Dr. Maite Melero, for her invaluable guidance, support, and expertise throughout this research journey. Her insightful feedback and continuous encouragement have been instrumental in shaping this thesis.

I would also like to extend my appreciation to Carlos Escolano for his exceptional patience and support. His willingness to answer my numerous questions has been greatly appreciated and it was his work which laid the groundwork for mine.

I am grateful to the Language Technology unit at BSC for their warm welcome and collaboration. I would like to thank Francesca De Luca Fornaciari for her assistance and understanding during my busy periods, as she graciously stepped in to support me.

I am deeply indebted to all the professors from the UPF Master in Theoretical and Applied Linguistics program. Their expertise and dedication have been instrumental in fostering significant personal and professional growth.

Finally, I would like to express my heartfelt appreciation to my husband, Rohan Schoeman, for his unwavering patience, love, and support. His understanding and encouragement have been vital throughout this demanding academic endeavor.

Thank you all for your invaluable contributions to the completion of this thesis.

Abstract

This study examines the impact of tokenization methods on gender bias in Neural Machine Translation (NMT). Unigram, BPE, Character, and Morfessor tokenization approaches are compared in terms of translation quality measured by BLEU scores and gender accuracy. Results show that Unigram achieves the highest BLEU scores, closely followed by BPE and Morfessor, while Character performs lower. However, all models display a bias towards generating masculine forms more frequently than feminine forms in gender accuracy analysis. They also overwhelming generate masculine forms when no context is provided. The Unigram method exhibits the highest accuracy for both feminine and masculine forms, surpassing BPE and Morfessor. These findings emphasize the need to address gender bias in MT systems and the complex relationship between tokenization methods, translation quality, and gender accuracy. Further research is warranted to explore additional factors influencing gender bias. This study contributes to the development of inclusive and unbiased translation technologies.

Key words: Machine Translation, Neural Machine Translation, Sub-word tokenization, Gender bias, Unigram, BPE (Byte Pair Encoding), Character-based tokenization, Morfessor

Contents

1. Introduction

Machine translation (MT) has undergone significant advancements since its inception in the 1950s. From rule-based approaches to statistical machine translation (SMT) and the introduction of neural machine translation (NMT) with the Transformer architecture, the field has witnessed remarkable progress in translation quality and reduced reliance on linguistic expertise. However, the complexity of deep learning algorithms has raised ethical concerns regarding the opacity and lack of interpretability of NMT models, commonly referred to as "black boxes." The recent rise of generative models, exemplified by ChatGPT, has further intensified the ethical discourse surrounding artificial intelligence (AI) and deep learning, particularly in natural language processing (NLP) and MT.

While generative models have received significant attention regarding ethical issues, NMT systems have not yet been subject to the same level of scrutiny. However, bias, in particular gender bias, has emerged as a prominent ethical concern within the field. Biases learned by NMT models can perpetuate stereotypes, reinforce existing inequalities, and lead to the invisibility or underrepresentation of certain groups.

This thesis aims to explore the phenomenon of gender bias in NMT and investigate the impact of different subword tokenization methods on gender bias in machine translation. By analysing the relationship between tokenization and gender bias, this study seeks to contribute to a more comprehensive understanding of the factors influencing gender bias in NMT systems. Based on the existing literature and identified research gaps, this thesis proposes a research question: To what extent does the choice of tokenization method for NMT impact gender bias in machine translation?

The research question aims to investigate the relationship between tokenization methods and gender bias in NMT. Different tokenization approaches can affect the granularity and representation of gender-related information in the input text, potentially influencing the translation output. By analysing and comparing the performance of NMT models trained with various tokenization methods, this study seeks to provide insights into how tokenization choices can impact gender bias in machine translation.

In conclusion, this thesis aims to contribute to the understanding of gender bias in NMT and its relationship with tokenization methods. By investigating the impact of different tokenization approaches on gender accuracy in machine translation, we can gain insights into how to develop more inclusive and unbiased NMT systems. The findings of this research will have implications for the development and deployment of AI systems, promoting fairness, and addressing ethical concerns in natural language processing and machine translation.

2. Background

MT has been through several evolutions since it first emerged as a viable research field in the 1950s. The demonstration of a small-scale system for English-Russian translation in 1954 sparked the public interest (Hutchins, 2004). Initially, rule-based methods dominated MT, relying on bilingual dictionaries and complex sets of rules for language conversion. However, this approach was labour-intensive, difficult to scale, and lacked the ability to transfer between languages (Wang et al., 2022). In the 2000s, SMT emerged as a more effective approach, leveraging large amounts of bilingual data to automatically extract translation knowledge. Nevertheless, SMT still required manually generated components (Lopez, 2008; Wang et al., 2022). The advent of NMT in the 21st century, particularly with the introduction of the

Transformer architecture, brought substantial improvements in translation quality and reduced the reliance on linguistic expertise.

The Transformer architecture, first introduced by Vaswani et al. (2017) revolutionised NMT by introducing the concept of attention, allowing models to focus on the relevant parts of the source text when generating translations. Such models, powered by deep learning techniques, achieved state-of-the-art performance and improved fluency in MT and other fields of Natural Language Generation. The rapid rise of NMT can be attributed to its ability to automatically learn translation patterns from large datasets, reducing the need for manual rule engineering; however, the complexity of deep learning algorithms means that the exact workings of the models are opaque, and so are often referred to as 'black boxes'. This lack of visibility on the inner workings of deep learning models has long raised ethical questions about the widespread adoption of AI across a variety of fields, and not only NLP or MT (Walmsley, 2021).

With the recent surge in publicity around generative models following the launch of ChatGPT and its competitors, NLP has become a focal point for the discussion of ethical issues around artificial intelligence and deep learning. The issues making global headlines have primarily related to the tendency of generative models to invent facts and even sources in order to provide responses (i.e. Moran, 2023), and questions of privacy around the vast quantities of data on which the models are trained which has led to ChatGPT being banned in Italy, with other EU countries considering following suit at the time of writing (Mukherjee et al., 2023). Translation software is an unquestioned aspect of many more people's daily lives than the novel GPT technology, and yet has received comparatively less ethical scrutiny. Despite this, NMT is not without its ethical issues, among the most visible of which is bias and more specifically gender bias (Ahmed et al., 2022; Savoldi et al., 2021; Sun et al., 2019).

i. Gender and Linguistic Diversity

The task of accurately translating gender without either perpetuating or amplifying gender bias is made more complex by a) the disconnect between social and linguistic categories of gender (Stanczak & Augenstein, 2021) and b) the range of ways in which diverse languages mark gender. Gender in linguistic terms refers to a noun class, which controls formal properties or agreements of other expressions headed by the noun phrase (McConnell-Ginet, 2013). The level of agreement varies immensely between languages. In so-called notional gender languages agreement is limited to pronouns and gender assignment is generally explicitly linked to sex, while in grammatical gender languages it is a morphosyntactic feature expressed over a range of parts of speech, including articles, adjectives and even verbal expressions, and with a less clear cut connection to sex (McConnell-Ginet, 2013). Further complicating the issue is the fact that over half of the world's languages do not have gender as a feature (Corbett, 2014a).

