The GENDER-GAP Pipeline: A Gender-Aware Polyglot Pipeline for Gender Characterisation in 55 Languages

Benjamin Muller, Belen Alastruey, Prangthip Hansanti, Elahe Kalbassi,
Christophe Ropers, Eric Michael Smith, Adina Williams, Luke Zettlemoyer,
Pierre Andrews* and Marta R. Costa-jussà*
FAIR, Meta
{benjaminmuller,alastruey,prangthiphansanti,ekalbassi,chrisopers
ems,adinawilliams,lsz,mortimer,costajussa}@meta.com

Abstract

Gender biases in language generation systems are challenging to mitigate. One possible source for these biases is gender representation disparities in the training and evaluation data. Despite recent progress in documenting this problem and many attempts at mitigating it, we still lack shared methodology and tooling to report gender representation in large datasets. Such quantitative reporting will enable further mitigation, e.g., via data augmentation. This paper describes the GENDER-GAP Pipeline (for **G**ender-**A**ware **P**olyglot Pipeline), an automatic pipeline to characterize gender representation in large-scale datasets for 55 languages. The pipeline uses a multilingual lexicon of gendered person-nouns to quantify the gender representation in text. We showcase it to report gender representation in WMT[1] training data and development data for the News task, confirming that current data is skewed towards masculine representation. Having unbalanced datasets may indirectly optimize our systems towards outperforming one gender over the others. We suggest introducing our gender quantification pipeline in current datasets and, ideally, modifying them toward a balanced representation.[2]

1 Introduction

Despite their widespread adoption, Natural Language Processing (NLP) systems are typically trained on data with social and demographic biases. Such biases inevitably propagate to our models and their generated outputs, e.g., by over-representing a given demographic group and under-representing others. It is, therefore, critical to measure, report, and design methods to mitigate these biases, before they can be encoded and potentially amplified during training (Foulds et al., 2020; Wang and Russakovsky, 2021).

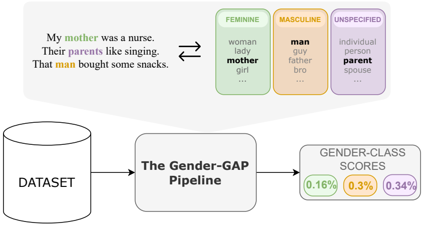

Figure 1: The Gender-GAP Pipeline works by identifying gendered lexical terms and reporting statistics on these lexical matching.

This paper focuses on quantifying gender representation in highly multilingual data (see Figure 1), in particular, for the task of machine translation. Gender is a complex concept that can be defined in many ways depending on the field of study, language or culture (Chandra et al., 1981; Hellinger and Bussmann, 2001; Kramer, 2020). We discuss and define gender in Section 3.1. However, briefly, we define gender bias as the systematic unequal treatment based on one's gender (Blodgett et al., 2020; Stanczak and Augenstein, 2021). Gender bias, when it impacts training data, may decrease the performance of the system on certain gender groups (Hovy et al., 2020). When impacting evaluation data, it may push the system designers to deploy a system that causes harm by favoring one group over others (Mehrabi et al., 2021). For example, a system that translates text that includes feminine nouns more poorly than text with masculine nouns may lead the end users to miss important information or misunderstand the sentence (Savoldi et al., 2021). A system that inaccurately translates a gender-neutral sentence in English e.g. *they are professors* to a sentence with a masculine noun *ils sont professeurs* in French may also lead to serious representational harm.

We propose the GENDER-GAP pipeline to quantify gender representation bias of multilingual texts

[1] http://www2.statmt.org/wmt23/

[2] The GENDER-GAP pipeline is available at https://github.com/facebookresearch/ResponsibleNLP/tree/main/gender_gap_pipeline

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 536–550
December 6–7, 2023. ©2023 Association for Computational Linguistics

using lexical matching as a proxy. Our pipeline can be seen as two main modules.

First, we build a multilingual gender lexicon: starting from a list of about 30 English nouns extracted from the HolisticBias dataset (Smith et al., 2022), split into 3 gendered classes—masculine, feminine, and unspecified. We manually translate them and reassign them to the appropriate gender class for each target language (e.g. "grandfathers", masculine in English, becomes "abuelos", masculine and unspecified in Spanish). Our list is restricted to nouns that refer to people (e.g. man, woman, individual) or to kinship relationships (e.g. dad, mom, parent). Most languages, including genderless languages (Prewitt-Freilino et al., 2012) (e.g. Finnish, Turkish) encode genders through kinship relationships and person terms (Savoldi et al., 2021). For this reason, focusing on a restricted list of kinship and person nouns allow us to scale our lexicon to 55 languages.

Second, we arrive at a straightforward and easily comparable gender distribution by using a word matching counter. Based on our newly collected multilingual lexicon, our pipeline segments each input sentence at the word-level using Stanza (Qi et al., 2020), a state-of-the-art word segmentation tool, and counts the number of occurrences of words in each gender class. As a result, we obtain a gender distribution across 55 languages. In summary, our contribution is threefold:

- We collect and release an aligned multilingual lexicon that can support measurement of the representation of genders in 55 languages.

- We introduce the Gender-Aware Polyglot pipeline (GENDER-GAP), a lexical matching pipeline, and describe the gender distribution observed in popular machine translation training and evaluation data. On average, all three analyzed datasets are biased toward the masculine gender. We find the gender representations to be domain- and language-specific. Additionally, using the GENDER-GAP pipeline, we can discover sentences that have been translated with a gender bias.

- We release our pipeline and recommend the reporting of gender representations in machine translation training and evaluation datasets to improve awareness on potential gender biases.

2 Related work

The study of biases in text has become more important in recent years, with Large Language Models (LLMs) displaying bias against people depending on their demographics and identity. As a testament to the importance of this topic, many recent papers, including those introducing GPT-3 and 4 (Brown et al., 2020; OpenAI, 2023), PaLM 1 and 2 (Chowdhery et al., 2022; Anil et al., 2023), LLaMa 1 and 2 (Touvron et al., 2023a,b), analyze how such biases affect their model outputs. Some works even discuss frequencies of gendered terms in their pretraining corpora (Anil et al., 2023; Touvron et al., 2023b), as this can affect downstream generation. Despite this acknowledgment of the issue, general purpose tools to measure demographic biases are still fairly rare, and so far have mainly been in English.

