Investigating the Roots of Gender Bias in Machine Translation: Observations on Gender Transfer between French and English

Anonymous ACL submission

Abstract

This paper aims at identifying the inner mechanisms that make a translation model choose a masculine rather than a feminine form, an essential step to mitigate gender bias in MT. We conduct two series of experiments using probing and comparing the predictions of translation and a language models to show that i) gender information is encoded in all decoder's and encoder's representations and ii) the translation model does not need to use information from the source to predict `his`.

1 Introduction

State-of-the-art machine translation models (TM) have been shown to suffer from gender-bias (Prates et al., 2019) and works trying to mitigate this problem constitute a very active line of research (e.g. (Costa-jussà and de Jorge, 2020; Saunders and Byrne, 2020; Savoldi et al., 2021)). We adopt here a different point of view and try to identify the inner mechanisms that make the translation model choose a masculine rather than a feminine form.

In the Transformer encoder-decoder architecture embedded in state-of-the-art MT systems, the target sentence is generated incrementally: target tokens are chosen one after the other and this choice relies on evidence coming both from the source sentence thanks to cross-attention and from the target context (i.e. target tokens that have already been generated). We strongly believe that an essential step to mitigate gender bias in MT is to better understand whether this problem is due to *i)* gender information being incorrectly captured in the source sentence *ii)* cross-attention being unable to properly transfer this information from the source to the target sentence or *iii)* target context pushing the model to make the wrong decision.

To identify the causes of gender bias among these (non-exclusive) possibilities, we focus on gender transfer from French into English building on the differences in gender expression between these two languages. Using a controlled test to precisely control where and how gender is expressed in sources and targets, we conduct two series of experiments to quantify the flow of information in an encoder-decoder architecture.

First, using linguistic probes (Alain and Bengio, 2017; Belinkov et al., 2020), we show that gender information is encoded in all the representations built and manipulated by the encoder and the decoder (§4). Second, we propose to compare the predictions of a language and of a translation models to distinguish information coming from the encoder from information from the target context when generating a target token. Our experiments (§5-6) rely on the different ways gender can be expressed in French and in English, to point out that the prediction of the feminine form `her` does not necessarily rely on the same information as the masculine form, which opens new perspectives to address the issue of gender-bias in MT.

The contribution of this work is twofold:

- we release a French-English controlled set to study gender bias in MT. Based on the same intuitions as Stanovsky et al. (2019), our test set covers a new language and is much larger than existing corpora which will allow us to control precisely different confounding factors such as the frequency in the training data or the way gender is expressed;
- we show that, counter-intuitively, the TM does not need to use information from the source sentence to predict `his` while this is necessary for the prediction of `her`.

2 A Controlled Set to Study Gender Transfer

We fill first describe the controlled test set used in our experiments and explain why (and how) we will use to identify the flow of information in an encoder-decoder architecture.

Test Set To study gender transfer from French into English, we consider a controlled test set made of 3,394 parallel sentences perfectly balanced between genders following the pattern:

- `[DET] [N]` a terminé son travail.
- The `[N]` has finished `[PRO]` work.

where `N` is an occupational noun chosen from (Dister and Moreau, 2014) that matches feminine and masculine professions and occupations in French. This list was automatically translated in English and manually corrected by the authors. `DET` is the French determiner in agreement with the `N` (the feminine form *la*$_F$, the masculine form *le*$_M$ or the epicene form *l'* that is be used for both genders when the job noun begins with a vowel) and `PRO` is the English possessive pronoun `her` or `his`.

In the English sentences, gender is unambiguously marked by the possessive pronoun; it may also be marked by the occupational noun that has different feminine and masculine forms for 5.5% of the sentences, (e.g. actress$_F$/actor$_M$). In most sentences, the occupational noun is epicene and gender can not be inferred from the surface form. In French, the gender can be expressed by the determiner, the occupational noun, or both; in rarer cases, both words are epicene, and the feminine and masculine versions are identical.[1]

TM as Conditional Language Model When translating sentences of our controlled set, the prediction of the English possessive pronoun can rely on two kinds of evidence: *i)* using cross-attention, the model can encode information about the French subject gender into the representation of the possessive pronoun;[2] *ii)* because of the decoder self-attention, this representation can also encode information from the target context, notably the English subject that encodes gender information either directly or because its representation depends on the French subject (through cross-attention).

