The Exacerbation of (Grammatical) Gender Stereotypes in English–Spanish Machine Translation

Nerea Ondoño-Soler
Mikel L. Forcada

revista.tradumàtica
tecnologies de la traducció

Nerea Ondoño-Soler
Universitat d'Alacant;
nereaondonosoler@gmail.com

Mikel L. Forcada
Universitat d'Alacant
mlf@ua.es

Abstract

The information required to select grammatical gender in machine translation of isolated sentences for gender-marking languages is frequently missing or difficult to extract. Our text-centric, black-box study demonstrates how the gender distribution of the training set is distorted at the output. Human evaluation reveals that gender clues are frequently absent from the source, resulting in stereotyped translations.

Keywords: machine translation, Spanish, English, grammatical gender, gender distribution distortion.

Resum

En la traducció automàtica d'oracions aïllades a llengües que marquen el gènere, la informació necessària per seleccionar el gènere gramatical sovint n'és absent o és difícil d'extraure'n. Aquest estudi de caixa negra i centrat en el text mostra com la distribució de gènere del conjunt d'entrenament es distorsiona en l'eixida. Una avaluació humana revela que sovint no hi ha cap pista de gènere en l'original, cosa que condueix a traduccions estereotipades.

Paraules clau: traducció automàtica, castellà, anglés, gènere gramatical, distorsió de la distribució de gènere.

Resumen

La información necesaria para seleccionar el género gramatical en la traducción automática de oraciones aisladas a lenguas que marcan el género suele estar ausente o ser difícil de extraer. Este estudio de caja negra y centrado en el texto demuestra cómo la distribución de género del conjunto de entrenamiento se distorsiona en la salida. Una evaluación humana revela que los indicios de género suelen estar ausentes en el original, lo cual da lugar a traducciones estereotipadas.

Palabras clave: traducción automática, español, inglés, género gramatical, distorsión de la distribución de género.

Rebuda: 18 de juliol de 2022 / Acceptació: 24 de novembre de 2022 / Publicació avançada: 15 de diciembre de 2022

Revista Tradumàtica. Número 20 (desembre 2022).
Cita recomanada: ONDOÑO-SOLER, NEREA; FORCADA, MIKEL L. (2022). «The Exacerbation of (Grammatical) Gender Stereotypes in English–Spanish Machine Translation». Revista Tradumàtica. Tecnologies de la Traducció, 20, 177-196. <https://doi.org/10.5565/rev/tradumatica.307>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

1. Introduction

1.1 Grammatical gender across languages

Many languages, like Spanish, French or German, use grammatical genders to classify their nouns; though generally the masculine and feminine grammatical genders, where present, correspond to men and women (or male and female animals) respectively, in some cases they do not: in German, *Mädchen* ('young lady') is grammatically neuter; in French, *sentinelle* ('sentinel', 'guard') is grammatically feminine regardless of the gender of the person. On the other hand, there are completely genderless languages, such as Turkish or Basque, or languages where there is one gender or word class for all people (Swahili). English presents grammatical gender only in 3rd person singular pronouns and possessives, showing a clear semantic alignment with personal genders. Readers are referred to Savoldi et al. (2021, §3.1.1) for a linguistic discussion of the encoding of gender across languages.

1.2 Grammatical gender as a challenge in professional and machine translation

Translation from a language that does not mark personal gender as grammatical gender into a language that does is challenging when context is insufficient, as when translating isolated sentences independently.

If the selection of grammatical gender in the target language is already a challenge for professional translators when the source language shows no grammatical gender or a weak grammatical gender system as English does, it comes as no surprise that machine translation (MT) faces the same challenge, or even harder, considering the fact that most state-of-the-art MT corpus-based systems are trained on corpora which are bags of shuffled sentence pairs without any context, and, once trained, translate each new sentence in isolation.

1.3 Social gender semantics and grammatical gender: encoding and decoding

In languages that mark gender, the encoding of personal or social gender as grammatical gender may vary. As said above, masculine forms are generally used to denote people of masculine personal gender, and feminine forms for people of feminine personal gender (this is indeed the reason for the naming of grammatical genders). This is generally true for nouns designating professions. Note that, to make things a bit more complicated, in some languages women designate their profession using the masculine grammatical gender (Spanish *médico*, masculine, 'doctor'), sometimes with a feminine agreement (*la médico*, 'the doctor' with a feminine article *la*). But importantly, and partially as the result of the distribution of gender roles in society, *traditional* encodings tend to use the grammatical masculine as the default to refer to people of both masculine and feminine gender. On top of that, the evolution of women in the workplace means, for example, that, in older texts, women will appear more often than in recent texts as *housewives* or non-wage-earning *caretakers* than as another worker in society.

In an attempt to use language to change this distribution of gender roles in society, a number of alternative gender encodings have emerged and are currently being applied (they are usually called *gender-inclusive language*).

As a result of the coexistence of different gender encoding styles, mixtures of them may be observed in the large corpora used to train machine translation systems, and quite often it is impossible to infer which gender encoding style was used in each isolated sentence.

There is also a problem with the representativity of corpora. Even if the encoding style were fixed, and again as a result of the distribution and visibility of gender roles in society, the distribution of (encoded) personal gender in the texts found in corpora cannot be expected to represent the demographics of personal genders.

Finally, there may be a mismatch between encoding during writing and decoding during reading. The semantics of gender decoding by a specific reader (*perceived* personal gender) may or may not match the semantics of gender encoding of a specific writer or translator (*intended* personal gender). For example, the word *profesores* ('teachers') could have been used to encode a mixture of male and female teachers by someone using the traditional *masculine-as-generic* encoding, but could be decoded as just male teachers (no one female) by a reader used to gender-inclusive doublets (*profesores y profesoras*).

Grammatical gender selection by modern corpus-based MT systems, a subject of debate in society as a result of the popularization of online MT, has also been the subject of various studies that have usually placed this in the framework of *gender bias*, or *gender discrimination*, which is clearly a socially relevant issue, and which is, in turn, expectedly encoded in written language; related work is discussed in section 1.6.