However, some have begun to measure demographic biases beyond English. Smith et al. (2022) built a comprehensive analysis dataset covering 13 demographic groups and Costa-jussà et al. (2023) extended it to the multilingual setting. Specific to Machine Translation, Savoldi et al. (2021) discussed best practices in reporting gender bias. Several works (Stanovsky et al., 2019; Prates et al., 2020; Renduchintala et al., 2021; Renduchintala and Williams, 2022) have explored metrics for exposing failures in automatically translating pronoun and occupations, and some have even explored MT model training (Escudé Font and Costa-jussà, 2019; Stafanovičs et al., 2020) or fine-tuning (Saunders et al., 2020; Corral and Saralegi, 2022; Costa-jussà and de Jorge, 2020) or both (Choubey et al., 2021) to lessen the effect of gender-related biases. More than this, there are initiatives that provide toolkits to generate multilingual balanced datasets in terms of gender (Costa-jussà et al., 2019) from Wikipedia and even balanced in gender within occupations (Costa-jussà et al., 2022).

However, despite the progress made, most of these resources only cover a handful of languages—the community still lacks easy to use, open-source toolkits to measure biases across a large number of languages. In this work, we address this need by showcasing, GENDER-GAP, a lexical matching pipeline to measure gender distribution across 55 languages.

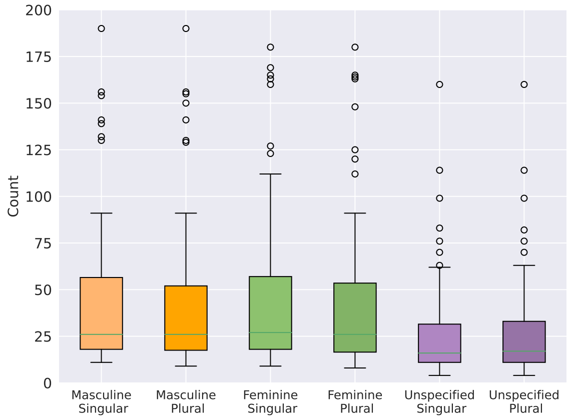

Figure 2: Distribution of the number of words in our proposed multilingual gender lexicon per language across gender-classes and number (i.e. singular and plural)

3 Proposed Data Collection and Pipeline

3.1 Defining Gender

Gender is a complex topic that can be defined in many different ways depending on the field of studies and the context (Hellinger and Bussmann, 2001). In this work, we approach gender from two perspectives:

First, linguistic gender (Corbett, 2013; Cao and Daumé III, 2020; Kramer, 2020; Stanczak and Augenstein, 2021) corresponds to the classification of linguistic units, such as words, into categories based on the gender information they provide. Linguistic gender refers to overlapping notions, such as *grammatical*, and *semantic* gender, depending on the properties of the language. Grammatical gender implies the classification of nouns, adjectives, and other parts of speech into categories based on their morphosyntactic properties. In many languages, grammatical gender morphology appears on all nouns, regardless of whether they refer to persons, animals, plants, or inanimate objects (e.g., “il libro” *the book* is a masculine noun in Italian). Semantic gender (Corbett, 1991) refers to the existence of lexical units whose meaning is associated with a specific cultural notion of peoples’ gender(s). For instance, in English, the word “men” associated with masculine traits, “woman” with feminine ones, etc. Semantic gender then may be present in languages that do not morphologically mark grammatical gender, such as English, Turkish, or Mandarin Chinese. In languages that do mark grammatical gender, grammatical and semantic gender do not always match: for example, in German, the word for girl “Mädchen” is grammatically neuter, but refers to a person which would fall into our ’feminine’ class based on its meaning. For our purposes, we use semantic gender classes in our multilingual lexicon, since we are interested in gender representation.

Our goal is to build and foster inclusive NLP technologies that do not carry, replicate, or amplify social gender biases, which can impact end users and societies negatively by affecting representations of specific groups. However, there are social meanings of gender that are not readily accessible in text, so, we use semantic gender on human words as a proxy for social gender.

Social gender refers to gender as a social construct based on cultural norms and identity (Ackerman 2019; Cao and Daumé III, 2020; Stanczak and Augenstein, 2021; Duignan, 2023). As highlighted by Ackerman 2019, social gender is defined as the internal gender experienced by a given human individual. For this reason, data-driven analysis of genders in large corpora can only relate to social gender indirectly through linguistic notions of gender(s).[3] We assume for our purposes that a list of gendered words can be used to approximate some important aspects of social gender for the purposes of measuring representation disparities.

3.2 Aligned Gendered Multilingual Lexicon

To measure gender distribution across 55 languages, we first build a multilingual lexicon. We want this lexicon to be as aligned as possible across languages while also encoding language-specific gender linguistic phenomena.

Languages Our lexicon is available in 55 typologically and phylogenetically diverse languages such as English, Finnish, Zulu, Vietnamese, Ganda, Japanese or Lithuanian, spanning 15 distinct scripts. We report the complete list of languages in Figure 6.

Gender Classes We define three semantic gender classes: masculine, feminine and unspecified. The unspecified class aggregates nouns of different sorts. It mainly capture nouns that do not explicitly encode any particular gender (e.g. “person” is considered unspecified in English). For this reason,

[3] We recall that gender is distinct from sex which refers to collections of biological properties of individuals such as genes (e.g., chromosomes), phenotypes (e.g., anatomy) (Council of Europe, 2023). See Butler (2011) for a discussion of additional factors that complicate this view.