To identify the information flow in the encoder-decoder architecture we compare the predictions of a translation model (TM) and of a language model (LM). Indeed, in a TM, the i-th target token t_i is chosen by taking information from the source sentence $\mathbf{s}$ and from $t_{<i} = t_0, ..., t_{i-1}$ the tokens of the target sentence already generated, while an LM only considers information from the target context. A TM can be viewed as a *conditional language model* which computes $p(t_i|t_{<i}, \mathbf{s})$ while an LM computes $p(t_i|t_{<i})$. By comparing the predictions of these two models, we can evaluate the impact of information coming from the source.

3 Experimental Setup

Translation and Language Models We use `JoeyNMT` (Kreutzer et al., 2019), an implementation of a translation system based on the `Transformer` model of Vaswani et al. (2017). Encoder and decoder are composed of 6 layers, each with 8 attention heads; the *feed-forward* layers have 2,048 parameters and the dimension of lexical embeddings is 512. Our model comprises a grand total of 76,596,736 parameters.

We consider our in-house implementation of a TRANSFORMER language model with the same dimensions as the MT decoder using the PYTORCH library (Paszke et al., 2019).[3] To mimic the decoder, we use an autoregressive ('causal') LM in which the representation of the i-th token is computed based on the $(i-1)$ previous tokens.

The two models are trained by optimizing the cross-entropy with ADAM on the same data (see below) and achieve a BLEU score of 34.0 and a perplexity of 43.0 on the WMT'14 test set.

Training Data We consider the English-French parallel corpus from the WMT'15 'News' taskthat contains 4,813,682 sentences and nearly 141 million French running words. All raw corpora were segmented into sub-lexical units using the unigram model of `SentencePiece` (Kudo, 2018); the vocabularies contain 32,000 units in each language.

Are training data gender balanced? We conduct two experiments to check if genders are well-balanced in our train set. First, we count the number of occurrences of `his` and `her` the prediction of which is at the heart of our evaluation. It appears that there are more than twice as many occurrences of `his` than of `her` in (108,364 versus 47,444 occurrences). Second, we looked at the number of times the French possessive pronoun `son` was translated by `his` or `her` in the training data: we align the train set with a Bayesian HMM model(Östling and Tiedemann, 2016) and use the alignment link to find all possible translations of the French `son` token. Results reported in Table 2

[1] See Appendix A for examples.

[2] The French subject can have either a direct impact through cross-attention or an indirect impact as the representation of all source tokens depends on it (encoder self-attention). We will not try to distinguish these two effects.

[3] Code is available in the supplementary material.

(Appendix B) show that translating `son` as `his` is three times more frequent than as `her`.

4 Probing

We use *probing* (Belinkov and Glass, 2019) to analyze which words in the source sentences convey gender information: a *probe* (Alain and Bengio, 2017) is trained to predict linguistic properties from the representations of language (i.e. token embeddings); achieving high accuracy at this task implies these properties are encoded in the representations.

Experimental Setup We collected the hidden representations at the output of the first and last layer of the encoder and the decoder of all tokens except the French subject and associate each of them to a label indicating whether the occupational noun in the French sentence refers to a woman or a man. For each of these examples, we randomly split all sentences between a train (75%) and a test (25%) set. Then, we used `scikit-learn` (Pedregosa et al., 2011) to learn a logistic regression to predict gender from a single token representation.

Results The probe achieve an average precision of 74.1% (resp. 87.9%) for the first (resp. last) layer of the encoder and of 80.5% and 86.2% for the decoder (results are detailed in Appendix C), showing that gender information is encoded in the representations of all source and target tokens.