Culturally, gender in many parts of the world has traditionally been thought of as male or female and connected to biological sex. However, even before recent cultural shifts in that sphere (such as the increasing adoption of non-binary pronouns in English) this was not the case for linguistic gender. Many (but far from all) languages which mark gender have a neutral gender class and many languages have more than the three genders represented by male, female and neutral. Corbett (2014b) gives the example of the Nakh language Batsbi, which depending on how it is analysed can be said to have either four or eight genders.

The differences between languages without gender, languages with notional gender, and languages with grammatical gender can be seen clearly in the examples below:

1) Hän on tyytyväinen arvosanoihinsa. **(Finnish)**
Pron3SG(M/F) is satisfied with-grades-POSS3PL

2) She is happy with her grades. **(English)**
Pron3SG(F) is happy with poss_pron3SG(F) grades

3) He is happy with his grades. **(English)**
Pron3SG(M) is happy with poss_pron3SG(M) grades

4) Ella està contenta amb les seves notes. **(Catalan)**
Pron3SG(F) is happy with poss_pron3PL(F) grades

5) Ell està content amb les seves notes. **(Catalan)**
Pron3SG(M) is happy with poss_pron3PL(F) grades

In these examples we see that Finnish, a language without gender, has no gender distinction in its pronouns and this sentence could refer equally to a male or female subject. The only way to clarify would be to use the noun *mies* (man) or *nainen* (woman) in place of the gender neutral pronoun *hän* (he/she), and these nouns would not trigger any morphosyntactic agreement. In contrast, English, a notional gender language, has gender agreement with the subject on both the personal and possessive pronouns but nowhere else. The noun 'grades' is ungendered and therefore does not require agreement from the possessive pronouns, leaving them free to agree with the head noun of the phrase. In our grammatical gender language, Catalan, the (optional) personal pronoun and the adjective both agree with the subject, while the article and the possessive pronoun agree with the head noun of their sub-clause.

Although these divisions between languages are necessarily rough, they give a broad outline of the typological distinctions which can be found and thus the problems which arise in translation. Translating a neutral noun to one which has a fixed gender (i.e. grades [en] → notes [ca]) does not pose a problem, but when the target language has a choice of genders and the

source language offers no indication which is correct (i.e. hän [fi]→ she/he [en], happy [en]→ content/contenta [ca]), how is the translator to assign an gender?

This would be a tricky question for a human translator, but a human frequently has access to a broader context and the possibility of additional research. An MT system is limited to the information provided in the window of context from which it can extrapolate, and if the information is not linguistically present, it has no further recourse. As a result, when provided with insufficient information MT systems will provide the statistically most likely gender inflection (Vanmassenhove et al., 2019), therefore perpetuating stereotypes and amplifying existing biases.

One possible solution to this is to train MT systems to produce gender neutral alternatives when no disambiguating information is provided. Sun et al. (2021) trained a model to re-write gendered sentences in English as gender neutral, demonstrating that this is a viable option. However, acceptance of gender neutral forms varies across languages and its wide adoption in English is not replicated in many other languages. This issue undoubtedly deserves further work and attention, but for the purposes of this paper we will be working within the gender binary.

ii. Gender Bias

Blodgett et al. (2020) looked at 146 recent papers investigating bias in NLP and identified a lack of consistency or clarity around the way in which the term ‘bias’ is conceptualised, and the harms which it is responsible for. Following their recommendations for best practice in the field, the following section will aim to make explicit what is meant in this work by gender bias and the negative implications of its presence in NMT.

In the context of this work, ‘bias’ is used to refer to information learned by models which is a) not necessary for the task for which they are being trained and b) potentially harmful (Caliskan et al., 2017; Stanovsky et al., 2019). In many ways this is an oversimplification of a complex topic with competing definitions (Savoldi et al., 2021), but for our purposes here it suffices. It is important to recognise that this categorization of bias is inherently normative, requiring as it does an agreement of what constitutes ‘harms’ to certain groups and a belief in the ethical value of debiasing (Deery & Bailey, 2022; Savoldi et al., 2021). Without entering into the philosophical debate around these issues, we accept this underlying principle.

Biases commonly deal with stereotypes which are present in society and therefore in the data on which the model is trained. Although they can take many forms, two of the most commonly identified in machine learning are racial and gender bias. If not corrected, both of these can lead to the exacerbation of existing inequalities (Mitchell et al., 2021). It has also been shown that rather than simply reproducing existing biases in the data or in society, models are instead prone to amplify biases found in the data, thus worsening the situation (Bolukbasi et al., 2016; Zhao et al., 2017). For example, in various tasks including visual recognition, language generation or MT models frequently assign women to stereotypically female activities or occupations such as ‘cooking’ or ‘nurse’, despite the presence of men in these roles in the data (Zhao et al., 2017).

Savoldi et al. (2021) provide an overview of gender bias in MT, highlighting the myriad ways in which it can affect users. Following earlier work, they classify the effects of bias into two categories; representational and allocational harms. Representational harms include ‘detraction from the representation of social groups and their identity’ (Savoldi et al., 2021, p. 2), whereas allocational harms relate to the allocation or withholding of resources. The representational harm which can be perpetuated by MT systems is clear, as automatically

translating gendered terms around stereotypically male or female occupations or attributes with the stereotypical gender inflections serves to reduce the visibility of, for example, female doctors or non-binary scientists. Allocational harm is less immediately obvious, but Savoldi et al. (2021) argue that the inferior quality of the service provided to groups who are rendered less visible can be considered an allocational harm. Other examples of allocational harm could include the biassed allocation of educational opportunities, limited access to healthcare services, disparities in legal outcomes, and exclusion from social and cultural inclusion initiatives.

Friedman and Nissenbaum (1996; expanded upon in Savoldi et al., 2021; Stanczak & Augenstein, 2021) identify three types of bias which can be found in computer systems, two of which are most relevant here. *Pre-existing bias* can be added to the model consciously or unconsciously and refers to the incorporation into the model of independently occurring biases. Such biases are frequently a result of the historical and cultural context in which the models are developed. In MT an example of this would be gender imbalances in the massive corpora on which models are trained, and which are often crawled from the web and are too large for manual inspection. These imbalances may lead for example to the model learning that doctors are most likely to be male and nurses female, and therefore always translate accordingly unless provided with clear markers to the contrary (Romaine, 2001).