1.4 A text-centric study

While necessarily acknowledging (a) the relevance of the intended or perceived personal gender semantics of grammatical gender in text under either traditional or gender-inclusive encoding styles, and (b) that either usage may still result in the representation in texts of gender biases and inequalities occurring in society, this work will therefore focus on grammatical gender as observed in the texts themselves, as this is what translation (and MT) materially deals with and can distinctly be observed and studied independently of semantic processes and their social implications. Therefore, we do not deal with the mitigation of any possible social harm caused by any shortcomings in grammatical gender selection by modifying the behaviour of the MT system to produce an output that tries to conform to commonly adopted language guidelines such as gender-inclusive language or avoid gender stereotypes. This is in stark contrast with other work published (see section 1.6), where both issues (encoding styles and gender semantics on the one hand and observed grammatical gender on the other) are not clearly separated.

1.5 Research hypotheses

Hypothesis H1: Machine translation systems translating from languages with little or no grammatical gender, such as English, into languages with grammatical gender, such as Spanish, are unable to even reproduce the distribution of grammatical gender choices observed in the translations used to train them. As a result, they usually fall back to the most frequently observed grammatical gender therefore, the distribution of grammatical gender in machine-translated text is a distorted version of the distribution of grammatical gender in the translations used to train the system, where grammatical gender imbalances are exacerbated. This may be due to a number of factors:

a) Machine translation systems are trained with sentence pairs which have been detached from their context in a larger text, so the source may not provide sufficient information for the grammatical gender choice observed in the target. In these cases, the system simply would end up learning the global distribution of grammatical gender in the target side of these sentence pairs.
b) Even when sufficient context is present in those isolated training examples, machine translation systems may not be able to learn the clues that led to that choice.
c) Once trained, machine translation systems also process their input sentence by sentence; if the source sentence does not provide the context for the target-language grammatical gender choice (as in (a) above), the system may simply *guess* and reproduce the global distribution of grammatical gender for that particular item by falling back to the most frequent one.
d) But even when the context is present in the isolated sentence (as in (b) above), the system may not succeed in identifying a clue it has learned to use to generate grammatical gender.

Phenomena (b) and (d) are closely related to shortcomings in the functioning of attention in state-of-the-art transformer-based (Vaswani et al. 2017) neural machine translation (NMT): when an NMT system predicts the next token of the target sentence being generated (and therefore its gender), it (ideally) does so by paying attention to the relevant tokens in the whole source sentence as well as to the relevant preceding target tokens already consolidated: training of the NMT system to attend to existing gender cues may be incomplete, partly due to not having enough examples (a possible reason for (b) above) or because the system may not be able to exploit unseen cues or modified versions of seen cues (case (d) above). Falling back to the most frequent gender is also a likely outcome of this lack of attention.

Hypothesis H2: The lack of sufficient context inside each sentence to decide the target-side grammatical gender (phenomena (a) and (c) above) does indeed contribute to the distortion of grammatical gender distributions. To prove this, we will use an indirect way to ask translators whether the source sentence in a training example contains enough context for the target-language grammatical gender choice made in the target side of the example. This part of the study does not try to deal with the shortcomings mentioned in (b) and (d) above.

1.6 Critical review of related work

Many recent papers have dealt with phenomena labelled as *gender bias in machine translation*; we will not try to review all of them, particularly as a very complete review was recently published (Savoldi et al. 2021); we will distinguish basically two groups of papers: those dealing with the *observation* of gender bias, and those dealing with *mitigation* strategies. Note that our article falls clearly into the first group. The review ends with a brief mention of work about a related phenomenon: the *loss of lexical diversity*.

Observing gender bias: Among the *observation* papers, there are a few that have received a lot of attention. In many of them, gender-ambiguous sentences are translated to a language which is more gender-marked and in which this ambiguity has to be solved and use sentences with professions, as we do. In a recent paper, Prates et al. (2020) studied how the Google MT system rendered the grammatical gender of professions when translating a challenge set of purposely-built gender-ambiguous source sentences involving professions in a non-gender-marking language –such as Hungarian, Chinese or Yoruba– of the form "[3rd person singular pronoun] is a/an [profession noun]") to a weakly-gender-marking language (English), and compared the statistics of the grammatical gender of 3rd person singular pronouns in MT output with the actual statistics about men and women practising those professions in the United States of America, as observed in the U.S. Bureau of Labor Statistics studies. Prates et al. (2020) seem to disregard that the reality of gender distribution in professions according to the statistics provided by the U.S. Bureau of Labor Statistics may radically differ from the reality encoded in the training corpus, made of a collection of texts that might have been written years or even decades ago under different gender encoding styles (*masculine-as-generic* or *gender-inclusive*), and probe how the output deviates from these statistics.

Prates et al. (2020) are not alone in the use of synthetic probes to study gender stereotyping in MT: for instance, Rescigno et al. (2021) use sentences similar to those of Prates et al. (2020), but do not compare the output with statistics of any kind.

Other work studies sentences which are not necessarily ambiguous in isolation. Renduchintala et al. (2021) machine-translate a set of purposely-built sentences in English and find that even if the sentence in isolation is clearly unambiguous (*That nurse is a funny man*), gender stereotypes still sometimes emerge (*Esa enfermera es un tipo gracioso*). Stanovsky et al. (2019) translate unambiguous English sentences that include pronouns and antecedents and cast participants into non-stereotypical gender roles such as, *The doctor asked the nurse to help her in the operation*, and observe their translation into a gender-marking language.

It is worth mentioning that Gonen and Webster (2020) take a different approach, which is in some respects similar to ours, particularly as they focus on grammatical gender distributions. Instead of using lists, they use a neural named-entity recognizer to mine a corpus for English sentences containing non-gendered designations of people (not only professions), use a neural gap-filling method to find the 10 most likely substitutions of them, run them through Google Translate, and select those in which the

grammatical gender changes in the target as probes for nouns *at risk* in a challenge dataset. As we do (see section 2.1), they also perform human filtering to discard erroneous examples and study the distribution of grammatical gender in their translations.