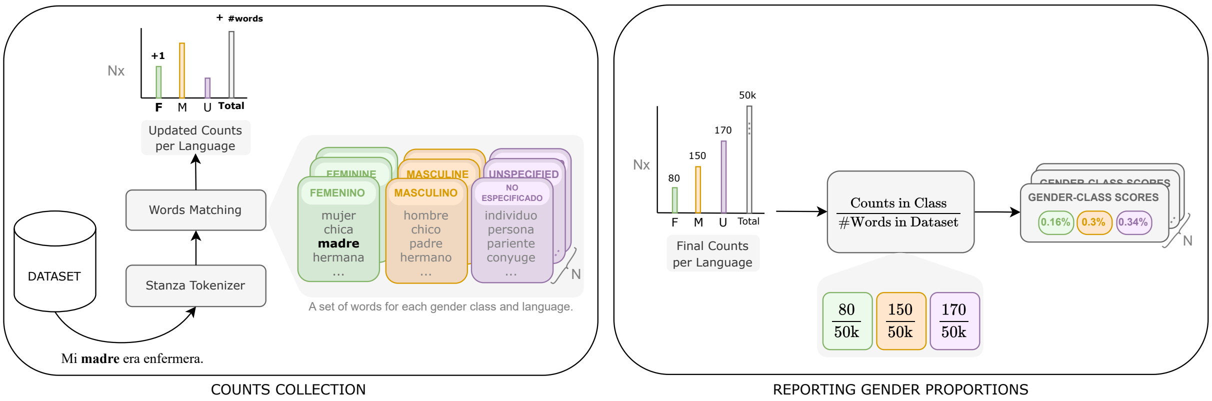

Figure 3: Diagram of the GENDER-GAP pipeline. In the first stage, we process each sentence of the 55 supported languages of the dataset and count the word matches for each category. Once this step is completed, we compute a gender-class score which corresponds to the proportion of gendered noun matched within all the words in the dataset.

"unspecified" can be seen as aggregating masculine, feminine and non-binary genders (Herdt, 2020).

While there exist more complex gender lexica as discussed in Stanczak and Augenstein (2021), they are focused on English and are not always easily translated. Because our goal is to provide a methodology that can be used to evaluate bias across multiple languages, we take a more pared down lexical approach.

Lexicon creation We start by defining a list of about ten, high frequency person nouns per gender class in English. Each noun is found in both its singular and plural form. To find a list of nouns that is as universal as possible, we restrict this list of persons such as masculine "man", feminine "woman", and "person" and synonyms (e.g. "individual") that we complement with kinship terms classified by gender (e.g., masculine "father", feminine "mother", neutral "parent"). Our list corresponds to the one defined in the previous work of HolisticBias (Smith et al., 2022), which is only available in English.[4]

We then translate these nouns into the other languages by reassigning them to the appropriate gender class. A noun in a given gender class may be part of another class (or multiple other classes) in another language. For instance "grandparents" (masculine, plural) becomes "abuelos" in Spanish which is both masculine and unspecified genders.

The English-language source list is passed on translators who are native speakers of the target language, with language proficiency at CEFR[5] level C2 in the source language. For all languages, translators are asked to provide equivalent singular and plural terms in their respective native language, except if any of the source concepts do not exist in the language. For example, not all languages use a distinctive, gender-agnostic term such as the English term *sibling*, distinctively from either *brother* or *sister*. We also consider that the reverse can be true (i.e. that the target language may have more than one term to translate one of the English terms in the source list), and give the translators the possibility to provide additional translations in such cases. For instance, when we translate *women* into Korean we get : "여성들" and "여인들".

Additionally, translators are asked to consider the terms in the source list as lemmas (or headwords in dictionary entries) and, if applicable to the given language, to provide relevant morphologically derived forms, including cases and gendered forms. Finally, translators are also encouraged to provide terms covering all language registers, which is necessary because some languages (e.g., Thai or Korean, among others) use several different terms at various levels of formality.

We are cognizant of the fact that this approach presents several limitations. The first limitation occurs when a term could be said to fall into both the unspecified and one of the gendered categories. For example, the term Spanish *padres* can be used to mean both *fathers* or *parents*. Some speakers also use the singular form to mean *parent* (and not necessarily *father*). The second limitation applies to

[4] We use the gender noun list v1.1 from HolisticBias

[5] https://coe.int/en/web/common-european-framework-reference-languages/level-descriptions retrieved 2023-07-24

languages that are closer to the synthetic end of the analytic-synthetic spectrum; i.e. languages that are agglutinative or highly fusional (e.g., Zulu, Uzbek, Estonian). This approach may not allow for the detection of many agglutinated or fused word forms. Finally, due to the templated, context-free nature of the lexicon, one term was particularly difficult to disambiguate: *veteran*, which can be used to refer to a soldier or a seasoned professional.[6] Cultural differences also had to be considered in addition to the above ambiguity; for example, Japanese translators mentioned the fact that the Japanese equivalent of the term was infrequently used with the first meaning cited above.[7]

Lexicon statistics In Figure 2 we can see the obtained data distribution across number and gender for the different languages. We notice a few outliers. As described above, translators are asked to provide relevant morphologically derived forms. This makes the number of nouns in Estonian to be 7 times larger than the average. For instance, "woman" is translated into *naine* "a woman", *naise* "of a woman", *naisele* "to a woman", etc.

3.3 Proposed Pipeline

Figure 3 shows a diagram of the GENDER-GAP pipeline. In the first stage or the counts collection, we work at the sentence level for the NTREX and FLORES-200 and at the document level for Common Crawl. We segment each sample at the word level using `Stanza` tokenizer available in the given language (Qi et al., 2020) except for Cantonese (yue) for which we reuse the model available for simplified Chinese (zh-hans) and Thai for which we use `PyThaiNLP`.[8] For the rest of the languages we use simple `nltk`[9] typographic tokenizer (based on white-space and punctuation marks). We then count and increment a gender-class counter anytime we match a word in the list of words representative of this class. For instance, in the sentence "my mother was a nurse" the pipeline will add +1 to the feminine counter (due to lexical match of "mother").