In the spirit of the experiments conducted to analyze monolingual representations (Marvin and Linzen, 2018), we have transformed source sentences to evaluate the robustness of our observation. Results in Appendix C show that the encoder is able to capture gender information even in complex situations (e.g. presence of *distractors*, only epicene French determiner, ...)

5 Gender Bias in LM and TM

Results reported in the previous section show all token representations include gender information. We will now investigate whether models use this information. Indeed, a well-known weakness of probes is that they can detect the presence of linguistic information in representations, but they cannot measure how much of this information is used in the model predictions (Ravichander et al., 2021).

Principle We investigate the ability of a TM or an LM to predict the correct form of the possessive pronoun in the English sentences by estimating $p(\texttt{her}|c)$ and $p(\texttt{his}|c)$ where c is either the prefix *The [occupational noun] has finished...* (for an LM) or the prefix and the source sentence (for a TM). These four probabilities can be easily estimated with a forced decoding of the English sentence. We evaluate the model preference to generate `her` over `his` by:[4]

$$b(c) = 1 - \frac{2 \times p(\texttt{her}|c)}{p(\texttt{his}|c) + p(\texttt{her}|c)} \quad (1)$$

Intuitively, the closer to -1 (resp. 1) b is, the larger (resp. smaller) $p(\texttt{her}|c)$ is with respect to $p(\texttt{his}|c)$ and if the two values are in the same ballpark, b will be close to zero. Considering the probabilities to generate the possessive pronouns rather than looking at the token actually predicted by a model allows us to consider all sentences in our test set, even those for which the another token than `his` and `her` is predicted.

Bias Evaluation We first consider the probabilities computed by an LM (i.e. without considering the source) and report, in Figure 1, the distribution of b for sentences with epicene occupational nouns. Would the model be unbiased, b would be close to 0, as no gender information has been expressed. It, however, appears that for the vast majority of sentences $p(\texttt{his}|c)$ is much larger than $p(\texttt{her}|c)$ even though, according to the distributional hypothesis, these two tokens should have very similar representations as they appear in similar contexts.

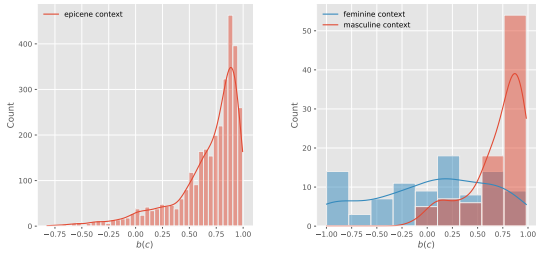
We have also looked at the distribution of $b(c)$ for the rarer English sentences in which the gender is lexically expressed in the noun. This distribution (Figure 1) shows that in masculine contexts the distribution of b is skewed to 1, meaning that not only the LM prefers `his` to `her` but also the probability of the former is much larger than that of the latter. On the contrary, for feminine contexts, $b(c)$ is spread over the whole domain. Similar conclusions can be draw for TM (see Figure 3 in Appendix E).

An Explanation for Bias These two observations show that the LM has a clear tendency to prefer generating `his` over `her`, which could result from the fact that the former is more frequent in the train set than the latter (§3). Counting the number of times the occupational nouns with the largest and smallest values of b co-occur with `his`

[4] Considering b rather than a simpler quantity like the ratio $p(\texttt{his}|c)/p(\texttt{her}|c)$ allows us to ensure that all values are bounded in $[-1, 1]$ which makes plotting distribution easier.

or her confirms this hypothesis:[5] for occupational nouns that are not epicene, values of b that are close to -1 or 1 are often made of several words, one of which is *woman* or *man*.

Epicene nouns for which b close to 1 all appear in the train set more often in sentences containing his than in sentences containing her, even though they both can describe a man or a woman. On the contrary, the second smallest values of b if achieved by *porn star*, one of the few epicene nouns for which the LM prefers to generate her.