One common limitation of all these studies is that they do not take into account the corpus used to train the MT system, which limits the understanding of the decisions made by the MT system. The use of simple synthetic sentences (except in the case of Gonen and Websters, 2020) without context is not representative of the contents of the training corpus. Even though these simple examples might exist, they will be a minor and isolated sample of what the training corpus is made of: corpora have more informative, more complex, and longer sentences as well. It is also important to note that these short, ambiguous sentences out of context would also be a challenge for the most experienced professional translators.

The study of gender bias in these works may also be affected by the fact that different gender encoding styles (see section 1.4) may coexist in the corpus; in particular, the *traditional* or *masculine-as-generic* style may be present. Using a *gender-inclusive* decoding of MT output may lead authors to unfairly *blame* machine translation for gender biases. It is, however, clearly reasonable to expect that social gender biases were encoded (through grammatical gender, but not only) in training texts when they were written and translated; therefore, the actual distribution of grammatical gender in the corpus is one source of the grammatical gender distribution observed in MT output. But, clearly, there is another source: the actual way in which the MT system is trained and used to translate (see hypothesis H1 in Section 1.5 above).

To avoid this, our study will deal with grammatical-gender bias *as observed in text*, regardless of gender encoding or decoding styles, and will use text from the training corpus to make sure that probe sentences are representative of the material the MT system has been trained on.

Mitigating gender bias: Mitigation is the subject of a growing body of active research (Saunders and Byrne, 2020; Štafanovičs et al. 2020; Tomalin et al., 2021; Vanmassenhove et al., 2018); for a systematic review of gender bias in MT and mitigating strategies, the reader is referred to Savoldi et al.'s (2021) work. Some strategies either add information to the input sentence during training or translation; for instance, Basta et al. (2020) observe a beneficial effect of adding the previous sentence and annotating the input with speaker information (an information that would not be available in general text-only sentence-by-sentence usage); Vanmassenhove et al. (2018) followed a similar gender-tagging approach; and Štafanovičs et al. (2020) show an improvement when randomly annotating the input with target gender during training (the annotation is not necessary during translation). Other approaches (Saunders and Byrne, 2020; Tomalin et al. 2021) involve retraining the system using handcrafted, de-biased sets to "domain-adapt" it.

Loss of lexical diversity and gender bias: Most state-of-the-art neural MT systems are trained to translate by showing them large corpora with pairs containing a source sentence and its translation, but usually shuffled and deprived of any supra-sentential or document-level context. This means that all the clues necessary to make decisions about possible translation equivalents (grammatical gender or otherwise) in the target

text should be present in the source sentence (and nowhere else), and that the system should be able to learn how to extract them and pay attention to them when producing the target words. If either the source sentence does not contain the information needed or the trained system does not manage to discover it and turn it into the adequate translation choice, it will tend to fall back to the most likely translation equivalent observed in the corpus (cfr. the discussion of hypothesis H1 in section 1.5). This leads to a loss of lexical diversity in the output of the MT system as compared to that observed in the training corpus (e.g., falling back to using more often the most frequent noun to name a specific profession, like *profesor* for *teacher* in English, instead of other less frequent choices, such as *maestro*).

This fall-back process may then result in a loss of lexical richness or diversity when a source word can be translated in more than one way (Vanmassenhove et al. 2019). Lexical simplification is, indeed, also observed in the work of professional translators as part of *translationese*, the language subset of translated texts that can be distinguished from natively produced source text in the same language (Volansky et al. 2015). Toral (2019) actually observed how professionally postedited MT output showed a special kind of *translationese*, which he called *posteditese*, where loss of lexical diversity was clearly observed.

The distortion of grammatical gender statistics may therefore be framed as a specific case of loss of lexical diversity: *doctor* may be translated as *doctor* or *doctora* depending on context, much in the same way as *cup* may be translated as *copa* or *taza* also depending on context. In the former case, solutions that may be expectedly interpreted as gender stereotyping could therefore be explained through fall-back in the absence of source-side cues or limitations of the system in retrieving and using them.

1.7 Paper structure

The rest of this article is organized as follows: Section 2 describes the methodology used to automatically analyse the exacerbation of the observed grammatical gender imbalances in nouns denoting professions using a novel but simple text-centric method that does not use any external information (section 2.1) and a series of experiments with (human) translators to ascertain whether the sentence pairs in the corpus contain enough information to make an adequate decision as regards the grammatical gender of profession nouns in the target language (section 2.2). Section 3 describes and discusses the results of the automatic (3.1) and manual (3.2) experiments. Conclusions are given and future work is outlined in section 4.

2 Methodology

To assess the extent of the exacerbation of grammatical gender statistics and to try to understand the underlying mechanisms, two experiments were designed: an automatic analysis of grammatical gender exacerbation and an annotation study with translation professionals. The first one tries (Section 2.1) to prove hypothesis H1 and the second one tries to prove hypothesis H2 (see section 1.5)

2.1 Automatic analysis of grammatical gender exacerbation

This section describes a method to prove Hypothesis H1, namely, that machine translation systems translating from languages with little or no grammatical gender into languages with grammatical gender are unable to even reproduce the target-side grammatical gender distribution observed in the corpus used to train them.

To perform a realistic evaluation of changes in grammatical gender statistics, we use a publicly available neural MT system and the corpus that was used to train it. The system is the transformer-based Helsinki-NLP OPUS-EN-ES English–Spanish system [1] provided through the Hugging Face[2] website, and it is locally executed using Hugging Face's Python library transformers;[3] the corpora used to train it are also available.[4] The Helsinki NLP setting makes it particularly easy for the experiments in this article to be reproduced in a variety of languages. This language pair was chosen for two reasons: firstly, MT is expected to work well with these two languages as they are both very common and well-resourced, and secondly, English is a clear example of a language without grammatical gender (except for third-person singular pronouns and possessives), whereas Spanish is clearly a language that marks grammatical gender. As in other work (Renduchintala et al., 2021; Prates et al., 2020; Saunders and Byrne 2020; Stanovsky et al., 2019; see section 1.6), we study how nouns referring to professional occupations or roles are translated.