Once this process has been done for each sentence in the dataset we move to the second stage or the reporting of gender proportions where we define a score for each gender-class by dividing the gender-class count by the total number of words in the dataset. By doing so, the final gender score does not depend on any defined linguistic macro-unit such as sentences or documents lengths but only on the word-level tokenization.

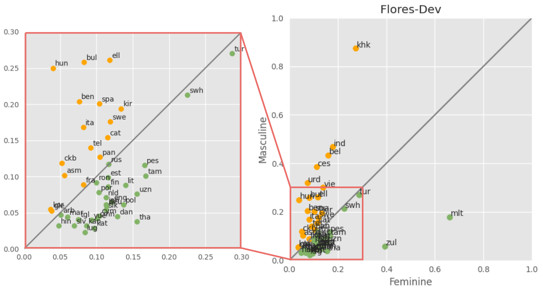

Figure 4: Gender Representation in % of the total tokens in the FLORES dataset dev split.

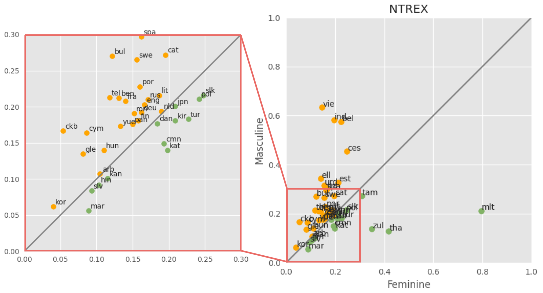

Figure 5: Gender Representation in % of the total tokens in the NTREX dataset.

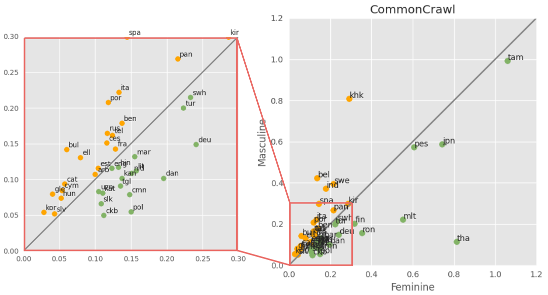

Figure 6: Gender Representation in % of the total tokens in the Common Crawl dataset.

4 Experiments

To showcase GENDER-GAP, we run it on Common Crawl raw data and two popular machine translation evaluation datasets: FLORES-200 (NLLB Team et al., 2022) and NTREX-128 (Federmann et al., 2022). FLORES is a Wikipedia-based dataset including 3001 sentences translated from English to 200 languages. NTREX-128 is made of 1997

[6] https://www.merriam-webster.com/dictionary/veteran retrieved 2023-07-24

[7] See https://en.wikipedia.org/wiki/Article_9_of_the_Japanese_Constitution retrieved 2023-07-24

[8] https://pythainlp.github.io/docs/2.0/api/tokenize.html

[9] https://www.nltk.org/api/nltk.tokenize.html

Lang	Fem.	Masc.	Uns.	Δ(|Fem.-Masc.|)	%doc.
Flores DevTest.					
eng	0.121	0.065	**0.379**	0.056 (0.0003)	11.2
avg.	0.128	0.144	**0.302**	0.097 (0.0003)	10.1
NTREX					
eng	0.166	0.203	**0.379**	0.037 (0.0003)	15.5
avg.	0.180	0.224	**0.329**	0.099 (0.0003)	13.4
CommonCrawl					
eng	0.120	0.115	**0.243**	0.005 (0.0000)	9.4
avg.	0.212	**0.260**	0.251	0.136 (0.0003)	12.0

Table 1: % Gender Distribution in WMT Evaluation dataset. We report the English distribution and the average across all languages (standard deviation indicated between parenthesis). The full table is available in the appendix Table 3-5. We **bold** the most represented gender class, and underline the second most represented gender class. We define the the gender gap Δ defined as the absolute difference between the Feminine and Masculine scores. %doc. refers to Coverage.

sentences from news documents originally collected for WMT 2019 (Barrault et al., 2019) translated from English into 128 languages. Both these datasets are part of the corpora provided by the WMT shared task. In addition, we run the pipeline on a sample of Common Crawl.[10] Common Crawl is a snapshot of crawlable web data that is widely used in the NLP community thanks to the release of the CCNET corpora (Wenzek et al., 2020), the OSCAR corpus (Ortiz Suárez et al., 2019) and the C4 corpus (Raffel et al., 2019). It is used to train NLP systems like language and machine translation models. We run our pipeline on 100k documents for each language. Our pipeline supports 55 languages, and we run it on the intersection of these datasets with the set of supported languages.

5 Analysis

5.1 Quantitative Analysis

We report the average coverage and gender distribution in Table 1 along with the complete tables for the 55 languages in Table 3-5.

Coverage We first look at the number of samples for which at least one noun is found (cf. %doc in Table 1). We find that, on average, about 10% of samples match with at least a noun (between 10.1 and 13.4% depending on the dataset). We find that the coverage is the largest for Vietnamese (with up 45.7% of samples matched) and Thai (28.9% of samples matched) and the smallest for Korean (between 1.7% and 2.5% depending on the dataset). This shows that even though our lexicon is restricted to person nouns and kinship relationships, we are still covering a very large number of samples based on which we measure gender representations.

Gender Distribution Table 1 shows gender representation for masculine, feminine and unspecified. For better visualization, Figures 4, 5 and 6 report the % of masculine and feminine representation of the total tokens in FLORES, NTREX, and Common Crawl respectively.

On average, the masculine gender is more represented than the feminine in all three datasets. We find that NTREX is the dataset with the highest bias toward the masculine gender on average. Accounting for uncertainty, using the standard error to define a confidence interval,[11] we find that 30/45 languages are biased toward the masculine gender for NTREX. This includes languages like English, Arabic, French, Spanish, Vietnamese, and Panjabi. The rest of the languages are either balanced between masculine and feminine (i.e. Δ(|Fem.-Masc.|) is inferior to the confidence interval length) or biased toward the feminine gender. In addition, we find 16/54 languages biased toward the masculine gender for all three datasets suggesting an inherent gender bias in these languages. This includes several romance languages such as Spanish, French, Catalan and Italian along with Belarusian, Indonesian, and Panjabi.