This differs from *technical bias,* which Friedman and Nissenbaum (1996) limit to bias that results from decisions made in the model design process. This may again apply to data, for example the test sets used to evaluate a model or any decisions made to debias (or not) the data. It may also apply to architectural choices in the models, such as those which lead to the amplification of existing biases (Savoldi et al., 2021; Vanmassenhove et al., 2019). The final type of bias noted is *emergent bias,* which develops when systems are used in contexts for

which they were not originally intended, therfore creating unexpected sources of bias which were not considered in the development process.

iii. Existing Approaches to Gender Bias in MT

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., 2018). In this method the training data is automatically annotated with gender information, either at sentence level (Elaraby et al., 2018; Vanmassenhove et al., 2018) or at word level (Stafanovičs et al., 2020).

Vanmassenhove et al. (2018) 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 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 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 (ST), 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. However, it requires a substantial effort to produce suitable training data, as well as 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. (2018). 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 NMT models with both Shared and Language-Specific Encoder-Decoders and found that the Language-Specific encoders-decoders exhibit less gender bias than the Shared encoder-decoder architecture while also achieving better BLEU scores.

Finally, Gaido et al. (2021) looked at the impact of tokenisation on gender bias in ST. As this forms the basis for the present study, their paper will be discussed in more detail below.

iv. Tokenization

As can be seen from the background provided above, much of the work which has been done on bias in MT to date has focused on data, and particularly on pre-existing bias which is present in society and therefore is also found in the training data. There have been attempts to remedy gender imbalances found in this data by providing curated, gender-balanced datasets for training; gender-tagging all gendered entities in the training set, and developing counter-factual datasets to counter stereotypes (i.e. masculine doctors and feminine nurses) often found in training data (Savoldi et al., 2021). This has led to improvements in output, but they do not account for the factors present in the models which lead to the amplification of the biases present in the data. At present, there is limited work which looks at ways in which the model architecture affects bias (Savoldi et al., 2021) and whether it can be used to mitigate the effects of biassed data or at least reduce its amplification.

NMT models, including Transformers, make use of an encoder-decoder architecture. At a high level, the encoder takes the 'input sequence and creates a contextualised representation of it [...] This representation is then passed to a decoder which generates a task-specific output sequence' (Jurafsky & Martin, 2023), in this case the input translated into the target language. Importantly, the input must be divided into tokens to be processed by the model, and the question of how this should be done is key.

Tokenization is the process of segmenting a textual input into smaller units. Early tokenization methods split text inputs into 'words' by splitting on whitespace, but this method

has several drawbacks, including the fact that it is not suitable for languages such as Chinese or Japanese which do not use whitespace as a typographic separator. The rise of neural language models has led to the rise of sub-word tokenization, in which tokens do not correspond to word-forms but to characters or sequences of characters.

Dividing text into smaller tokens allows for open-vocabulary processing, in which rare words not seen in training can be better handled by the model, while maintaining a fixed-size vocabulary which allows for faster processing. The underlying theory of subword translation is that many unknown or rare words are potentially translatable based on morphemes and phonemes. Types of words that are often transparent in this way include NEs, cognates/loanwords and morphologically complex words (i.e. compounds, affixation, inflection) (Sennrich et al., 2016). Therefore, if words can be broken down appropriately the NN can produce translations.

The dominant algorithms in current NMT systems are Byte-Pair Encoding (BPE) and the Unigram method, but Domingo et al. (2019) show that there is not necessarily one single form of tokenisation which is best for all languages and all contexts. They investigated the impact of different tokenization methods on overall translation quality and found that the optimal tokenization method differed across language pairs, and that even within a language pair the results could differ depending on the direction.

Savoldi et al. (2021) note that there has been little work done to date on the impact the chosen method of subword tokenization may have on the preservation of the meaning of morphologically more complex feminine forms. In the field of ST, Gaido et al. (2021) conducted a study to examine the effects of different subword tokenization approaches, namely 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.

It cannot be assumed that similar results would be obtained in the field of MT due to the inherent differences between ST with an Automatic Speech Recognition (ASR) component, and MT. Unlike MT, ASR deals with audio input and transcribes it into text, requiring the model to accurately place word boundaries and transcribe spoken language. Additionally, audio data in ASR 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 ASR compared to 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. Further research is needed to explore this question.

v. Research Question

In light of the existing literature and the identified research gap, this study aims to investigate the influence of different methods of subword tokenization on gender bias in NMT. Specifically, we seek to explore how various subword tokenization approaches, namely character-based tokenization, BPE, unigram tokenization and Morfessor, affect the translation

of gendered terms in English - Catalan machine translation. By examining the impact of tokenization on MT, this research intends to shed light on an overlooked aspect of gender bias mitigation in NMT models, contributing to a more comprehensive understanding of the factors influencing gender bias in machine translation systems. This investigation aims to address the following research question:

To what extent does the choice of tokenization method for NMT impact the gender bias of the model's output?

This research question focuses on investigating the association between subword tokenization methods and gender bias, with a specific emphasis on gender accuracy. Given that NMT operates solely on text data without access to audio clues or speaker information, we anticipate that the results of this study will differ from those of ASR-related research. By extending the work of Gaido et al. (2021), this study aims to explore whether similar effects can be observed in the context of NMT, where the data is solely textual in nature.

3. Methodology

i. Subword Tokenization

This section briefly introduces the four tokenization approaches compared in our experiments. BPE and Unigram tokenization methods were selected based on their current prominence in the field of NMT. We also chose to include the most popular morphological tokenization method, Morfessor, and Character Segmentation in order to compare their performance here with their performance in ST. Despite having been the best performing method for gender accuracy in ST, it is expected that Character will perform poorly here due to the limited amount of semantic information contained in each token. Morphological methods

of tokenisation, on the other hand, should contain a larger quantity of semantic information and therefore can be expected to perform well

Character-Based

One of the drawbacks of segmenting texts into word-forms is that systems are, for practical reasons, limited to vocabularies of finite-size which necessarily exclude many rare words, leading to relatively frequent out-of-vocabulary tokens (Libovický et al., 2022). To circumvent this, Costa-jussà & Fonollosa (2016) proposed using character-level embeddings. This method of tokenization limits the vocabulary size to the number of characters, ensuring all tokens are covered without the possibility of OOV tokens.

Tokenization systems based on words or sub-words do not take into account information about the orthographic representation of the token (Costa-jussà & Fonollosa, 2016; Libovický et al., 2022), which results in a loss of all morphological information which is internal to the token, and leads to the possibility that character-level relations between tokens are not reflected in the model. On the other hand, the use of character-based tokenization ensures that relationships at the level of orthography are fully captured. In morphologically rich languages where base forms are subject to extensive inflection, this may allow for better preservation of information regarding grammatical features such as tense or, for our purposes, gender. Character-based systems also exhibit greater robustness towards source-side noise in the training data (Libovický et al., 2022), enabling them to handle spelling variations and typos more effectively compared to other sub-word tokenization methods.

The counter-balance to this is that character-based tokenization is associated with much higher computational costs, meaning that it requires more training time, more memory and more computational resources than other methods (Libovický et al., 2022). Also, as characters

are used in an immense variety of different contexts, each character contains far less semantic information than a larger sub-word token might. Using character level representations in the encoder side of the model has therefore had less success than using them in the decoder (Libovický et al., 2022).