(a) Unmarked gender (b) Marked gender

Figure 1: Measures of gender bias in an LM.

6 Investigating Information Transfer from the Source

We now focus on sentences for which the TM is able to give a higher probability to the correct form of the possessive pronoun but the LM is not. These sentences correspond to cases in which the correct prediction is due to the source information.

Results in Table 1 are consistent with observations in §5: both the LM and TM models have a clear tendency to prefer his, whatever the context (hence the perfect accuracy for masculine sentences). More interestingly, for feminine sentences, taking the source into account strongly increases the number of times $p(\texttt{her})$ is greater than $p(\texttt{his})$. There is thus an effective transfer of information from the source even if it is not perfect: overall, her gets a probability higher than his in only 43.0% of the sentences.

We have tried to characterize the cases where the TM succeeds in correcting the LM estimation. Our analysis shows that this happens when:

- the French determiner is not epicene: in this case, only 4,8% of sentences are corrected, versus 44,4% when this is not the case;
- the probabilities of his and her estimated by the LM are close: the average ratio between $p(\texttt{her})$ and $p(\texttt{his})$ is 0.226 in sentences that are corrected and 0.175 in sentences that are not corrected and the median 0.146 and 0.099;
- the occupational noun appears in the train set: the TM corrects 23.5% when this is the case and only 15.6% when it is not.

[5] See Tables 6 and 7 in Appendix D.

Gender Marked		Target Gender			
		Fem.		Masc.	
English	French	LM	TM	LM	TM
✔	✖	0.50	0.50	1.0	1.0
	✔	0.43	**0.81**	0.99	1.0
✖	✖	0.038	0.0	0.96	0.99
	✔	0.053	**0.39**	0.95	1.0

Table 1: Precision achieved by an LM and a TM when generating the possessive pronoun.

Another hypothesis we have explored is the possibility that the segmentation into sub-lexical tokens of French sentences creates gender-specific suffixes that facilitate the transfer of gender information. We have represented (Figure 2) the 20 most frequent suffixes in feminine occupational nouns. It appears that these suffixes are in their vast majority feminine markers but their very presence does not guarantee that the gender of the translated sentence is correct: they appear in as many "correct" sentences as "incorrect" sentences.

Above all, these results show that the TM does not need to use the same information to decide between her and his: for the latter, it can rely on the target context only; but, for the former, it must learn to transfer information from the source, which it only does imperfectly (at least for our system). This suggests that gender bias could partly results from using cross-entropy as a loss function: indeed, the model appears to be quite capable of learning to predict his for the wrong reason and without necessarily taking into account the gender information present in the source, which weakens the estimation of the parameters necessary to take into account gender information (e.g. cross attention).

7 Conclusions

The experiments reported in this study shed a new light on the cause of gender bias in MT systems: they show that the prediction of the her does not necessarily rely on the same information as his, which opens new perspectives to address the issue of gender-bias in MT. In future work, we consider using other loss function to force the TM to consider source information when predicting gender.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Anne Dister and Marie-Louise Moreau. 2014. *Mettre au féminin : guide de féminisation des noms de métier, fonction, grade ou titre*, 3e édition edition. Fédération Wallonie-Bruxelles.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2019. Assessing gender bias in machine translation – a case study with google translate.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

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

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

A Examples of Gender Expression in our Corpus

In French gender can be expressed in four different ways:

- in 53.0% of the sentences, it is marked by both the determiner and the occupational noun:
  - (la_F $boulangère_F$/le_M $boulanger_M$) a fini son travail — the baker

  - (la_F *banquière*$_F$/le_M *banquier*$_M$) a fini son travail — the banker

- in 24.2% of the sentences, it is marked only by the determiner the occupational noun having the same feminine and masculine form:
  - (la_F *cinéaste*/le_M *cinéaste*) a fini son travail — the film-maker
  - (la_F *médecin*/le_M *médecin*) a fini son travail — the doctor
- in 14.9% of the sentences, it is marked only by the noun but not the determiner :
  - (*l'astrophysicienne*$_F$/*l'astrophysicien*$_M$) a fini son travail — the astrophysicist
  - (*l'ouvrière*$_F$/*l'ouvrier*$_M$) a fini son travail — the worker
- in 7.9% of the sentences, the gender is not marked at all :
  - *l'artiste* a fini son travail — the artist
  - *l'ethnographe* a fini son travail — the ethnographer

  l'artiste and *l'ethnographe* designate either a female or a male person.