Impact of Domains We find that 14/55 languages for which, the gender representation changes drastically across the different datasets. For instance, the gender differences are much larger in NTREX than in Common Crawl data. More specifically, in Lithuanian the distribution is skewed toward the masculine class for NTREX data, while it is skewed toward the feminine for Common Crawl data. For Danish, the gender representation is balanced for NTREX but skewed toward the Feminine class for Common Crawl data. This shows that domains highly impact gender representation. NTREX is based on news data, while

[10] https://commoncrawl.org/

[11] We consider that a given dataset in a language is biased toward a specific gender when the gap Δ(|Fem.-Masc.|) is higher than two times the standard error (ste.). This is equivalent to defining a confidence interval as $[r_g - 2ste, r_g + 2ste]$ given the gender score r_g with $g \in \{masc., fem.\}$. If Δ(|Fem.-Masc.|) is inferior to $2ste$, we consider the dataset to be gender balanced. ste is defined as $\frac{\sigma(fem-masc)}{\sqrt{n}}$ with n the number of words in the dataset and σ the standard deviation. See (Bulmer, 1979) for more details on these definitions.

Sentence 1: Omission of words/lexical variation	
Eng: shark injures 13-year-old on lobster dive in california	`masc.+= 0`
Spa: tiburón hiere a un **niño** de 13 años que buceaba en busca de langostas en california	`masc.+= 1`
Cat: un tauró fereix un **nen** de 13 anys mentre buscava llagostes a califòrnia	`masc.+= 1`
Sentence 2: Multiple translations and variation in part of speech	
Eng: [...] something increasingly demanded by younger shoppers.	`unspecified.+= 0`
Cat: [...] un aspecte cada cop més demanat pels consumidors més **joves**.	`unspecified.+= 1`
Sentence 3: Robust to typographic differences	
Eng: **mother**-of-three willoughby and **husband** dan baldwin have been close to jones and his **wife**	`fem.+= 2,masc.+= 1`
Cmn: [...]个孩子的母亲的威洛比及其丈夫 dan baldwin 十年来与琼斯及其妻子 tara保持[...]	`fem.+= 2,masc.+= 1`
Sentence 4: Synonyms	
Eng: [...] the owner of the lloyds pharmacy chain, for £125m, three years ago.	`masc.+= 0`
Vie: [...] chù sõ hũu cũa chuõi nhà thuòc lloyds, vói giá 125 triẽu bàng vào **ba** năm trùóžc.	`masc.+= 1`

Table 2: Selected examples of gender representation across parallel sentences between English and multiple target languages (based on the NTREX dataset). Detected gendered nouns in bold/underlined. We indicate the counter incremented by the pipeline for the three gender classes (feminine, masculine and unspecified) next to each sentence when there is at least a match in one of the languages.

Common Crawl includes a large diversity of domains from the Web.

Comparing Genders across Languages In addition, we find a large variability across languages. Some languages like Belarus (`bel`) and Swedish (`swe`) are highly skewed toward the Masculine gender class, while other languages are much more balanced such as Mandarin Chinese (`cmn`) or Hindi (`hin`).

We note that gender distribution cannot be compared across languages quantitatively. Indeed, first, our lexicon is based by design on nouns that are not entirely parallel across languages. Second, our metric highly depends on the number of words in each dataset, which is not comparable across all languages due to their differences in morphology and syntax. However, as discussed below (§ 5.2), our pipeline allows us to highlight qualitative differences in how gender is encoded in different languages.

5.2 Qualitative Analysis: Gender representation variation in parallel data

To understand the cause of these gender representation differences across languages, we present several examples in Table 2. We dicuss them here:

- Omission of words: When comparing English with Romance languages, we observe cases where the gendered word is omitted in English while being translated as a masculine noun in the target language, like Spanish or Catalan. This leads to larger gender representation gaps in these languages.
- Multiple translations and part-of-speech: Sentence 2 shows the impact of how a single English word corresponds to multiple words in other languages. The unspecified word "kid" is translated in 10 words in Catalan: unspecified "jove, criatura"; feminine "minyona, menuda, nena, marreca"; masculine, "minyó, menut, nen, marrec", augmenting the coverage in that second language. In addition, some words in Catalan have multiple part-of-speech, like "jove, menuda, menut" which can act as nouns or adjectives.
- Sentence 3 illustrates that even with typologically different languages such as English and Mandarin Chinese, our lexical matching approach successfully highlights cases where gender is preserved across languages.
- Finally, in Sentence 4, we illustrate the limit of the context-free approach. Indeed, the noun “ba” means both *father* and *three* in Vietnamese, leading to over-estimating the masculine class on some samples.

In summary, the differences in gender representation across languages point to four distinct phenomena: First, the inherent limit of our context-free lexical approach. Gender is, in some cases, incorrectly estimated by a by-design restricted lexical-matching method (e.g., Sentence 4). Second, different domain distributions may lead to diverse gender representation. As reported in the previous section, for some languages, the gender scores highly vary depending on the domains (e.g., News vs. Web crawled data). This suggests that when we analyze non-parallel data, the domain may be a prevalent factor that explains gender representation differences across languages. Third, as we observe when analyzing parallel data, gender representation differences may come from biases in the translation itself. For instance, in Sentence 1, the translation explicitly encoded the masculine gender in Spanish and Catalan while being gender unspecified in English. Other translations could have preserved the gender. Fourth, the way gender is encoded is, partly at least, unique to each language. Some languages are inherently biased toward the masculine gender (e.g. “padres”, which may mean both *fathers* and *parents* in Spanish). Other languages do not always have genderless nouns. For instance, *siblings* can only be translated onto Lithuanian as “broliai ir seserys” *Brothers and Sisters*.