Byte Pair Encoding

Byte Pair Encoding (BPE) (Sennrich et al., 2016) has become one of the most popular tokenisation methods in the field of NMT today, and is used in many popular models such as GPT-3 (Brown et al., 2020). Like all the sub-word methods of tokenisation considered here, BPE provides a solution to the question of vocabulary limitations.

BPE starts by initialising the vocabulary at the character level. The algorithm only requires one hyperparameter: the number of merge operations. In each merge operation, the most frequent pair of characters or character sequences in the input is replaced with a new symbol. The final vocabulary size is determined by the desired number of merge operations plus the initial character-based vocabulary size (Sennrich et al., 2016).

This mix of character and sequence level representations gives BPE its advantage. The resulting sequence-level representations make BPE more efficient and less computationally demanding than character-based tokenization. At the same time, the character-level representations ensure that BPE can process any input text even if a word-form was not seen during training and cannot be handled by the larger sub-word units.

Despite BPE's popularity in NLP applications, it has some limitations. One weakness is the possibility of one sequence being tokenized multiple different ways by the same model, and being treated as completely different inputs as a result (Kudo, 2018). Additionally, the sub-word units generated by BPE do not always correspond to linguistically meaningful units,

which Gaido et al (2021) suggest may result in a loss of semantic information conveyed in the morphology,

Unigram model

The Unigram model of tokenisation gained popularity when it was proposed that it be combined with subword regularisation to provide greater robustness to noise and segmentation errors such as those to which BPE could be vulnerable (Kudo, 2018). Although the method of subword regularisation proposed is no longer limited to Unigram and can also be used with BPE through the SentencePiece library, the unigram algorithm remains a popular method of subword tokenization.

Unigram looks at the frequency of individual characters and character sequences (n-grams) and determines their probability in the training data. In contrast to BPE which begins with a limited vocabulary (characters) and iteratively increases the size of the vocabulary, the Unigram algorithm initializes with a large seed vocabulary composed of multiple possible segmentations of the input text, and at each step removes the least probable n-grams (Kudo, 2018). The desired vocabulary size is a required parameter and the process ends when the vocabulary size approximately matches this parameter, but the match between final and desired vocabulary size is not as exact as in BPE.

Despite its advantages, the Unigram algorithm also has limitations. Domingo et al. (2019) found that it performed worse than other tokenization methods with a limited training corpus. Additionally, the fact that it begins from a large seed vocabulary and iteratively reduces the vocabulary size may result in worse handling of rare or unseen words than BPE, which begins from character level and is therefore able to break down almost any unseen token, provided its consituent characters were present in the training corpus.

Morphological (Morfessor)

Although statistical methods have risen to prominence in NMT, there remains the possibility that morphological information is lost in these approaches. This may impact languages with richer morphology, or affect translation of specific features which are conveyed through inflection or affixation and which are not specifically assessed with current evaluation standard practice such as BLEU scores. Morphological methods of tokenisation 'are designed with the goal of "producing subword segments that are closely aligned to the true morphs constituting a word" (Park et al., 2021, p. 265), with the idea that these subwords will then contain the same or similar information across contexts.

The most commonly used morphological tokenization approaches are based on the recursive Minimum Defined Length (MDL) model Morfessor (Creutz & Lagus, 2002), updated to Morfessor 2.0 (Smit et al., 2014). These are generative probabilistic models which can be used either semi-supervised, in conjunction with annotated training data, or unsupervised, to segment the corpus (Smit et al., 2014). Although the 'morphs' generated will not have an exact correspondence with linguistically recognised morphemes (Creutz & Lagus, 2002), they are intended to be closer to a true morphological segmentation than other statistical methods.

There has been little evidence that Morfessor and other morphological-based methods of tokenisation are able to perform as well or better than data-driven methods (Mielke et al., 2021), although some studies suggest that in particular circumstances, such as low-resource or highly-agglutinative languages, they may offer benefits, particularly if either they are either semi-supervised or combined with a statistical method (Mielke et al., 2021; Park et al., 2021; Vania & Lopez, 2017).

ii. Models

Preprocessing & Training[1]

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 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. 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

1 All scripts and data are available at https://github.com/audreyvm/tfm_gender_bias

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.

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). The Transformer model is built on the concept of self-attention, which allows it to compute the relative significance of words within a sentence, producing more accurate translations. It has an encoder-decoder architecture, with both composed of multiple layers. The encoder contains an embedding table which is used to encode each token of the input. It also incorporates positional information. These embeddings are then passed into the self-attention mechanism and feedforward layers. This allows the model to focus on different parts of the input and in each layer decide which new information to preserve and which to discard.

In the decoder part of the Transformer model, the encoded representation of the source sentence is used to generate the translated output. The decoder also consists of multiple layers, similar to the encoder. At each layer of the decoder, self-attention is applied to the previously generated target tokens. This enables the model to attend to different parts of the translated sequence while considering the relationships between the target words. The self-attention mechanism in the decoder helps capture the dependencies between the generated tokens and facilitates the generation of accurate translations.

Following the attention mechanisms, the decoder utilises feedforward layers to process the attended representations and generate the final translations. These feedforward layers transform the information captured through attention into the desired output format, producing the translated sentence.

Our models have an embedding table with 512 dimensions, increased to 2048 in the 6 feed-forward layers. The models employ 8 attention heads.

iii. Training Corpus

The MT systems are trained 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 which was used to train the Transformer models consisted of 16,437,038 sentence pairs.

iv. Evaluation Datasets

The models were evaluated using two datasets. The first, the Flores 101 dataset (Goyal et al., 2021), serves as a benchmark specifically designed to evaluate the translation

performance of MT systems across 101 languages. It aims to address the need for multilingual evaluation and analysis of MT models in a diverse linguistic context. The dataset is named after "Flores," which stands for "FLOw of REpresentations for Multilingual Machine Translation." The Flores 101 dataset consists of 3001 sentences taken from the English Wikipedia and manually translated into the target languages by professional translators (Goyal et al., 2021). The BLEU score, which allows comparison of generalised translation quality with other models, is calculated on both the dev (997 sentences) and devtest (1012 sentences) splits of the Catalan Flores 101 corpus. This dataset is used purely for evaluating the overall quality of the output and does not provide any information regarding the performance in terms of gender bias.

The second dataset used is derived from MuST-SHE (Bentivogli et al., 2020), 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 ST systems, consisting of audio, transcript and translation triplets, with English as the source language. As we are only interested in MT for this project, we have discarded the audio and worked solely with the transcript and translation.

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 variable gendered translation in the target language. Sentences were selected so as to contain consistent gendering throughout.

These segments have been human-annotated with information relating to gender. All gender-marked expressions in the translations have been manually marked, and then an

alternative translation created which is identical except that the gender of all marked words has been changed for the opposite (Bentivogli et al., 2020). Each translation (both the 'reference' translation and the 'wrong-reference' translation) has also been annotated with a list of the gender-marked terms which it contains. Each entry in the text based dataset therefore contains a triplet of 'source', 'reference' and 'wrong reference'.

