# Middlebury CS 457: Gender Bias in Machine Translation
## Gender Inference in Spanish to English Translations

**Lauren Clarke**
Middlebury College '24.5
lclarke@middlebury.edu

**Christopher Fridlington**
Middlebury College '24
cfridlington@middlebury.edu

## 1 Introduction

Despite bias in machine translation being well studied in the field of Natural Language Processing, research often focuses on the gender bias in translations from English to other romance languages, as many parts of language (e.g. adjectives, nouns, etc) that are mostly or completely gender-neutral in English are inherently gendered in other romance languages. As a result of this, much of the existing research focuses on machine translation of English and Spanish is specifically analyzing the bias from English to Spanish, leaving a gap in the research from Spanish to English.

In contrast, our research project focuses on gender bias specifically from Spanish to English, particularly on the inference of pronouns in three contexts (`No`, `Limited`, and `Full Context`), with varying amounts of gender information embedded in each sentence.

For `No` and `Limited Contexts`, we calculate the percentage of inferred gender for both masculine and feminine, i.e. how often the male gender is inferred versus the female.

For `Full Context`, we calculate the percentage of total correct vs incorrect translations. For the incorrect translations, we then analyse both the specific sentence classification of the mistranslation (e.g. `male-male`, `female-male`, etc.), and calculate a bias differential by averaging—for all mistranslations within that sentence classification—the difference in percent likelihood of the actual output gender versus the expected output gender.

After all of these calculations, we analyze the results to see what gender bias, if any, there is, and why we think it may manifest in the way it does.

## 2 Literature Review

Two of the most important initial questions to conduct a study such as ours are "how do you define gender bias in machine translation" and "how can you measure it?" These questions are both introduced and addressed in (Farkas and Németh, 2022). They describe gender bias as "an amplification of existing occupational gender segregation," and strive to measure the extent of this bias in machine translation by comparing against a benchmark, for which they use three metrics: "the proportion of male and female workers of each occupation in Hungary..."—as their study looked at translations from Hungarian to English—"...the proportion of male and female workers of each occupation in the USA, and the masculinity/ femininity score of occupations" (Farkas and Németh, 2022).

Another article we took inspiration from was (Lopez-Medel, 2021). Lopez Medel analyzed how different contexts of the language affect a translation, specifically how the translation is gendered. One example of this was the use of an adjective in a variety of "I am a [adjective] + [profession noun]" sentences: the simple "I am a [profession noun] offered both feminine and masculine translations, while for some reason the adjective only chose one translation from those two. Other examples of these varied contexts are unfinished personal noun sentences to see with what gender the machine would finish the sentences, e.g. "Es ingenier. . . >He is an engineer", as well as looking at the suggested words for "He is a. . . " versus "She is a. . . ". Another interesting note she made, as a throw-away note mentioned in one sentence, is that "gender results varied with punctuation, [so] all sentences were capitalised and a full stop added (es colega>is a colleague, but Es colega.>He's a colleague.)" (Lopez-Medel, 2021). We made sure to make note of this, as it is a problem we had run into, and now include a step to account for this in our pre-processing.

The article (Attanasio et al., 2023) proposes some methods for actually mitigating the bias in machine translation. One of these is using instruction-tuned models and using prompting in conjunction with the interpretability scores they created to determine bias in machine translation. While this is perhaps a step beyond what we will be doing in our project, it gives a variety of good next steps to take once the bias has been identified. We may even get to implementing some of the proposed methods to use as a check for our bias identification.

The paper (Stanovsky et al., 2019) presents an analysis of the bias in machine translation. To test bias in different translations, the authors created a composite set of data, from the Winogender and WinoBias text sets, that they have dubbed WinoMT. They translated the sentences in this set and calculated the accuracy biased on the number of sentences that correctly preserved the original gender present. They ultimately found that all existing systems exhibit biases. Interestingly, they also experimented "fighting bias with bias" by introducing adjectives that are sterotypically gendered such as "handsome" or "pretty" (Stanovsky et al., 2019). This research was especially helpful given that we used WinoMT as one of our initial testing datasets. This research also provided us an understanding of the circumstances that led to the dataset's creation, and the way it was used to analyze bias.