6 Conclusion

In this work, we presented GENDER-GAP, a large scale multilingual pipeline to compute gender distribution across 55 languages. We find that broadly used datasets are biased toward masculine gender. Based on this finding, our primary recommendation for multilingual NLP practitioner is to report the gender distribution along with the performance score. This allows reader and systems adopters to be aware of these biases in order to integrate this in their system deployment. Secondly, based on our multilingual lexicon, many directions could be taken to mitigate biases in the performance of the systems (due to biases in the data). Qian et al. (2022) developed a perturbation-based technique to build NLP systems that are less biased toward specific group. We envision using our multilingual lexicon to adapt this technique beyond English.

Limitations

English-centric We designed the list of gendered nouns starting from the English language and then scaled it to multiple languages. This means that our approach may cover incompletely the nuances in different language families regarding gender or only cover them partially and from an English-centric perspective.

Non-Binary Gender Modeling To favor scalability across 55 languages, we chose to use a three gender class lexicon. However, this restrict our approach to binary genders (masculine and feminine) and we only measure imperfectly non-binary genders distribution (Haynes et al., 2001; Herdt, 2020) with the “unspecified” class. We leave for future work the refinement of our lexical categories in order to measure more granularly genders across languages.

Lexical Matching The core assumption of this work is that our predefined lexicon defined in Section 3.2 gives us a proxy to account for gender distributions in large datasets. Although our lexicon is obviously not exhaustive, it is simple enough to scale to highly multilingual environments. Future work could consider other types of nouns (beyond family relations or persons) such as gendered occupations nouns, pronouns, etc.

References

Lauren Ackerman 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

M.G. Bulmer. 1979. *Principles of Statistics*. Dover Books on Mathematics Series. Dover Publications.

Judith Butler. 2011. *Bodies that matter: On the discursive limits of sex*. Taylor & Francis.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Greville G. Corbett. 2013. Number of genders (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Ander Corral and Xabier Saralegi. 2022. Gender bias mitigation for NMT involving genderless languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 165–176, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marta Costa-jussà, Christine Basta, Oriol Domingo, and André Rubungo. 2022. Occgen: Selection of real-world multilingual parallel data balanced in gender within occupations. In *Advances in Neural Information Processing Systems*, volume 35, pages 1445–1457. Curran Associates, Inc.

Marta R Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. *arXiv preprint arXiv:2305.13198*.

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

Marta Ruiz Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2019. Gebiotoolkit: Automatic extraction of gender-balanced multilingual corpus of wikipedia biographies. In *International Conference on Language Resources and Evaluation*.

Council of Europe. 2023. Sex and gender. `https://www.coe.int/en/web/gender-matters/sex-and-gender`. [Accessed: July 17, 2023].

Brian Duignan. 2023. gender continuum. Encyclopedia Britannica.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

J. R. Foulds, R. Islam, K. Keya, and S. Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, Los Alamitos, CA, USA. IEEE Computer Society.

Felicity Haynes, Tarquam McKenna, and E McWilliam. 2001. *Unseen genders: Beyond the binaries*. Peter Lang Publishing.

M. Hellinger and H. Bussmann. 2001. *Gender Across Languages: The Linguistic Representation of Women and Men*. Number vol. 2 in Gender Across Languages: The Linguistic Representation of Women and Men. J. Benjamins.

Gilbert Herdt. 2020. *Third sex, third gender: Beyond sexual dimorphism in culture and history*. Princeton University Press.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Ruth Kramer. 2020. Grammatical gender: A close look at gender assignment across languages. *Annual Review of Linguistics*, 6:45–66.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