As in English, gender can be expressed in two different ways by the noun:

- in 5.5% of the sentences, the gender is marked by the occupational noun:
  - *the policewoman*$_F$/*the policeman*$_M$ has finished her/his work
  - *the cowgirl*$_F$/*the cowboy*$_M$ has finished her/his work
- in 94.5% of the sentences, the gender is not marked by the occupational noun:
  - the *user* has finished her/his work
  - the *tailor* has finished her/his work

B Are gender well-balanced in the training data?

Table 2 shows an excerpt of the translation table of the `son` token. Alignment was performed by `eflomal` (Östling and Tiedemann, 2016) after BPE tokenization and symmetrized using the `grow-diag-final-and` heuristic. Note that, in French, `son` can either be a possessive pronoun or a noun meaning 'sound'.

translation	frequency
_its	27.94%
_his	**18.28%**
_the	7.24%
_her	**6.42%**
_a	3.34%
_their	2.92%
_it	2.45%
_sound	1.37%
s	1.33%
_he	0.76%
`___OTHER___`	22.13%

Table 2: Most frequent translation of the French token *son* according to the word alignment links. *son* is aligned with 3,658 different types. Those which do not appear in the table are grouped in the special token `___OTHER___`.

layer	decoder	
	the	all tokens
1	89.5% ±0.2	71.6% ±0.6
2	92.0% ±0.1	76.3% ±0.7
3	91.8% ±0.1	78.1% ±0.6
4	90.9% ±0.2	79.1% ±0.6
5	89.3% ±0.2	82.4% ±0.5
6	87.7% ±0.2	84.7% ±0.3

Table 3: Precision of a probe predicting the gender of the French occupational noun given the decoder representation.

C Probing Results

Detailed results of our probing experiments are in Table **??** for the encoder and in Table **??** for the decoder. For the latter, the diversity of the translation structures makes it impossible to carry out a position-by-position analysis.

We have also carried out a series of syntactical modifications of our pattern to test the transmission of gender information through the encoder:

- "gender weakening": by neutralising the marked gender information of the French determiner `le`/`la`, by replacing them with `chaque` (each);
- "gender reinforcement" : by inserting an adjective of the same gender as of the noun between the determiner and the noun;
- "syntactical distancing": by increasing the

layer	encoder						random labels
	a	terminé	son	travail	.	eos	son
1	80.4% ±1.1	75.1% ±0.3	80.6% ±0.3	76.4% ±0.6	59.5% ±1.0	73.3% ±1.0	45,3% ±0.9
2	85.8% ±1.0	80.8% ±0.2	81.6% ±0.3	78.3% ±0.7	87.6% ±0.6	88.3% ±0.7	50,7% ±0.8
3	89.5% ±0.6	88.2% ±0.2	89.2% ±0.2	82.0% ±1.1	86.5% ±1.0	87.6% ±0.6	48,8% ±0.9
4	90.8% ±0.4	89.3% ±0.2	90.6% ±0.2	85.9% ±0.9	85.7% ±1.0	85.6% ±0.7	48,6% ±0.8
5	90.4% ±1.0	89.3% ±0.2	90.4% ±0.2	85.5% ±0.8	86.4% ±0.8	85.2% ±1.2	49,6% ±0.8
6	91.0% ±0.6	89.3% ±0.2	90.0% ±0.2	86.0% ±1.0	86.4% ±1.1	85.1% ±0.8	49,2% ±0.8

Table 4: Precision of a probe predicting the gender of the French subject given the encoder representations.

paradigmatic distance between the occupational noun and the rest of the sentence;

- "distractor" : by adding a subordinate clause which echoes with the gender information of the occupational noun.