One paper that served as a comprehensive guide to the study of bias in machine translation was (Savoldi et al., 2021). The authors first begin with a history and discussion of the term bias itself, how it is understood in the NLP field, and different types of bias that can be incurred. Building on this foundational knowledge of bias, the authors shift to exploring the ways researchers understand bias in the field of NLP. Finally, the paper moves to a discussion of existing metrics used to highlight bias, a selection of research on different types of bias in NLP, and methods for removing biases from NLP systems. This analysis provided us with a broader framework to use when thinking about bias in machine translation, which will guide us in our own project. Additionally, the paper's broad survey of the field introduced us to a significant amount of information that that we investigated to see what was relevant to our project.

Another paper that we found useful, (Stewart and Mihalcea, 2024), discusses another form of bias in machine translation—that against same gender relationships. In the paper, they assess the extent of the bias, which proved to be quite significant. While the findings of the paper were intriguing, the methods used to create the sentence data was perhaps the most useful to us. Instead of relying on human-labeled data, they created a series of template sentences with parameters that could be changed. Each parameter in the sentence was then changed to produce a series of sentences that represent every combination of occupation, gender, and relationship type. This method of generating texts to translate informed us in the creation of our own translation data as it showed us how to easily create template sentences for our three different translation contexts, from which we could generate a multiple of each template sentence by changing the parameters of occupation and gender.

## 3 Data

We use a .csv file of sentence parameters to generate sentences, all of which were written by Chris and/or Lauren, then translated to Spanish by Lauren.

The sentence parameter file is split into 10 rows:

- `es_male_occ`
- `es_fem_occ`
- `en_occ`
- `amb_names`
- `es_verb_phrase_occ`
- `en_verb_phrase_occ`
- `es_verb_phrase_male_rel`
- `es_verb_phrase_fem_rel`
- `en_verb_phrase_male_rel`
- `en_verb_phrase_fem_rel`

There are a total of **25** individual occupations used, for which the English genderless versions can be found in `en_occ`, and the gendered Spanish equivalents in `es_male_occ` and `es_fem_occ`. There are **14** gender-ambiguous names in `amb_names`.

There are **22** occupational verb phrases, the English in `en_verb_phrase_occ` and the Spanish in `es_verb_phrase_occ`. There are **17** gendered romantic verb phrases, for which the English versions can be found in `en_verb_phrase_male_rel` and `en_verb_phrase_fem_rel`, and the Spanish equivalents in `es_verb_phrase_male_rel` and `es_verb_phrase_fem_rel`.

The full combination of these subjects to verb phrases resulted in a total of **3013** sentences for translation, of which **1536** were `Full Context`, **1424** were `Limited Context`, and **53** were

No Context. We classify the Full Context sentences as sentences which contain a gender for both the subject and the verb phrase (i.e. the occupational subjects with the relationship phrases). The Limited Context sentences are sentences which contain either an occupational subject (with "su", the gender-neutral possessive, before it) or a gender-ambiguous name and a relationship phrase. These are Limited Context because there is some information—either a name or a gendered occupation—but the name is not explicitly gendered, and the gendered occupation is not related to the gender neutral "su" put before it. The No Context sentences are simply the two types of verb phrases by themselves. All of these sentences were post-processed to have correct capitalization and punctuation as to get the best translation results (Lopez-Medel, 2021).

Example sentences were:

Full Context

- El maestro besó a su marido.
  [The teacher kissed his husband.]

- La ingeniera fue de vacaciones con su novia.
  [The engineer went on vacation with her girlfriend.]

Limited Context

- Su jefe se tomó unos días para navegar alrededor del mundo.
  [His/Her/Their boss took a few days off to sail around the world.]

No Context

- Salvó a la familia del edificio en llamas.
  [He/She/They saved the family from the burning building.]

# 4 Methods

## 4.1 Batching

As mentioned in the previous section, the data for the project is generated from a set of parameters and saved to a .csv file. We then read in this data, translate it, and analyze the results. Additionally, to aid in development, several versions of the sentences file are created with varying lengths to allow for quick tests of functionality. These scopes are passed as a command line argument automatically when the mini_test.sbatch, test.sbatch, and main.sbatch are called. The resulting argument points to the associated .csv file, which is opened and read in as batches. This batching process organizes the inputted file into a dictionary with labeled contexts and sentence classifications. For example, it stores the sentence "**El** abogado besa a su **novio**" [The lawyer kisses **his boyfriend**] in the full_context section of the dictionary under the male-male relationship classification.