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

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Jennifer L. Prewitt-Freilino, T Andrew Caswell, and Emmi K. Laakso. 2012. The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles*, 66:268–281.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 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)*, pages 99–109, Online. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (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.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *International Conference on Machine Learning*, pages 10882–10893. PMLR.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Lang	Feminine	Masculine	Unspecified	Δ(| Fem.-Masc. |) (ste.)	# words	% matched sentences
Flores DevTest.						
eng	0.121	0.065	**0.379**	0.056 (0.0003)	23211	11.2
arb	0.051	0.047	**0.094**	0.004 (0.0002)	25549	4.1
asm	0.056	**0.102**	0.093	0.046 (0.0003)	21610	4.5
bel	0.161	0.434	**0.444**	0.274 (0.0005)	21174	12.7
ben	0.076	**0.204**	0.142	0.128 (0.0004)	21101	7.2
bul	0.083	**0.258**	0.114	0.175 (0.0004)	22834	9.1
cat	0.115	**0.154**	0.146	0.038 (0.0003)	26005	9.4
ces	0.113	**0.385**	0.153	0.271 (0.0005)	20284	10.6
ckb	0.052	0.119	**0.152**	0.066 (0.0003)	21073	4.3
cmn	0.101	0.042	**0.794**	0.059 (0.0002)	23676	17.6
cym	0.104	0.046	**0.146**	0.058 (0.0002)	26013	6.4
dan	0.129	0.045	**0.160**	0.085 (0.0003)	22471	6.3
deu	0.114	0.059	**0.301**	0.055 (0.0003)	21922	9.2
ell	0.118	**0.261**	0.253	0.143 (0.0004)	24548	12.8
est	0.116	0.099	**0.519**	0.017 (0.0003)	18107	11.0
fin	0.116	0.086	**0.147**	0.031 (0.0004)	16314	4.9
fra	0.082	0.089	**0.234**	0.007 (0.0003)	26910	9.6
gle	0.038	0.053	**0.479**	0.015 (0.0002)	26517	12.3
hin	0.048	0.032	**0.104**	0.016 (0.0002)	25094	3.8
hun	0.040	**0.250**	0.060	0.210 (0.0004)	19977	6.0
ind	0.179	**0.468**	0.193	0.289 (0.0006)	20728	14.5
ita	0.082	0.168	**0.223**	0.086 (0.0003)	25583	10.2
jpn	0.113	0.061	**0.716**	0.052 (0.0002)	31000	20.4
kan	0.086	0.032	**0.102**	0.054 (0.0002)	18593	3.1
kat	**0.097**	0.029	0.068	0.068 (0.0002)	20527	3.0
khk	0.274	**0.874**	0.270	0.599 (0.0007)	21861	22.6
kir	0.134	0.194	**0.482**	0.060 (0.0004)	20120	12.7
kor	0.037	**0.055**	0.012	0.018 (0.0002)	16341	1.7
lit	**0.140**	0.088	0.125	0.052 (0.0003)	19246	5.4
lug	0.084	0.023	**0.606**	0.061 (0.0002)	21457	12.6
mar	**0.060**	0.044	0.055	0.016 (0.0002)	18281	2.5
mlt	**0.661**	0.179	0.191	0.482 (0.0005)	25104	18.3
nld	0.113	0.071	**0.236**	0.042 (0.0003)	21229	7.5
pan	0.105	**0.127**	0.087	0.022 (0.0003)	27651	6.5
pes	0.166	0.116	**0.310**	0.050 (0.0003)	24157	10.0
pol	0.137	0.061	**0.544**	0.076 (0.0003)	21143	13.4
por	0.103	0.078	**0.338**	0.025 (0.0003)	24269	10.9
ron	0.100	0.092	**0.240**	0.008 (0.0003)	25046	8.9
rus	0.117	**0.117**	0.098	0.000 (0.0003)	21431	5.4
slk	0.113	0.054	**0.508**	0.059 (0.0003)	20292	11.5
slv	0.069	0.032	**0.069**	0.037 (0.0002)	21586	3.3
spa	0.104	0.201	**0.260**	0.097 (0.0003)	26896	12.3
swe	0.119	0.176	**0.200**	0.057 (0.0004)	20969	8.9
swh	0.225	0.213	**0.689**	0.013 (0.0004)	23964	20.4
tam	**0.168**	0.101	0.123	0.067 (0.0003)	17862	4.5
tel	0.092	**0.140**	0.122	0.049 (0.0004)	16373	3.9
tgl	0.075	0.041	**0.373**	0.034 (0.0002)	29518	11.1
tha	0.156	0.038	**0.439**	0.118 (0.0003)	28922	12.7
tur	0.287	0.270	**0.293**	0.017 (0.0005)	17775	8.4
urd	0.074	**0.320**	0.234	0.245 (0.0004)	26887	9.2
uzn	0.156	0.076	**0.260**	0.080 (0.0003)	21181	8.3
vie	0.139	0.301	**1.441**	0.162 (0.0004)	25263	30.6
yue	0.093	0.040	**0.837**	0.053 (0.0002)	24728	19.1
zul	0.394	0.059	**0.653**	0.335 (0.0005)	18532	17.0
avg.	0.128	0.144	**0.302**	0.097 (0.0003)	22572	10.1

Table 3: % Gender Distribution in FLORES-200 dataset (NLLB Team et al., 2022). We **bold** the most represented gender class, and underline the second most represented gender class for each language. We report Δ the gender gap defined as the absolute difference between the Feminine and Masculine scores along with the standard error (ste.). % matched sentences refers to the coverage of our pipeline (cf. § 5.1).

Lang	Feminine	Masculine	Unspecified	Δ(| Fem.-Masc. |) (ste.)	# words	% matched sentences
				NTREX		
eng	0.166	0.203	**0.379**	0.037 (0.0003)	48254	15.5
arb	0.105	0.107	**0.206**	0.002 (0.0002)	51388	8.7
bel	0.224	**0.574**	0.397	0.350 (0.0004)	44597	16.9
ben	0.131	0.212	**0.311**	0.081 (0.0003)	40505	11.3
bul	0.122	**0.270**	0.095	0.148 (0.0003)	49283	10.5
cat	0.195	**0.272**	0.235	0.077 (0.0003)	54401	15.6
ces	0.248	**0.454**	0.190	0.206 (0.0004)	43623	16.3
ckb	0.054	0.167	**0.244**	0.113 (0.0002)	42554	6.5
cmn	0.193	0.149	**0.944**	0.044 (0.0003)	50326	24.8
cym	0.086	**0.164**	0.154	0.078 (0.0002)	52540	8.8
dan	0.184	0.177	**0.186**	0.007 (0.0003)	45684	10.7
deu	0.162	0.192	**0.276**	0.030 (0.0003)	46398	12.3
ell	0.141	**0.344**	0.170	0.203 (0.0003)	51204	14.4
est	0.212	0.328	**0.458**	0.116 (0.0004)	37794	15.8
fin	0.158	0.181	**0.196**	0.024 (0.0003)	33617	7.9
fra	0.140	0.208	**0.258**	0.068 (0.0003)	54336	13.9
gle	0.081	0.135	**0.493**	0.054 (0.0002)	54205	16.2
hin	0.103	0.092	**0.147**	0.011 (0.0002)	55207	8.1
hun	0.110	**0.140**	0.072	0.030 (0.0002)	42834	6.6
ind	0.195	**0.581**	0.213	0.386 (0.0004)	45071	18.1
ita	0.166	**0.301**	0.229	0.135 (0.0003)	51884	14.8
jpn	0.209	0.201	**0.868**	0.008 (0.0002)	59704	25.2
kan	0.115	0.101	**0.131**	0.014 (0.0002)	36574	4.9
kat	**0.198**	0.140	0.103	0.058 (0.0002)	39912	5.1
kir	0.209	0.181	**0.310**	0.028 (0.0003)	38682	12.0
kor	0.040	**0.062**	0.059	0.022 (0.0002)	32204	2.5
lit	0.187	**0.216**	0.153	0.029 (0.0003)	41190	9.4
mar	**0.089**	0.056	0.069	0.033 (0.0002)	35980	3.6
mlt	**0.795**	0.212	0.284	0.583 (0.0004)	51466	24.7
nld	0.190	0.194	**0.196**	0.004 (0.0003)	48003	11.2
pan	0.150	**0.176**	0.100	0.026 (0.0002)	53845	9.9
pol	0.242	0.211	**0.525**	0.030 (0.0003)	42638	17.9
por	0.160	0.228	**0.244**	0.067 (0.0003)	50482	13.8
ron	0.152	0.191	**0.367**	0.039 (0.0002)	54463	15.5
rus	0.171	**0.210**	0.089	0.039 (0.0003)	46295	8.5
slk	0.248	0.216	**0.420**	0.033 (0.0003)	43063	16.0
slv	**0.093**	0.084	0.077	0.009 (0.0002)	45339	4.8
spa	0.162	0.297	**0.344**	0.135 (0.0003)	52579	15.9
swe	0.156	**0.265**	0.240	0.109 (0.0003)	42980	12.3
tam	**0.308**	0.273	0.068	0.035 (0.0002)	36960	7.0
tel	0.118	**0.213**	0.086	0.095 (0.0003)	31427	5.0
tha	0.418	0.128	**0.870**	0.290 (0.0003)	57923	23.1
tur	0.227	0.183	**0.252**	0.044 (0.0003)	36163	8.1
vie	0.146	0.633	**2.166**	0.487 (0.0004)	52577	45.7
yue	0.133	0.173	**0.933**	0.041 (0.0002)	54233	26.6
avg.	0.180	0.224	0.329	0.099 (0.0003)	46231	13.4