Results on the modified source sentences are in Table 5.

D Pronouns Predicted by an LM

In Table 6, we report the 10 sentences with the smallest and largest ratio (see Section 5) as well as the corresponding occupational noun. In Table 7, we have reported the number of sentences that contains both the occupational noun and either `his` or `her` for the ten epicene occupational nouns with the largest value of b. With a few rare exceptions (*porn star*), smallest values of b corresponds to occupational nouns that do not appear in the train set.

To have a point of comparison, we have also computed $b(c)$ for four stereotypical contexts in which the occupational noun was *man*, *woman*, *boy* and *girl*. Surprisingly enough, $b(\text{boy})$ is only 0.3779, while for the other three contexts, $b(c)$ is close to its maximal value (for masculine contexts) and to its minimal values (for feminine contexts): -0.9720 for *girl*, -0.9739 for *woman* and 0.9609 for *man*.

E Gender Bias in TM

We have represented in Figure 3 the distribution of $b(c)$ for a translation model. The conclusions to be drawn from these results are very similar to the observations for a language model: there is a clear tendency to favor the masculine form: even when gender is marked both in the source and the target sentences, the values of $b(c)$ are spread all over the domain for feminine sentences while, for masculine sentences, for most sentences, $b(c)$ is close to 1.

F Identifying Sentences Corrected by the TM

We have represented, in Figure 4, the distribution of the ratio between $p(\texttt{her})$ and $p(\texttt{his})$ for sentences for which the LM assigns a larger probability to the 'wrong' possessive pronoun; 'corrected' correspond to sentences in which the translation model assigns a higher probability to the correct pronoun and 'not corrected' when this is not the case.

G Distribution of prefixes

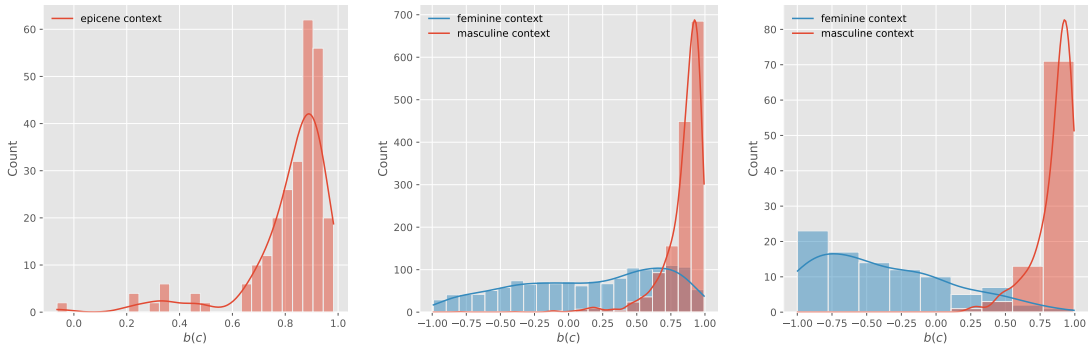
We have represented in Figure 2 the distribution of the most frequent prefixes in the sentences corrected by the TM.