## 4.2 Translation

When the batching is complete, the sentences are sequentially iterated through to be translated. For the translation, a model and tokenizer object are created from the opus-mt-es-en model. Each sentence is then passed to a function that uses the tokenizer to create input_ids and attention_masks that are used by the model for generation. The output of this model is then iterated through by passing a list of decoder_ids consecutively to the model. This produces logits to which a softmax is applied to get the probability for each token_id. These are sorted into a list of the top 20 possibilities, and saved to a dictionary that represents the likelihood of each word generation. The most likely token_id is decoded to get the next word in the translation. It is also used as the key in this dictionary for convenient retrieval during analysis. After iterating through each sentence, we have the decoder_id, which we used to retrieve and return dictionary with the sentence translations and the generated probabilities.

## 4.3 Analysis

After the text is translated and the associated probabilities are generated, we analyze the results of each sentence translation. There are two cases that we use to characterize the three contexts: one for full_context, and the other for limited_context and no_context.

To begin with the second case that encapsulates the two limited_context and no_contexts, analysis is simple. There is no gender explicitly provided in the original sentence, so we simply search the translated sentence for pronouns using a master list of all possible gendered pronouns and possessives, and recorded if the pronoun inferred is masculine or feminine. The inference for each sentence is then totaled to produce a ratio of masculine-to-feminine inferred sentences, and a percentage of each.

For the first case, which, despite being solely focused on the `full_context` translations, is much more complicated, the pronouns found in the translated sentences are compared to a set of pronouns representing the expected gendered output. For example, if the sentence is `male-male`, it checks to see if the output contains "he/him/his" pronouns. If it does, it was correctly translated. If not, it was incorrectly translated. If translated incorrectly, we calculate the difference in the probability of the translated pronoun and the actual pronoun by subtracting the actual pronoun's probability from the translated pronoun's probability. This difference was then averaged over all of the incorrect translations for that specific sentence classification, e.g. `female-female`, to get the average difference in probability of the output pronoun vs the correct pronoun. This is shown as Diff. Prob. in table 1.

### 4.4 Output & Summary

The results from the analysis functions are organized, formatted, and written to a markdown file. To summarize, the sentences are read from the initial .csv file, batched into contexts and classifications, translated, analyzed, and then output to a .md file. A graphical depiction of the methods can be seen in fig. 1.

For future work, we would like to consider more than just the male/female binary. As it stands, we also do not account for mistranslations of the gendered noun in the relationship verbs, e.g. if the sentence "**El** maestro ama a su **esposo**" [The teacher loves **his** **husband**] was mistranslated as "...loves his wife" rather than the more common mistranslation of "...loves her husband", we currently do not account for that mistranslation in our model. This is because doing so would require cataloging every individual translation and associating the correct individual translation with it, rather than just searching for a known pronoun/possessive, which would take a lot more manual work.

For further possible future work, this same structure could be utilized to look at gender bias in other languages, models, and datasets. We would also like to optimize our methods more and parallelize our work for a faster run-time. One last thing we would like to do is analyze which specific sentences from the `No Context` and `Limited Contexts` are inferred masculine, and which are inferred feminine, and see if there is a pattern to them. It would be interesting to see if certain types of sentences are typically inferred as a certain gender.

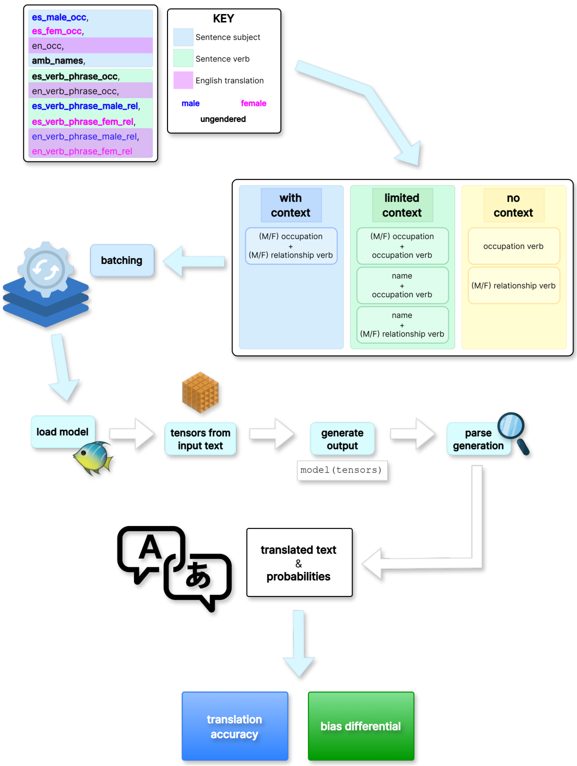

Figure 1: Project Workflow

## 5 Results

All of the results of our tests can be found in their original /md output in `./results/` in our final project folder, but are detailed below for analysis and completion.