Creation of Catalan MuST-C Dataset

As MuST-SHE does not contain En-Ca data, it was necessary to create a synthetic Catalan dataset for evaluation of our models. To do so both the 'reference' and 'wrong-reference' 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 and was trained on an aggregated dataset of approximately 92 million sentences. It was evaluated across various domains and received an average BLEU score of 47.6.

Without resources to manually annotate all gender-terms present in the synthetic Catalan data, it was necessary to create a script to automatically extract them. This was done using the Catalan morphologizer available from spaCy (*Catalan · SpaCy Models Documentation*, n.d.). This pipeline parses strings (in this case individual tokens) and returns both the part of speech and morphological features following the Universal Dependencies annotations guidelines. Our script iterated through tuples of 'reference' and 'wrong reference' sentences to identify tokens which shared an index, were not identical, and both had the feature 'Gender'. All such pairs of tokens were extracted and appended to the gender terms list for that sentence pair.

130 sentences contained no gender terms in the Catalan translation. This was partly due to terms which take gender in Spanish but do not in Catalan, as in the case of *tonto/tonta* (es) vs *ximple* (ca) in Table 1. However, during analysis it was also discovered that some sentences had been truncated during the translation process, removing parts of the sentence which contained the target gender terms. This was identified as being due to a problem with the translation engine which has since been rectified, but there was insufficient time to reprocess the dataset. All triplets from which gender terms were not extracted were therefore discarded from the dataset at this stage, leaving 1034 triplets.

The heuristic used for the extraction of gender terms relied on the automatic translations of both sentences being identical in all other respects, which was not consistently the case. In cases where the translator had output differently gendered synonyms (i.e. *el meu treball / la meva feina* as translations of *mi trabajo*), non-target tokens were extracted. There were also cases where the phrasing of the two translations differed in the number of words used, throwing out the alignment of the two sentences and resulting in unconnected tokens being output as gender terms as can be seen in Table 1.

	English	Catalan Reference	Catalan Wrong Reference	Desired Gender Terms	Retrieved Gender Terms
Correctly Processed Triplet	The most ambitious and most competent leader on the international stage today is Chinese President Xi Jinping.	El líder més ambiciós i competent en l'escena internacional avui, és el president Xi Jinping.	La lideressa més ambiciosa i competent en l'escena internacional avui, és la presidenta Xi Jinping.	El La;líder lideressa;ambiciós ambiciosa;el la;president presidenta	El La;líder lideressa;ambiciós ambiciosa;el la;president presidenta
Incorrectly Processed Triplet (misaligned)	They had big muscles, supermodel good looks, and phenomenal cosmic powers. And me? I kind of looked like this, except shorter and with frizzier hair, and I never felt powerful.	Tenien grans músculs, eren atractives com a supermodels i posseïen poders còsmics fenomenals. I jo? Jo era així, només que més petita i amb el cabell més arrissat i mai em vaig sentir poderosa.	Tenien grans músculs, eren atractives com supermodels i posseïen poders còsmics fenomenals. I jo? Jo era així, només que més petit i amb el cabell més arrissat i mai em vaig sentir poderós.	petita petit;poderosa poderós	poders còsmics;el cabell
Triplet Without Gender Terms	I know what people think when they see this. They go, "Well, he's certainly not dumb enough to stab himself through the skin to entertain us for a few minutes. So, let me give you a little peek. How's that look out there?	sé el que la gent pensa quan veu això. diuen, "bé ell no és tan ximple com per tallar-se així mateix la pell tan sols per entretenir-nos per uns minuts." Així que, deixin-me mostrar-los una mica. Com van?	sé el que la gent pensa quan veu això. diuen, "bé ell no és tan ximple com per tallar-se així mateix la pell tan sols per entretenir-nos per uns minuts." Així que, deixin-me mostrar-los una mica.		

Table 1: Sample Processed Sentences

A manual revision of the gender terms output identified 76 cases where non-parallel tokens were extracted as gender terms. In these cases, the entire sentence was reviewed with the assistance of a native Catalan speaker. The translations were aligned, selecting the more natural phrasing, and the correct gender terms were manually extracted. During the manual revision it was also recognized that contracted word-forms had been split by spaCy and so they were changed back to the full token. This was the case when certain prepositions (e.g. *a* and *de*) were followed by the masculine determiner, creating the contractions *al* and *del*. spaCy had extracted only *l* into the gender terms, while we required the full word as our evaluation script splits on white space.

The resulting evaluation set consists of 1034 triplets of English sentences with their 'reference' and 'wrong-reference' Catalan translations. All other, non language specific annotations from the En-Es dataset have been preserved with the language tag changed to 'ca' where relevant.

In the original MuST-SHE corpus, all sentences are divided into 4 categories:

1) Gender information is not present in the sentence, but aligns with the speaker's preferred expression of gender, which can be obtained through vocal characteristics.
2) Gender information is present in the text, through either lexically gendered words, pronouns or proper names.
3) Gendered terms refer to two individuals and require that both gender identities are taken into account, combining textual context and vocal characteristics.
4) No gender-disambiguating information is present. (Savoldi, 2021)

As categories 1 and 3 rely on vocal characteristics to identify gender, when evaluating MT they are indistinguishable from category 4. We have therefore created separate categories for evaluating text:

1) No gender-disambiguating information is present *in the text* (an amalgamation of categories 1, 3 and 4 for ST)
2) Gender information is present in the text, through either lexically gendered words, pronouns or proper names. (This is unchanged)

The dataset is well-balanced between the two categories and between the number of male and female segments:

Category	Female	Male
1	234	245
2	278	278

Table 2: Distribution of Masculine and Feminine sentences between categories

The distribution is also reasonably well balanced in terms of gender terms per category as seen in Table 3. There are more gender terms present in category 2 than category 1, but in both categories there is no significant discrepancy between the number of male and female tagged sentences.

Category	Term Count
1F	407
1M	414
2F	475
2M	530

Table 3: Number of gender terms per category

v. Metrics

Evaluation of Translation Quality

BLEU scores (Papineni et al., 2002) are the most commonly used metric for evaluating and comparing MT output quality. They aim to provide a holistic comparison with a human created reference translation by comparing *n*-grams of the MT translation with *n*-grams of the reference and outputting a score between 0 and 1. Although these scores have become the industry standard for evaluating MT quality and are thus useful tools for making comparisons across models and systems, they do not provide any insight into the types of errors being made in the translations, and so for investigation of a specific phenomenon such as gender bias, require supplementation with additional metrics. BLEU scores for all models have been calculated on the Flores dev set, the Flores test set and the MuST-SHE data.

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. As the Catalan dataset has different categories for ST and MT, this has been added as an optional argument to the script.