To test hypothesis H1, and to avoid the limitations of using synthetic sentence sets as probes (as in Prates et al., 2020) which may not follow the distribution in the training data, we run instead random samples of the actual training data through the MT system. Our reasoning behind this is that the corpus represents the material "textual reality" learned by the MT system, which is what the system bases its decisions on. This is, therefore, the main difference between our study and Prates et al.'s (2020), and other studies using synthetic probes (see section 1.6). Prates et al.'s (2020) use of synthetic, intentionally ambiguous sentences without context studies the combined effect of a number of phenomena: (a) how grammatical gender is used to encode social gender in the texts found in the corpus, an encoding that may have changed across documents; (b) the fact that the social gender distribution of a list of professions in a specific country or time may differ from the distributions that the texts in the corpus describe; and (c) the actual limitations of the machine translation system and the training process to reproduce the observed grammatical gender distributions found in the corpus. By using samples taken from the training corpus instead, we expect to be able to study (c) separately, as one would expect that the system should be able to reproduce their distribution more accurately. We use training data to give the maximum possible advantage to the MT system, and study how it still fails. Note that some sentences may

[1] https://huggingface.co/Helsinki-NLP/opus-mt-en-es
[2] https://huggingface.co
[3] Details about the model may be found in the training log provided with the model at https://github.com/Helsinki-NLP/OPUS-MT-train/tree/master/models/ca-es
[4] The version of the corpus used is available at https://object.pouta.csc.fi/Tatoeba-Challenge-v2020-07-28/eng-spa.tar it contains 154,168,790 sentence pairs in the training set, and 195,195 sentences were used as the development set.

contain the necessary information for a target language gender choice, and some may not (see section 2.2 below). By giving the system this advantage, one would expect the amount of distortion of grammatical gender distribution to be a kind of lower bound to the amount of distortion that would occur when translating unseen text, in particular when translating synthetic, context-deprived sentences. That is, the value obtained for the amount of distortion could be seen as a reference or baseline, since the grammatical gender distribution of these examples is as similar as possible to that of the training corpus.

Therefore, to focus as much as possible on the actual problem addressed, namely, whether grammatical gender biases already observed in the corpus are preserved or amplified when translating English profession names, the method we used was as follows:

1. **Initial filtering**: The corpus was filtered to avoid long lines (we selected sentence pairs having no more than 1500 characters in either side, that is, about a hundred words on either side); the corpus did indeed contain a few very long lines as the result of corpus segmentation issues.
2. **Sampling sentences containing professions**: After this filtering step, carefully handcrafted regular expressions were used as filters to determine an initial set of professions in English which were very common in the training set. The regular expression, obtained by iteratively improving it to try to harvest as many relevant sentences as possible, describes a set of *safe* contexts for this kind of nouns. The actual regular expression used was:

```
\(selected\|sworn in\|drafted\|enrolled\|appointed\|
recruited\|employed\|trained\|hired\|works\|worked\|working\|
moonlighting\|moonlights\|moonlighted\) as an\?[^[:alpha:]]
\+[a-z]\+[^[:alpha:]]
```

   This produced a list of 250 English nouns, from which we selected the 100 most frequent that are professions expressed with a single noun. Scanning the whole corpus is not straightforward, and we are aware that we may be missing some frequent profession nouns. However, we believe we extracted a sample of professions which is representative enough of professional grammatical-gender distribution observed in the corpus. Note that we focus on frequent professions which are well represented in the training corpus itself, instead of using external lists as Prates et al. (2020) did. We argue there is no point in using professions for which we do not know whether the system has even been trained on before, and this was the reason to look for them directly in the corpus. While there are some professions in their lists which are frequently mentioned in our corpus (such as *artisan* or *playwright*), and our regular expression has missed, we believe that our list is representative enough to test hypotheses H1 and H2 above. Indeed, we remind the reader of the fact that professions (and their grammatical gender distribution) in the training corpus (and an ideal system that has learned them) may not correspond to the reality of the current world.
3. **Selecting relevant, unambiguous example translations**: This list of English nouns is curated as follows. (a) Source nouns which already carry a semantic gender mark (such as in the pair *waitress/waiter*) are eliminated. (b) Source nouns which can function also as an adjective (such as *volunteer*, *executive*, or *associate*) are eliminated. (c) The nouns kept are manually paired to all possible Spanish translations, but only those that do show a variation with grammatical gender

(removing words such as English artist → Spanish *artista*, where the Spanish noun can be masculine and feminine). As a result, a list of 350 English profession noun – Spanish profession noun pairs that meet all the required criteria is used to prepare a dataset in which each record has the form "`[Tt]eacher [Tt]eachers [Pp]rofesora [Pp]rofesoras [Pp]rofesor [Pp]rofesores`", ready to be used in regular expressions; the first two are the English forms (singular and plural) and the last four are the Spanish forms (singular and plural, feminine and masculine).