	layer	encoder a	terminé	son	travail	.	eos
Gender weakening							
chaque surveillant a terminé son travail.	1	73.1	73.6	65.7	63.5	53.9	56.7
	6	71.0	71.4	70.4	68.2	71.2	69.7
Gender reinforcement							
le surveillant *français* a terminé son travail.	1	99.9	98.5	95.0	80.6	62.0	80.4
	6	100.0	99.7	99.7	98.9	98.8	96.9
Gender switch on direct object							
le surveillant a terminé son *travail*.	1	79.4	74.6	79.0	75.0	58.8	72.0
	6	90.3	88.8	89.2	85.3	86.2	83.3
le surveillant a terminé son *activité*.	1	80.5	75.5	78.6	62.6	57.6	67.2
	6	89.7	88.3	89.6	84.3	86.1	84.1
Syntactical distancing							
le surveillant *qui a chanté formidablement hier* a terminé son travail.	1	71.1	66.3	68.8	81.1	56.8	65.4
	6	91.5	91.0	90.5	86.8	81.2	82.1
Distractor							
.without gender weakening							
le surveillant *que cette femme critiquait* a terminé son travail.	1	65.7	66.6	69.3	79.50	62.8	68.5
	6	90.6	89.6	89.1	85.91	81.9	80.2
le surveillant *que cet homme critiquait* a terminé son travail.	1	65.4	67.0	68.7	80.0	63.4	68.2
	6	90.3	89.3	89.7	86.6	81.0	79.9
.with gender weakening							
chaque surveillant *que cet homme critiquait* a terminé son travail.	1	63.1	63.5	64.3	62.4	56.2	55.8
	6	72.1	71.4	69.7	69.9	71.8	69.2
chaque surveillant *que cette femme critiquait* a terminé son travail.	1	63.3	64.6	65.9	63.4	55.4	55.2
	6	71.8	71.8	70.0	69.2	70.2	69.5

Table 5: Precision of probes for manipulations of the sentences of our corpus

Unmarked gender in prefix				Marked gender in prefix			
occupational noun	$p(\texttt{his}\|c)$	$p(\texttt{her}\|c)$	$b(c)$				
Smallest value of $b(c)$							
_church ward en	0.0159	0.1638	-0.8231	_princess	0.0004	0.5320	-0.9985
_porn star	0.0020	0.0147	-0.7565	_duchess	0.0010	0.5148	-0.9962
_bill - poster	0.0012	0.0072	-0.7099	_unemployed _woman	0.0031	0.3992	-0.9845
_bill poster	0.0286	0.1403	-0.6614	_baroness	0.0016	0.1870	-0.9827
_ty le _layer	0.0026	0.0126	-0.6600	_office _lady	0.0025	0.2646	-0.9813
_bill _sticker	0.0012	0.0049	-0.6021	_mistress	0.0029	0.2178	-0.9738
_act uary	0.0024	0.0090	-0.5774	_shoes h ine _girl	0.0071	0.3027	-0.9543
_re stituto r	0.0109	0.0396	-0.5676	_literary _woman	0.0102	0.4206	-0.9526
_motel ier	0.0058	0.0201	-0.5534	_quilt ing _woman	0.0079	0.2325	-0.9346
_ped ic ure	0.0019	0.0066	-0.5475	_actress	0.0193	0.4927	-0.9246
Largest value of $b(c)$							
_subscriber	0.0494	0.0003	0.9861	_quarry man	0.0791	0.0023	0.9434
_hydraulic _engineer	0.1740	0.0012	0.9861	_mid ship man	0.1943	0.0053	0.9473
_energy _engineer	0.2265	0.0014	0.9878	_self - taught _man	0.3497	0.0089	0.9506
_golf er	0.1807	0.0010	0.9886	_delivery _guy	0.1944	0.0047	0.9526
_visitor	0.1904	0.0011	0.9887	_repair man	0.0982	0.0023	0.9533
_cellar _worker	0.3030	0.0017	0.9887	_coal maker	0.1074	0.0023	0.9573
_user	0.0684	0.0004	0.9894	_railway man	0.1087	0.0021	0.9616
_dealer	0.1135	0.0006	0.9900	_emperor	0.2765	0.0050	0.9642
_buyer	0.0850	0.0004	0.9909	_baron	0.1851	0.0031	0.9674
_player	0.0782	0.0002	0.9952	_salesman	0.1724	0.0015	0.9829

Table 6: Occupational nouns with the smallest and largest values of b. We keep the segmentation into sub-lexical units.