During the results stage we discovered that the script was producing inaccurate results due to sentences containing multiple instances of the same term, but with different genders. This particularly affected common function words such as determiners. As originally written the script would search the sentence for the correct gender term. If it was found, it stored the index of the term to prevent duplication, but would still search the string for the wrong gender term. Where sentences contained both masculine and feminine forms of the same word, this led to results claiming that more gender terms had been found than were present in the sentence. We therefore amended the script to only search for the wrong gender term if the correct term had not been found.

We also identified an issue with word-forms followed by punctuation not being extracted as the script worked on whitespace separation. We amended this to strip following periods but in a future experiment would make further changes to take other punctuation such as commas and semi-colons into account.[2]

2 The orginal script is https://github.com/audreyvm/tfm_gender_bias/blob/main/mustshe_acc_v1.1.py while our amended script is https://github.com/audreyvm/tfm_gender_bias/blob/main/mustshe_acc_v1.2.py

4. Results and Discussion

i. Overall Translation Quality

BLEU SCORES				
	Uni	BPE	Char	Morf
Flores Dev	**41.6**	41.0	38.0	41.2
Flores DevTest	**42.0**	41.3	37.7	41.5
MuST-SHE	40.0	**40.1**	33.2	39.6
AVG	**41.2**	40.8	36.3	40.8

Table 4: BLEU Scores for all models

The BLEU scores, as depicted in Table 4, serve as informative indicators of the overall translation quality achieved by the machine translation (MT) models trained using distinct tokenization methods. Notably, the unigram method achieves the highest score, with a BLEU of 41.2, closely followed by both BPE and Morfessor, which both score 40.8. In contrast, character-based tokenization exhibits a notable performance gap, with a BLEU score of 36.3.

The better outcomes obtained by the Unigram and BPE models align with their status as the current state-of-the-art tokenization approaches in machine translation. Their use of sub-words to capture and retain semantic information manifests in superior translation quality. Conversely, the lack of semantic information captured in character-level tokenisation undermines its translation performance.

While Morfessor achieves a performance level similar to 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 balance between translation quality and computational efficiency.

If we are only taking into account the BLEU score, it appears that the unigram tokenization method emerges as the best approach. Its superiority in terms of BLEU scores underscores its capacity to generate accurate and contextually appropriate translations, thus justifying its place as the optimal tokenization method. However, although the BLEU score is a useful tool of comparison between translation methods, it does not give us insight into the specific areas in which a method performs well or poorly. To understand the gender-specific performance of the tokenization methods we need to look more closely at the output.

ii. Term and Part of Speech Coverage

Only target gender terms which appear in the translated output can be evaluated for coverage. 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.

Predictably, the levels of term coverage exhibit a strong correlation with the BLEU scores. The unigram tokenization method achieves the highest level of coverage, with 63.6% (1164 tokens out of a possible total of 1831). BPE and Morfessor are close behind with coverage rates of 63.2% (1159) and 63.1% (1155) respectively. On the other hand, character-based tokenization, reflecting its overall weaker performance, displays a lower level of term coverage at 61.1% (1118).

To better understand the alignment of word choice in the translated output compared to the reference text, all gender terms were subjected to part-of-speech tagging using the Catalan

spaCy morphologizer. The evaluation of term coverage considered both the masculine and feminine forms, regardless of the desired output; in other words if a term was expected in the feminine form but it was found in the masculine form, it was still considered found. Across all tokenization methods, nouns, adjectives, and verbs displayed the lowest probability of occurrence. This trend can be ascribed to the greater lexical synonymy available for these categories compared to grammatical and function words.

The output of the morphologizer does not exhibit exact alignment with the gender-specific terms due to spaCy's contraction splitting mechanism, as discussed in Section 3 iv. The adpositions *a* and *d-* were discarded from our analysis. There were also two terms (*senadora*, *prim*) which were miscategorised as prepositions and one (*psiquiatra*) which was miscategorised as an adverb. These have also been removed from the below table.

	BPE		UNI		CHAR		MORF	
	Found	Not Found	Found	Not Found	Found	Not Found	Found	Not Found
ADJ	322	253	367	255	385	281	375	268
AUX	1	1	2	2	1	1	1	1
DET	310	77	143	80	162	82	139	82
NOUN	300	272	355	265	382	283	359	256
NUM	8	1	4	1	3	0	4	1
PRON	177	41	46	36	49	33	53	37
PROPN	11	0	1	0	3	3	2	1
VERB	25	27	38	28	38	29	37	29

Table 5: POS counts for all models

iii. Gender Accuracy with Context

First we look at the results of gender accuracy, as calculated through the F1 score, for sentences where all necessary gender information is available in the reference text. These results, as presented in Table 5, reveal an inherent bias in the translation models. All models show a substantial disparity in translation accuracy between feminine and masculine forms, with a tendency across the board to generate masculine forms more frequently than their feminine counterparts. With all F1 scores for masculine gender references in the high 90% range, it seems that the models have a greater ability to preserve the knowledge of the intended masculine gender. Despite the availability of contextual information, F1 scores which range from 65.26% to 72.79% mean that the models are not able to preserve knowledge of feminine gender in the same way.

		Uni	**BPE**	**Char**	**Morf**
F1 scores for gender translation with provided context	F	**72.79%**	72.04%	65.26%	71.38%
	M	96.31%	**96.82%**	95.26%	97.11%
	Overall	**84.55%**	84.43%	80.26%	84.25%

Table 5: Gender Accuracy (F1 Scores)

Among the tokenization methods employed, the unigram approach demonstrates the highest level of accuracy, surpassing both BPE and Morfessor in producing accurate translations for both feminine and masculine forms. With an overall gender translation accuracy of 84.55%, the unigram method performs slightly better than BPE, which achieves 84.43%, and Morfessor, which achieves 84.25%. On the other hand, the character-based tokenization method exhibits the lowest performance, with an accuracy rate of 80.26%.

As we had made substantial amendments to the original evaluation script, for comparability we are also reporting the results without those changes. With the original script we saw a smaller difference between Character and the other methods, but it still came out as the worst performing method. The most noticeable difference here is that before our changes, BPE performed better that Unigram in accuracy for both feminine and masculine terms. However, with both metrics the difference between these two methods is slight and cannot be taken as conclusive evidence of the superiority of one method over the other.

		Uni	BPE	Char	Morf
Accuracy of gender translation with provided context	F	63.61%	**64.33%**	58.77%	62.58%
	M	83.61%	**83.76%**	84.52%	84.15%
	Overall	73.61%	**74.05%**	71.64%	73.36%

Table 6: Gender Accuracy (unamended evaluation script)

The poor performance of character based tokenization in NMT is the opposite of the results obtained by Gaido et al. (2021) in ST. We hypothesised that this would be the case due to the lack of semantic information contained in individual characters. A possible explanation of the good performance of Unigram and BPE is the fact that their methods of sub-word tokenization may result in tokens which reflect morphological boundaries despite the fact that they are not explicitly designed to capture semantic information. However, this leaves unexplained the slightly poorer performance of the explicitly morphologically based tokenization method, Morfessor. Further investigation would be necessary to understand the reasons behind the slightly lower performance of the Morfessor method compared to Unigram and BPE, but the difference in performance here was so slight that a manual investigation of the output was not able to reveal any systematic differences.

iv. Context-Free Gender

The second category we investigate is the frequency with which our models produce feminine and masculine forms when no overt indication of gender is given in the source text. In these "context-free" sentences, all models predominantly generate masculine forms. The Morfessor model shows the highest percentage of feminine forms at 18.66%, while the Unigram model exhibits the lowest, producing feminine forms only 13.96% of the time.