4. **Extracting a sample of sentence pairs from the corpus for each profession**: These records were used to build a set of cascaded regular expressions that selected sentence pairs from the training corpus such that: (a) the English noun appeared only once in the source side, (b) the Spanish noun appeared only once as a single word or twice in masculine-plus-feminine doublets (often used in Spanish gender-inclusive language) of the form *maestra o maestro* or *maestras y maestros* ('teachers'), in either order. The first matching examples in the corpus (which already comes randomly ordered), up to a maximum of 1,000 occurrences (provided that a minimum of 300 could be collected) were selected.[5] We argue that 300 examples for each profession is a reasonable sample, and 1,000 sentences was chosen as a limit to keep the computational cost of neural machine translation reduced. As we did not initially consider other gender-inclusive solutions found in writing like *maestro/a*, *maestr@s*, *maestrxs*, or newer gender-neutral proposals like *maestres* (o *maestris*), our work initially considered single nouns in the four possible forms given above. We eventually focused on single-word translations, excluding doublets, due to their low presence in the corpus. This step reduced the list from 350 to 249 English profession nouns with their respective and different possible translations.
5. **Additional filtering after examining extracted sentence pairs:** At this point, we looked at all the remaining pair nouns, and we deleted those in which (a) the Spanish equivalent could be ambiguous (such as a female *technician* in Spanish, which could be translated as *técnica*, but this noun also means 'technique' and, as an adjective, it means 'technical'), (b) the selected translations into Spanish were not really professions (*esposo* and *esposa* as translations of *partner*) and (c) the noun could be something else than a noun — for example, *laywer* could be translated in the feminine singular as *letrada*, but this form could also appear as an adjective in expressions such as *asistencia letrada* ('legal assistance'), as in the following example, where *lawyer* is translated in the masculine as *abogado* but the feminine word *letrada* also appearing in the sentence is not a noun but an adjective:

    - English: *He adds that he was denied legal assistance in his appeal proceedings, as the* ***lawyer*** *refused to continue to represent him.*
    - Spanish: *Añade que en sus procedimientos de apelación no contó con asistencia* ***letrada*** *de oficio, habida cuenta de que su* ***abogado*** *se había negado a seguir representándole.*

[5] Drawing m samples from a binomial distribution yields a relative frequency p' which estimates the underlying probability p with a standard deviation $\sqrt{[p(1-p)/m]}$, which is maximum for an equiprobable distribution (p=½) and always smaller than $½/\sqrt{m}$: sampling 300 samples gives a standard deviation for p' of 0.014 (1.4%); 1000 samples reduce that to 0.008 (0.8%).

After this, the first 15 sentences extracted for each of the remaining noun pairs were manually examined with the aim of making sure that the profession in English had been translated as one of the considered equivalents in Spanish (e.g., attorney as abogada). If the noun had been translated at least once differently than desired, the noun pair was removed from the dataset. The resulting dataset consists of 300 to 1,000 sentences for each of the remaining 88 English noun – Spanish noun pairs.

This "cleaning" process mitigates but does not completely exclude possible ambiguities that may still creep into the probes. The following is an example of what that was removed from the list after inspecting 15 sentence pairs:

- English: *For the best Manitowoc Wisconsin Internet Web Defamation lawyers and Manitowoc Wisconsin Internet Web Defamation* ***attorneys*** *in the business, AttorneysDelivered will "deliver".*
- Spanish: *Manitowoc Wisconsin Internet Difamacion Abogados y* ***Procuradores*** *que ganar! Para obtener el mejor Manitowoc Wisconsin Internet Difamacion abogados en la empresa, AttorneysDelivered se "entregue".*

We were looking to have *attorney* translated as *procuradora*, *procurador*, *procuradores* or *procuradoras*. In the example extracted from the corpus, we analyse an advert that has *attorneys* in the source sentence, and *procuradores* in the target, just as the regular expression requested. Nonetheless, *attorneys* is not being translated as *procuradores*, but as *abogados*. At this point, it is important to mention once again that we did not judge the correctness or quality of the sentences analysed throughout the whole study due to being out of our scope (this sentence actually seems to have been machine-translated).

The result of this filtering step yields, a set of 88 English profession noun – Spanish profession noun pairs —in which the English nouns have no gender mark and the Spanish words show grammatical gender variation—, and, for each one, a relatively clean sample of 300 to 1,000 sentence pairs from the corpus containing that translation. These sets may be considered representative of the grammatical-gender distribution observed in the corpus and may therefore be used to perform a statistical study to assess hypothesis H1 above.

1. Obtaining the grammatical gender distribution in the training examples: For each noun pair in the dataset, statistics of masculine and feminine translations of the English nouns in the extracted examples, using appropriate regular expressions.
2. Obtaining the grammatical-gender distribution of the machine-translated version of the source side of the examples: Next, the English side of the same sentence pairs was machine-translated using the system trained on the corpus, and the same statistical analysis of grammatical gender was performed on the Spanish machine-translated output.
3. Comparative analysis: Finally, the grammatical gender statistics on the Spanish side in the training corpus and in the machine-translated output are collected as in (6) above to detect any sign of exacerbation or polarization of gender statistics in the corpus.

2.2 Human assessment of the lack of source-side context

We now describe a method to test hypothesis H2, namely, that the lack of sufficient context inside each source sentence to decide the target-side grammatical gender does indeed contribute to the distortion of grammatical gender distributions.

The fact that the MT system tends to offer the stereotypical grammatical gender more often than it actually appears in the corpus (as will be shown in Section 3.1) that was used to train it deserves further analysis. We are aware that part of the exacerbation observed may be due to the fact that, in some English sentences, the absence of context may render a Spanish translation using either gender equally adequate, with the MT system "falling back" or "defaulting" to the most common translation observed in the corpus, regardless of context. Bear in mind that the MT system translates each sentence in isolation, and, as a result, it can only rely on information that is present in the English sentence when assigning a grammatical gender to the Spanish noun. If there is not enough information, or the system is unable to extract it from the English sentence, it is reasonable to expect that it will fall back to the most frequent grammatical gender. This is, as mentioned in section 1.5, due to the fact that MT systems are basically target-language models conditioned on (representations of) the source sentence and are trained to produce the most likely output as observed in the corpus (these sentences are closely related to what Gonen and Websters (2020) call at risk sentences, see section 1.6). Analogous mechanisms are possibly involved in the reduction of target-language lexical diversity, as observed by Vanmassenhove et al. (2019) and also in our experiments (see the end of section 3.1).