(a) No gender mark in the source and target

(b) Gender mark only in the source

(c) Gender mark in the source and target

Figure 2: Distribution of $b(c)$ for a translation model according to the presence of gender marks in the source and the target sentences.

occupational noun	cooc. with `her`	cooc. with `his`
Preference for `his`		
_subscriber	1	8
_hydraulic _engineer	71	284
_energy _engineer	71	284
_golf er	1	15
_visitor	77	206
_cellar _worker	144	247
_user	118	617
_dealer	5	66
_buyer	6	93
_player	131	602

Table 7: Epicene occupational nouns that generate `his` with a high probability and the number of sentences in which they co-occur with each personal pronoun.

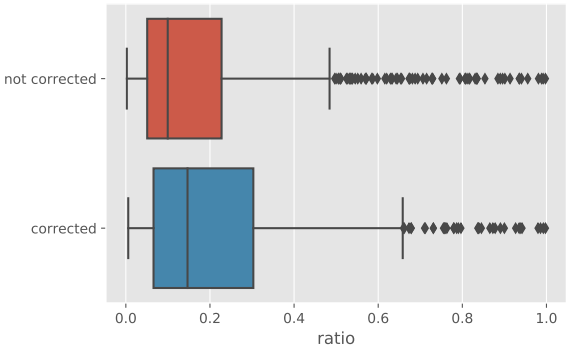


Figure 3: Distribution of ratio between $p(\texttt{her})$ and $p(\texttt{his})$ for sentences in which the LM has assigned the highest probability to the 'wrong' pronoun; 'corrected' correspond to sentences in which the translation model assigns a higher probability to the correct pronoun and 'not corrected' when this is not the case.

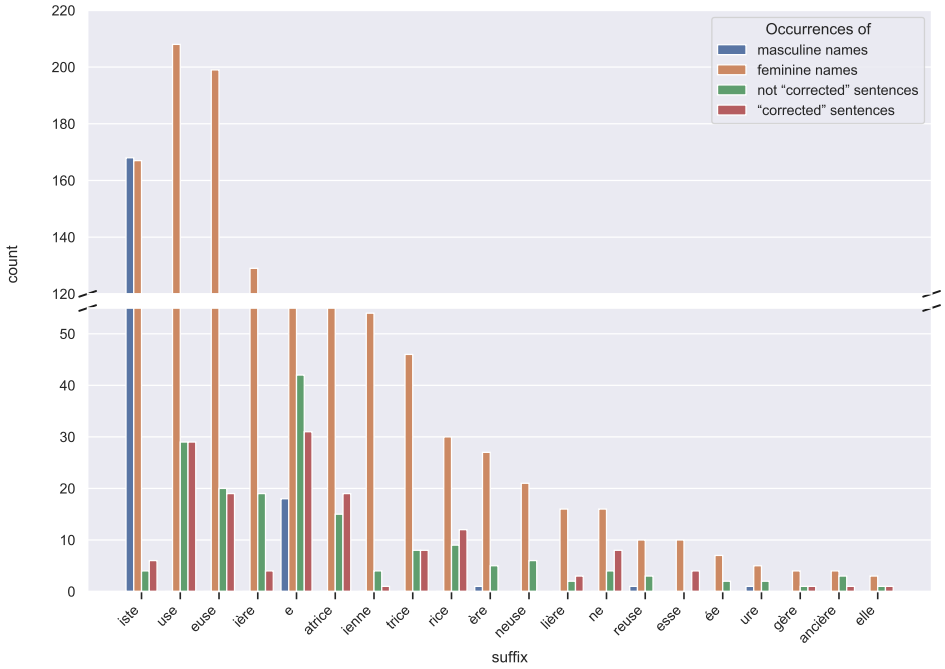


Figure 4: Most frequent BPE suffixes in our test set broken down according to the occupational name gender and whether taking into account the source sentence improve the prediction of the possessive pronoun (“corrected” sentences) or not (“not corrected” sentences).