	Uni		BPE		Char		Morf	
Female	3.96%	67	5.48%	74	4.98%	71	**8.66%**	**89**
Male	86.04%	413	84.52%	404	85.02%	403	81.34%	388

Table 7: Generation of Feminine/Masculine terms with no provided context

An initial investigation of these ‘context free’ sentences revealed a limitation of the evaluation script. In many cases the feminine gender term which had purportedly been generated was a feminine determiner. However, the script mistakenly identified another feminine determiner in the sentence, while the intended target determiner was either not generated or produced as masculine. This issue with the script's evaluation led to inaccurate counts of feminine target terms in the output. Resolving this issue would present considerable challenges and was not possible in the scope of this project.

As the number of feminine forms generated by each tokenization method was so low - ranging from 67 for Unigram to 89 for Morfessor - once false feminines from incorrectly identified determiners had been removed from the data, there were insufficient terms remaining to identify patterns in the data. The sample size is also too small to be taken as conclusive evidence that Morfessor does produce more feminine forms when no context is provided. This would be an interesting area for further investigation.

One clear trend that did emerge from the data is the fact that the feminine *infermera* or *infermeres* was generated in every context that 'nurse' occurred in the source text, across all tokenisation methods, despite the fact that for 7 out of 9 of those occurrences no gender disambiguation was present in the source text. *Professora* also occurred as feminine with relative frequency when no context was provided, although still substantially less than it did as male. This reinforces the fact that although architectural changes to models can reduce gender bias, as long as it remains present in the training data it cannot be eliminated.

5. Future Work

In order to further advance the field of gender-aware NMT and build upon the findings of this study, there are several potential areas for future exploration and improvement.

Firstly, further work is needed to finalize the Catalan version of the MuST-SHE dataset created in this study. Re-translation using the latest version of the es-ca model, which has been re-trained to address the truncation issues we experienced, is necessary. This process will involve manual revision of the dataset to ensure accurate extraction and alignment of gender terms. Once completed, the dataset will be made freely available for use as an evaluation corpus, facilitating future research in gender-aware machine translation.

Once that dataset has been created, further work to investigate the impact of different vocabulary sizes in the tokenization could be warranted. We initially hoped to explore different vocabulary sizes for BPE and Unigram, but were limited by time constraints. Changes to the vocabulary size might impact the amount of morphological information captured by the tokens.

It would be useful for future research to focus on a more fine-grained investigation of gender bias in machine translation. While this study examined gender accuracy and bias at a

general level, a more carefully curated evaluation dataset or fully reworked evaluation script could avoid the problems we encountered with misidentified target terms, allowing researchers to delve deeper into the translation of specific gendered terms or expressions. This could provide insights into the underlying biases present in MT systems. Analyzing how different tokenization methods affect the translation of these terms can guide the development of more inclusive and unbiased models.

Finally, future research should investigate the use of gender-neutral language as a strategy to mitigate gender bias in machine translation. Adapting translation models to recognize and employ gender-neutral alternatives when translating gendered terms can help promote inclusive and unbiased language use. Our research, and the majority of other work in the field, has maintained a strict gender binary, but moving beyond this can help contribute to the creation of machine translation systems that actively support gender-inclusive communication and contribute to the broader societal goals of equality and non-discrimination.

6. Conclusion

In conclusion, the present research contributes to the field of NMT by investigating the impact of different tokenization methods on gender bias. Our study builds upon previous research conducted by Gaido et al. (2021) in ST and explores the effects of tokenization approaches on translation quality and gender accuracy in text-based MT.

Gaido et al.'s study focused on subword tokenization methods and their influence on gender bias in ST. Their findings challenged the assumption that the best tokenization method for addressing gender bias aligns with overall translation quality, as Character Segmentation demonstrated the best gender accuracy despite lower BLEU scores. However, it is important

to note that Speech Translation and Machine Translation have fundamental differences in terms of input data and task requirements.

In our research, we specifically investigated the performance of Unigram, BPE, Character, and Morfessor tokenization methods in the context of MT. Our comprehensive analysis revealed that Unigram achieved the highest translation quality, closely followed by BPE and Morfessor, while Character-based tokenization exhibited weaker performance. Regarding gender accuracy, all models exhibited a tendency to generate masculine forms more frequently. Unigram showed the highest accuracy for both feminine and masculine forms, surpassing BPE and Morfessor, while Character-based tokenization performed the least effectively. Architectural changes can reduce gender bias, but the persistence of bias in training data highlights the need for ongoing efforts to address gender imbalance.

Comparing our findings with Gaido et al.'s research, we observed that the optimal tokenization method for addressing gender bias in MT differs from that in ST. The limited semantic information conveyed by individual characters in MT challenges the effectiveness of Character-based tokenization, whereas it proves more beneficial in ST due to its ability to capture phonetic patterns. These distinctions emphasize the importance of domain-specific investigations and the need to tailor tokenization methods to the unique requirements of each task.

By expanding our understanding of the relationship between tokenization methods, translation quality, and gender accuracy in MT, we can strive towards developing more inclusive and unbiased translation systems. Further research in this area is warranted to explore additional factors that may influence gender bias. Ultimately, this research contributes to advancing the field of MT and promotes the development of ethically sound and culturally sensitive translation technologies.

Despite limitations in the analysis of context-free gender, the consistent bias in generating feminine forms without explicit context warrants further investigation. These findings emphasize the complex interplay between tokenization, translation quality, and gender bias, underscoring the importance of adopting tokenization methods that capture semantic information and striving for gender-neutral translations in NMT systems.

7. References

Ahmed, Md. A., Chatterjee, M., Dadure, P., & Pakray, P. (2022). The role of biased data in computerized gender discrimination. *Proceedings of the Third Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering*, 6–11. https://doi.org/10.1145/3524501.3527599

Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., & Turchi, M. (2020). Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6923–6933. https://doi.org/10.18653/v1/2020.acl-main.619

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). *Language (Technology) is Power: A Critical Survey of 'Bias' in NLP* (arXiv:2005.14050). arXiv. http://arxiv.org/abs/2005.14050

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (arXiv:1607.06520). arXiv. https://doi.org/10.48550/arXiv.1607.06520

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

https://doi.org/10.1126/science.aal4230

Catalan · spaCy Models Documentation. (n.d.). Catalan. Retrieved 15 June 2023, from https://spacy.io/models/ca

Corbett, G. G. (2014a). Introduction. In G. G. Corbett (Ed.), *The Expression of Gender*. De Gruyter Mouton.