To test hypothesis H2, and to get an indication of how often this might be the case —that is, when the system cannot extract any gender clue from the source sentence— we designed a web-based questionnaire addressed to human annotators. For the human assessment to take a reasonable time, we first selected five different profession nouns in our list for which the training corpus already showed an unbalanced distribution which favoured the feminine in one of them and the masculine in the remaining four: *nurse*, *engineer*, *scientist*, *researcher*, and *lawyer* (both in singular and plural). We then randomly selected 10 sentence pairs from the training corpus (in the same way as they were selected for the statistics described in Section 2.1) and modified them by reversing the grammatical gender of the corresponding Spanish noun (and where needed, modifying any agreeing words such as articles or adjectives). The rationale behind this reversal is the following: if translators judge that the modified target sentence would still be an adequate translation of the source sentence in at least one imaginable context, this would clearly indicate that there is no clue in the source sentence to choose a specific grammatical gender in the target. Bear in mind that one important effect of sentence pair randomization already present in the training corpus is that respondents cannot use the preceding and succeeding sentences as context.

We found that in the case of lawyer, as shown in 2.1, the corpus contained numerous versions of the same sentence from an advert, each one adapted to mention different company names and regions. Only the first such example was selected to be part of the sentences to be analysed by the different annotators.

The resulting sentence pairs were presented to 16 annotators (5 professional translators, 10 translation graduates and 1 translation student, all of whom translate from English to Spanish): annotations were collected for 49 examples in total, as one of the 5 x 10 = 50 examples shown to annotators was later found to be incorrect (an adjective did not agree with the gender-reversed noun) and was excluded from the analysis. Translators were not aware of the gender-reversing editing, and received the following instructions:

> Each question shows an English sentence and a Spanish translation. The English sentence contains one noun, which designates a profession. Your task is to examine how that profession has been translated into Spanish and think whether there would be at least one context where that translation would be correct. Please do not judge the correctness of the rest of the Spanish sentence: focus on that particular noun.

Then, for each sentence pair, the question was:

> The Spanish translation of the English sentence contains a translation of the word [English profession noun]. Would that translation be correct in at least one context you could think of for the sentence? ("No" means you cannot imagine such a context).

We did not tell annotators what *context* meant, but we reasonably assumed they would understand it referred to possible text surrounding the sentence or to possible circumstances external to the text. An example of such a sentence pair would be:

- English: *A three feet sea level rise is already considered by scientists to be "locked in."*
- Spanish: *Las científicas ya consideran la subida del nivel del mar de tres pies como «asegurada».*

If annotators imagined, for instance, two female scientists being mentioned just before this sentence, they would deem it to be correct in that context (Gonen and Webster's (2020) at risk sentences).

As said above, if human annotators, observing the sentence in isolation, judge that the Spanish sentence where the gender of the profession noun has been reversed "would [...] be correct in at least one context [they] can think of", this means that translators cannot detect any information in the (isolated) English sentence that would render the gender-reversed Spanish sentence as not correct in any conceivable context; and that, possibly, a machine-learning algorithm would not be able to extract any gender-determining information from the source sentence either. These, therefore, would likely be the cases in which the neural MT system would probably go for the majority gender in the training corpus.

3 Results

3.1 Automatic assessment

As expected, the distribution of grammatical gender in the training corpus encodes the prevailing personal gender stereotypes. In the corpus, 86 out of the 88 nouns lean towards the masculine grammatical gender, except for *nurse* as *enfermera* and *dancer*

English	Spanish feminine	Spanish masculine	Feminine in corpus	(%)	Masculine in corpus	(%)	Feminine in MT	(%)	Masculine in MT	(%)	Feminine percentage change
teacher	*maestra*	*maestro*	110	11%	871	89%	45	5%	776	95%	–6%
cook	*cocinera*	*cocinero*	208	21%	771	79%	122	14%	754	86%	–7%
illustrator	*ilustradora*	*ilustrador*	155	16%	838	84%	97	10%	830	90%	–5%
dancer	*bailarina*	*bailarín*	371	38%	617	62%	301	31%	670	69%	–7%
trainer	*capacitadora*	*capacitador*	72	7%	916	93%	0	0%	4	100%	–7%
bartender	*camarera*	*camarero*	114	12%	868	88%	20	3%	670	97%	–9%
curator	*conservadora*	*conservador*	159	18%	703	82%	0	0%	16	100%	–18%
curator	*curadora*	*curador*	237	24%	750	76%	157	17%	754	83%	–7%
singer	*cantaora*	*cantaor*	143	20%	561	80%	5	5%	86	95%	–15%
counselor	*orientadora*	*orientador*	84	20%	339	80%	0	0%	4	100%	–20%
choreographer	*coreógrafa*	*coreógrafo*	270	27%	726	73%	212	21%	781	79%	–6%
nurse	*enfermera*	*enfermero*	626	63%	360	37%	669	68%	310	32%	5%
dancer	*bailaora*	*bailaor*	392	59%	276	41%	36	80%	9	20%	21%
student	*alumna*	*alumno*	34	4%	934	96%	4	5%	73	95%	2%
curator	*comisaria*	*comisario*	274	28%	717	72%	64	30%	151	70%	2%
hairdresser	*peluquera*	*peluquero*	285	29%	692	71%	262	31%	595	69%	1%

Table 1: Comparison of the number (and percentage) of times in which a selection of frequent English gender-ambiguous professional noun is translated as a feminine or masculine Spanish noun in the first 1,000 sentence pairs extracted from the corpus (see text) and in a machine-translated version of the same English sentences. The last column indicates the increase or decrease in the percentage of feminine translations. The already stereotypical distributions observed in the corpus are sometimes grossly exacerbated.

as *bailaora*. Columns "in corpus" in Table 1 shows statistics for a selection of English noun–Spanish translation pairs. This basic result remains the same after the MT process (see the "in MT" columns). Examples shown include (a) pairs for which masculine forms

predominate in the corpus and for which the percentage of feminine forms decreases by more than 5% when MT is applied (unshaded) (b) pairs for which feminine forms predominate in the corpus (light shading); and (c) the only pairs for which masculine predominance in the corpus is significantly reversed (by at least 1%; darker shading).