The results of the `Full Context` sentences can be seen in fig. 1. The results of the `Limited Context` sentences are that **59.43%** of the sentences were inferred masculine, while **40.57%** were inferred feminine. The results of the `No Context` sentences are that **68.63%** of the sentences were inferred masculine, while **31.37%** were inferred feminine.

The `No Context` and `Limited Context` results show an almost a **38%** greater likelihood to be inferred as masculine than feminine without any context, and a **19%** greater likelihood if there is even some context. This is interesting as it shows a heavy masculine leaning in a binary translation of a genderless sentence.

Some of the most interesting results, however, are those of the `Full Context` calculations. Our findings were that the model of machine translation we used seemed to be very biased against same-sex relationships, especially so for `male-male`. The majority of `male-male` sentences were mistranslated **(65/384)**, and had a high bias differential probability of **61.28%**. This means that for the incorrectly translated male-male sentences, the model considered the feminine possessive **61.28%** more likely than the masculine possessive.

The result that puzzled us most was the negative number for the female-female probability differential. This meant that the model was not always selecting the most probably word (in other words, it was not a greedy model), and instead was performing some other post-processing on the translation that introduced a bias in gender. While this assumption proved correct, subsequent attempts to make the model greedy ultimately produced translations without pronouns or with other possessives that were unexpected and unusual in the context, such as "your". Therefore, even though this solved our problem of negative differential probabilities, it caused other unforeseen issues during analysis which impacted our ability to understand if a sentence was gendered correctly. In these cases, our analysis functions were unable to parse for a pronoun and thus marked all those sentences as incorrect. This skewed the results significantly, in a way that was not meaningful (see ./results/results_all_6785.md), so we ultimately reverted to our original implementation.

Throughout the project, we also experimented with performance optimizations to decrease runtime. The most significant of these involved reading over the Ada Cluster Guidelines and the Ada GitHub to use multiple threads on the cluster for parallelization. Whilst we got an example with the tokenizer working, we learned that the MarianMTModel is not compatible with the the Pool function from Python's multiprocessing package. This meant that it was out of scope to get the translations working on multiple threads, so we instead focused our efforts on optimizing model initiation, which was previously in a loop, to improve performance. The result of this, along with other small performance improvements, led to a much faster model that now runs in approximately thirty-five minutes, as opposed to a previous two-plus hours.

These results were sadly fairly in line with our initial hypothesis that there would be both a misogynistic and homophobic leaning to the machine translations. The results were, however, much more drastically biased against same-sex relationships than we thought they would be, especially those of male-male.

| Classification | Corr. | Incorr. | Diff. Prob. |
|---|---|---|---|
| Male-Male | 65 | 319 | 61.28 % |
| Male-Female | 382 | 2 | 51.11 % |
| Female-Male | 384 | 0 | 0.0 % |
| Female-Female | 223 | 161 | -12.53 % |

Table 1: Full Context Translation Results

## 6 Ethical Considerations

As it stands, we only account for the masculine/feminine binary, and do not include any of the few gender neutral words in Spanish. We also therefore never expect an output of they/their, and do not account for this edge case, only ever expecting he/him/his or she/her/hers. This is a big gap in our research, and excludes nonbinary people. The main reason for this oversight is the limited time that we had to complete this project; we reasoned that given a gendered noun in Spanish, we would see a genderless translation so infrequently in English that the case was, for our purposes, negligible. This is, however, a considerable oversight that, if we had more time, we would like to correct.

We do not think there is any bias inherent in the way we generated the sentences from these parameters, as we merely combine every possible combination per sentence classification. There is, however, also possible bias in the parameters that we came up with to generate sentences. We tried to think about occupations or actions that we felt might have some gender associated with them, and specifically include those phrases or occupations for what we felt would create interesting sentences with an output that may show bias. However, in doing so, we showed our own gender biases toward the occupations, and it may show in the results of our work.

## References

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. *Preprint*, arXiv:2310.12127.

Anna Farkas and Renáta Németh. 2022. How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences & Humanities Open*, 5(1):100239.

Maria Lopez-Medel. 2021. Gender bias in machine translation: an analysis of google translate. *Academia Letters*.

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. *ArXiv*, abs/1906.00591.

Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? assessing bias against same-gender relationships in machine translation. *Preprint*, arXiv:2401.04972.