Corbett, G. G. (Ed.). (2014b). *The expression of gender*. De Gruyter Mouton.

Costa-jussà, M. R., & de Jorge, A. (2020). Fine-tuning Neural Machine Translation on Gender-Balanced Datasets. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 26–34. https://aclanthology.org/2020.gebnlp-1.3

Costa-jussà, M. R., Escolano, C., Basta, C., Ferrando, J., Batlle, R., & Kharitonova, K. (2020). *Gender Bias in Multilingual Neural Machine Translation: The Architecture Matters* (arXiv:2012.13176). arXiv. https://doi.org/10.48550/arXiv.2012.13176

Costa-jussà, M. R., & Fonollosa, J. A. R. (2016). Character-based Neural Machine Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 357–361. https://doi.org/10.18653/v1/P16-2058

Creutz, M., & Lagus, K. (2002). *Unsupervised Discovery of Morphemes* (arXiv:cs/0205057). arXiv. https://doi.org/10.48550/arXiv.cs/0205057

Deery, O., & Bailey, K. (2022). The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing. *Feminist Philosophy Quarterly*, *8*(3/4), Article 3/4. https://ojs.lib.uwo.ca/index.php/fpq/article/view/14292

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., & Turchi, M. (2019). MuST-C: A Multilingual Speech Translation Corpus. *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), 2012–2017. https://doi.org/10.18653/v1/N19-1202

Domingo, M., Garcıa-Martınez, M., Helle, A., Casacuberta, F., & Herranz, M. (2019). *How Much Does Tokenization Affect Neural Machine Translation?* (arXiv:1812.08621). arXiv. https://doi.org/10.48550/arXiv.1812.08621

Elaraby, M., Tawfik, A. Y., Khaled, M., Hassan, H., & Osama, A. (2018). Gender aware spoken language translation applied to English-Arabic. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 1–6. https://doi.org/10.1109/ICNLSP.2018.8374387

Escudé Font, J., & Costa-jussà, M. R. (2019). Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. In *ArXiv e-prints*. https://doi.org/10.48550/arXiv.1901.03116

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3).

Gaido, M., Savoldi, B., Bentivogli, L., Negri, M., & Turchi, M. (2021). *How to Split: The Effect of Word Segmentation on Gender Bias in Speech Translation* (arXiv:2105.13782). arXiv. https://doi.org/10.48550/arXiv.2105.13782

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., & Fan, A. (2021). *The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation* (arXiv:2106.03193). arXiv. https://doi.org/10.48550/arXiv.2106.03193

Hutchins, W. J. (2004). The Georgetown-IBM experiment demonstrated in January 1954. In R. E. Frederking (Ed.), *Machine Translation: From Real Users to Research* (Vol. 3265, pp. 208–216). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-

30194-3_23

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing (3rd ed. Draft)*. https://web.stanford.edu/~jurafsky/slp3/

Kudo, T. (2018). *Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates* (arXiv:1804.10959). arXiv. https://doi.org/10.48550/arXiv.1804.10959

Libovický, J., Schmid, H., & Fraser, A. (2022). *Why don't people use character-level machine translation?* (arXiv:2110.08191). arXiv. https://doi.org/10.48550/arXiv.2110.08191

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, *40*(3), 1–49. https://doi.org/10.1145/1380584.1380586

McConnell-Ginet, S. (2013). ` Gender and its relation to sex: The myth of 'natural' gender. In ` *Gender and its relation to sex: The myth of 'natural' gender* (pp. 3–38). De Gruyter Mouton. https://doi.org/10.1515/9783110307337.3

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP* (arXiv:2112.10508). arXiv. http://arxiv.org/abs/2112.10508

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Moran, C. (2023, April 6). ChatGPT is making up fake Guardian articles. Here's how we're responding. *The Guardian*. https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-

technology-risks-fake-article

Mukherjee, S., Pollina, E., & More, R. (2023, April 3). Italy's ChatGPT ban attracts EU privacy regulators. *Reuters*. https://www.reuters.com/technology/germany-principle-could-block-chat-gpt-if-needed-data-protection-chief-2023-04-03/

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., & Schwartz, L. (2021). Morphology Matters: A Multilingual Language Modeling Analysis. *Transactions of the Association for Computational Linguistics*, *9*, 261–276. https://doi.org/10.1162/tacl_a_00365

PlanTL-GOB-ES/mt-plantl-es-ca · Hugging Face. (n.d.). Retrieved 15 June 2023, from https://huggingface.co/PlanTL-GOB-ES/mt-plantl-es-ca

Romaine, S. (2001). English: A corpus-based view of gender in British and American English. In M. Hellinger & H. Bußmann (Eds.), *Gender Across Languages*. J. Benjamins.

Savoldi, B. (2021). *Data Statement for MuST-SHE*.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). *Gender Bias in Machine Translation* (arXiv:2104.06001). arXiv. http://arxiv.org/abs/2104.06001

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 21–24. https://doi.org/10.3115/v1/E14-2006

Stafanovičs, A., Bergmanis, T., & Pinnis, M. (2020). Mitigating Gender Bias in Machine Translation with Target Gender Annotations. *Proceedings of the Fifth Conference on Machine Translation*, 629–638. https://aclanthology.org/2020.wmt-1.73

Stanczak, K., & Augenstein, I. (2021). *A Survey on Gender Bias in Natural Language Processing* (arXiv:2112.14168). arXiv. http://arxiv.org/abs/2112.14168

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. https://doi.org/10.18653/v1/P19-1164

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. https://doi.org/10.18653/v1/P19-1159

Sun, T., Webster, K., Shah, A., Wang, W. Y., & Johnson, M. (2021). *They, Them, Theirs: Rewriting with Gender-Neutral English* (arXiv:2102.06788). arXiv. https://doi.org/10.48550/arXiv.2102.06788

Vania, C., & Lopez, A. (2017). *From Characters to Words to in Between: Do We Capture Morphology?* (arXiv:1704.08352). arXiv. http://arxiv.org/abs/1704.08352

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting Gender Right in Neural Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in*

Natural Language Processing, 3003–3008. https://doi.org/10.18653/v1/D18-1334

Vanmassenhove, E., Shterionov, D., & Way, A. (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. *Proceedings of Machine Translation Summit XVII: Research Track*, 222–232. https://aclanthology.org/W19-6622

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & SOCIETY*, *36*(2), 585–595. https://doi.org/10.1007/s00146-020-01066-z

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in Machine Translation. *Engineering*, *18*, 143–153. https://doi.org/10.1016/j.eng.2021.03.023

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. https://doi.org/10.18653/v1/D17-1323

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). *Learning Gender-Neutral Word Embeddings* (arXiv:1809.01496). arXiv. http://arxiv.org/abs/1809.01496