The results reveal that MT not only preserves the stereotypes in the corpus (nurses are feminine, engineers are masculine), but rather amplifies them: the percentage of the majority grammatical gender increases in their machine-translated counterparts (the "in MT" columns), for all nouns but three: *curator* as *comisaria*, *hairdresser* as *peluquera*, and *student* as *alumna* (darkest shading, last rows in Table 1), but with small differences, considering the size of the sample (see section 2.1, footnote to step (4)). Note the total number of both masculine and feminine translations decreases when MT is applied; this is due to the fact that alternative translations of the English professional noun are produced.

There are indeed clear signs of lexical impoverishment in the MT process. When a noun in English may be translated into Spanish in different ways, the system tends to have a preferred translation. An example of this is illustrated by the translation of the word *teacher*. We select the first 1,000 sentences where teacher is translated as a form of *maestro* or *profesor*; the statistics in the corpus show 598 cases of *maestro* forms and 402 cases of *profesor* forms. If the source English sentences are then translated into Spanish, the number of *maestro* forms increases to 623, and that of *profesor* forms decreases to 366; of the remaining 11 cases, 8 correspond to forms of the alternative translation *docente* and the remaining 3 are paraphrased translations where no profession noun is found. This happens to many other nouns and is related to the loss of lexical diversity discussed in section 1.6.

As a final note, the nouns *nurse* and *teacher* we have studied appeared among the most frequent nouns suffering from grammatical gender stereotypes in English–Spanish *at risk* sentences by Gonen and Webster (2020). Their feminine to masculine ratios are not directly comparable in view of the different method to build probe sentences (see section 1.6) but are compatible with our observation.

3.2 Translator assessment

Table 2 collects the responses of 16 translators when they were asked (see section 2.2) to assess whether they could think of a context where one could accept 49 corpus translations in which the grammatical gender of the professional noun had been reversed in Spanish.

Percentage of translators agreeing	Sentences adequate with reversed gender	
	Count	Percentage
≥50%	39	79.6%
≥75%	23	46.9%
≥90%	11	22.5%
100% (unanimous)	6	12.2%

Table 2: Agreement, in a group of 16 annotators, about the adequacy, in some conceivable context, of translations from the corpus in which the gender of a noun-denoting profession has been reversed.

Most annotators agree that in many of the Spanish sentences where the grammatical gender of the nouns denoting a profession was reversed with respect to that in the corpus would indeed be adequate translations of the English sentence in a certain context. Half (50%) of the 16 annotators would agree that 79.59% of the Spanish sentences analysed in the questionnaire would be adequate translations within a certain context with the grammatical gender change. This result is already significantly high.

If the agreement between the annotators is required to obtain more qualified majorities, they still agree that a gradually decreasing percentage of the sentences could be adequate translations within a certain context, but even when all 16 annotators are asked to unanimously agree, it is still the case that in 12.24% of the sentences (6 out of 49) the gender-reversal still renders the Spanish translation adequate in a certain context. In these 6 sentences, all 16 annotators could think of an authorizing context, and the neural MT system processing the ambiguous sentence would be expected to fall back to stereotypes. We have computed Krippendorf's alpha,[6] a measure of the reliability of annotators, and the value happens to be positive ($\alpha>0$, annotators do agree above chance) but still low, $\alpha=0.261$.

These results, combined, are a clear indication that, while annotators do not seem to strongly agree on many sentences, there is an important fraction (12.24%) of the sentences from the training corpus that are unanimously considered as not having enough context for the MT to be able to choose one grammatical gender over the other. Sentences in that set would definitely contribute to the amplification of grammatical-gender stereotypes through the falling back process described. For some of the remaining sentences, disagreement may be due to the fact that not all annotators are able to evoke a context that would authorize the gender-reversed translation, but some are.

The lack of context associated to the sentence-level granularity of MT is a major challenge, and mitigation strategies explored in the literature (see section 1.6) either annotate the input during training (Basta et al., 2020; Štafanovičs et al., 2020) or retrain (Saunders and Byrne, 2020) the system on de-biased data. These strategies are

[6] https://en.wikipedia.org/wiki/Krippendorff's_alpha

unavailable in the customary configuration in which the MT system with random bags of sentence pairs deprived of any document context or annotation, and the human assessment attests to that, by giving a clear indication of the lack of that context beyond the isolated sentence.

Note that we do not try to give a precise quantitative assessment of the contribution of lack of context to the distortion of grammatical-gender distributions, but rather a clear indication of the existence of a significant contribution.

4 Concluding remarks

We have performed a straightforward black-box text-centred study to examine how MT modifies the distribution of grammatical genders observed when translating a sample of sentences containing 88 frequent translations pf professional nouns from English to Spanish. The main novelty of our study is that, instead of using synthetic sentences to probe possible learning biases, we machine-translate text which is as similar as possible to that used for training, using probe sentences extracted from the training corpus itself, filtered to isolate the grammatical-gender problem studied. The first finding is consistent with those of other authors (see Savoldi et al. 2021): the system almost invariably amplifies the grammatical gender imbalance observed in the corpus, reinforcing the social stereotypes encoded (hypothesis H1; see section 1.5). This amplification or polarization may be seen as part of a more general tendency of MT to reduce the lexical diversity observed in the corpus. As this exacerbation of grammatical gender imbalance is likely to be due to the fact that an MT system which is trained on isolated sentences and translates isolated sentences does not have access (or cannot identify and utilize) the context necessary to reproduce the grammatical gender observed in the corpus, we have devised an experiment where professional translators are requested to validate an edited version of sentence pairs in the corpus where the grammatical gender of the Spanish professional noun has been reversed. The second finding is a widespread agreement that contexts where many of these translations are still adequate can be imagined; this would clearly indicate that the amplification of grammatical gender imbalances is simply the result of the MT system falling back to the default translation in the absence of clear cues (hypothesis H2; see section 1.5).

Note that in our study we have deliberately decided to decouple the study of the exacerbation of grammatical gender imbalances in MT (which can be directly observed in text) from its actual social implications, as the latter would involve an explicit modelling of the processes involved in the textual encoding of personal gender by writers and the subsequent decoding by readers, two processes that are necessarily mediated by implicit or explicit encoding and decoding conventions (traditional masculine-as-default versus gender-inclusive usages, among others).

We are aware of the limitations of the preliminary black-box study presented here; therefore, in future work, we plan to:

- Model separately the extra-textual processes involved in gender bias, namely, the role of encoding and decoding conventions during writing and reading respectively, following the steps of Savoldi et al. (2020).
- Study how gender manifestations beyond the masculine–feminine binary scheme are encoded, as well as an analysis of indirect non-binary language solution and emerging direct non-binary language proposals.
- Improve the detection and extraction of sentence pairs containing profession nouns that are gender-ambiguous in English but need to be disambiguated so that it is as exhaustive and representative as possible.
- Refine our sentence filtering and sampling method, using natural-language processing tools more powerful than regular expressions: part-of-speech taggers, word aligners, etc.
- Collaborate with neural MT researchers exploring the explainability (Belinkov et al., 2020) of the decisions made by these systems.
- Generalize the methodology so that it can be used for other language pairs.

Acknowledgements

The authors thank Jörg Tiedemann for his availability when questions about the corpora and the systems arose, and Antonio Toral for interesting suggestions. We also thank the sixteen participants in the human study.

References

Basta, Christine; Costa-jussà, Marta R.; Fonollosa, José A. R. (2020). Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In: *Proceedings of the Fourth Widening Natural Language Processing Workshop, Seattle, U.S.A., July 2020*, pp. 99–102. <https://aclanthology.org/2020.winlp-1.25>. [Accessed: 20221209].

Belinkov, Yonatan; Durrani, Nadir; Dalvi, Fahim; Sajjad, Hassan; Glass, James (2020). On the linguistic representational power of neural machine translation models. *Computational Linguistics*, v. 46, n. 1, pp. 1–52. <https://direct.mit.edu/coli/issue/46/1>. [Accessed: 20221209].

Gonen, Hila; Webster, Kellie. (2020). Automatically identifying gender issues in machine translation using perturbations. In: *Findings of the Association for Computational Linguistics: EMNLP 2020, (November 16–20),* pp. 1991–1995. Cornell University. <https://arxiv.org/abs/2004.14065>. [Accessed: 20221209].

Prates, Marcelo O. R.; Avelar, Pedro H.; Lamb, Luís C. (2020). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, n. 32, pp. 6363–6381. <https://doi.org/10.1007/s00521-019-04144-6>. [Accessed: 20221209].

Rescigno, Argentina A.; Vanmassenhove, Eva; Monti, Johanna; Way, Andy (2021). A Case Study of Natural Gender Phenomena in Translation: A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and

Spanish. In: *Proceedings of the 7th Italian Conference on Computational Linguistics, CLiC-it 2020, in CEUR Workshop Proceedings* vol. 2769, pp. 62-90. <https://aclanthology.org/2020.amta-impact.4> [Accessed: 20221209].

Renduchintala, Adithya; Diaz, Denise; Heafield, Kenneth; Li, Xian; Diab, Mona (2021). Gender Bias Amplification During Speed-Quality Optimization in Neural Machine Translation. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 99–109. <https://aclanthology.org/2021.acl-short.15>. [Accessed: 20221209].

Saunders, Danielle; Byrne, Bill (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. In*: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020, July 5–10)*. Association for Computational Linguistics, pp. 7724-7736. <https://aclanthology.org/2020.acl-main.690/>. [Accessed: 20221209].

Savoldi, Beatrice; Gaido, Marco; Bentivogli, Luisa; Negri, Mateo; Turchi, Marco (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, v. 9, pp. 845-874. <https://doi.org.10.1162/tacl_a_00401>. [Accessed: 20221209].

Štafanovičs, Arturs; Bergmanis, Toms; Pinnis, Mārcis (2020). Mitigating Gender Bias in Machine Translation with Target Gender Annotations. In: *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, pp. 629–638. <https://aclanthology.org/2020.wmt-1.73>. [Accessed: 20221209].

Stanovsky, Gabriel; Smith, Noah A.; Zettlemoyer, Luke (2019). Evaluating gender bias in machine translation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019, Florence, Italy, July 28–August 2, 2019)*, pp. 1679–1684. <https://doi.org/10.18653/v1/P19-1164>. [Accessed: 20221209].

Tomalin, Marcus; Byrne. Bill; Concannon, Shauna; Saunders, Danielle; Ullman, Stefanie (2021). The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics and Information Technology,* n. 23, pp. 419-433. <https://doi.org/10.1007/s10676-021-09583-1>. [Accessed: 20221209].

Toral, Antonio. (2019). Post-editese: an Exacerbated Translationese. In: *Proceedings of Machine Translation Summit XVII: Research Track, Dublin, Ireland.* European Association for Machine Translation, pp. 273–281. <https://aclanthology.org/W19-6627/>. [Accessed: 20221209].

Vanmassenhove, Eva; Hardmeier, Christian; Way, Andy (2018). Getting Gender Right in Neural Machine Translation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, *Brussels, Belgium.* Association for Computational Linguistics, pp. 3003–3008. <https://aclanthology.org/D18-1334>. [Accessed: 20221209].

Vanmassenhove, Eva; Shterionov, Dimitar; Way, Andy (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In: *Proceedings of Machine Translation Summit XVII: Research Track, Dublin, Ireland*. European Association for Machine Translation, pp. 222.232. <https://aclanthology.org/W19-6622>. [Accessed: 20221209].

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; Polosukhin, Illia (2019). Attention is all you need. In: *Conference on Neural Information Processing Systems (NIPS 2017),* pp. 1-11. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. [Accessed: 20221209].

Volansky, Vered; Ordan, Noam; Wintner, Shuly (2015). On the features of translationese. *Digital Scholarship in the Humanities,* v. 30, n. 1, pp. 98-118. <https://doi.org/10.1093/llc/fqt031>. [Accessed: 20221209].