Table 4: % Gender Distribution in NTREX data (Federmann et al., 2022). We **bold** the most represented gender class, and underline the second most represented gender class for each language. We report Δ the gender gap defined as the absolute difference between the Feminine and Masculine scores along with the standard error (ste.). % matched sentences refers to the coverage of our pipeline (cf. § 5.1).

Lang	Feminine	Masculine	Unspecified	Δ(| Fem.-Masc. |) (ste.)	# words	% matched documents
CommonCrawl						
eng	0.120	0.115	**0.243**	0.005 (0.0000)	2529756	9.4
arb	0.101	**0.106**	0.085	0.005 (0.0000)	6078083	9.5
bel	0.122	**0.447**	0.358	0.325 (0.0000)	2430561	14.1
ben	0.158	**0.199**	0.140	0.041 (0.0000)	4603054	14.4
bul	0.072	**0.145**	0.142	0.073 (0.0000)	2708232	7.7
cat	0.079	0.141	**0.152**	0.062 (0.0000)	3157729	9.1
ces	0.117	0.146	**0.165**	0.030 (0.0000)	2366804	7.9
ckb	0.108	0.049	**0.124**	0.059 (0.0000)	5341945	10.2
cmn	0.170	0.097	**0.519**	0.072 (0.0000)	5484451	23.8
cym	0.079	0.082	**0.164**	0.003 (0.0000)	2777579	7.4
dan	0.182	0.102	**0.201**	0.080 (0.0000)	2310993	7.9
deu	0.144	0.099	**0.187**	0.044 (0.0000)	2148705	6.8
ell	0.068	**0.143**	0.142	0.075 (0.0000)	2855903	7.7
est	0.112	0.152	**0.429**	0.040 (0.0000)	1943773	10.3
fin	**0.294**	0.201	0.155	0.094 (0.0001)	1621020	7.3
fra	0.110	0.136	**0.151**	0.025 (0.0000)	2857434	8.3
gle	0.044	0.101	**0.406**	0.057 (0.0000)	2634719	12.2
hin	**0.176**	0.124	0.065	0.052 (0.0000)	2675603	7.4
hun	0.058	**0.097**	0.075	0.038 (0.0000)	2572506	4.5
ind	0.183	**0.367**	0.184	0.185 (0.0000)	2227691	12.1
ita	0.131	**0.195**	0.070	0.064 (0.0000)	2961219	8.2
jpn	0.858	0.724	**0.963**	0.134 (0.0000)	5964414	27.4
kan	**0.103**	0.094	0.093	0.009 (0.0000)	3772755	7.0
kat	**0.129**	0.089	0.116	0.040 (0.0000)	3977699	6.2
khk	0.301	**0.948**	0.248	0.647 (0.0000)	4996882	32.1
kir	0.269	**0.308**	0.270	0.039 (0.0000)	3895597	20.0
kor	0.032	0.047	**0.047**	0.015 (0.0000)	2364450	2.4
lit	0.148	0.117	**0.243**	0.031 (0.0000)	2293338	8.4
mar	**0.133**	0.112	0.051	0.021 (0.0000)	1531197	3.8
mlt	**0.554**	0.179	0.213	0.375 (0.0001)	2437212	20.0
nld	0.127	0.101	**0.201**	0.027 (0.0000)	1921934	6.3
pan	0.236	**0.308**	0.074	0.072 (0.0000)	6772503	22.4
pes	1.459	1.425	**1.514**	0.034 (0.0000)	3881584	14.7
pol	0.175	0.074	**0.290**	0.101 (0.0000)	2453053	9.9
por	0.110	**0.160**	0.158	0.050 (0.0000)	2846706	9.2
ron	0.207	0.138	**0.257**	0.068 (0.0000)	2555624	10.1
rus	0.107	**0.139**	0.117	0.031 (0.0000)	2565203	6.4
slk	0.111	0.066	**0.324**	0.045 (0.0000)	2269033	8.7
slv	0.057	0.071	**0.142**	0.014 (0.0000)	2373967	5.3
spa	0.122	**0.255**	0.183	0.133 (0.0000)	3046193	11.7
swe	0.179	**0.372**	0.157	0.193 (0.0000)	2346273	11.5
swh	0.221	0.194	**0.492**	0.027 (0.0000)	2385794	19.9
tam	**0.766**	0.676	0.073	0.091 (0.0000)	1691612	11.7
tel	0.127	**0.165**	0.056	0.038 (0.0000)	1277513	3.5
tgl	0.145	0.108	**0.419**	0.038 (0.0000)	5035687	21.2
tha	0.735	0.107	**0.932**	0.628 (0.0000)	7142646	28.9
tur	**0.228**	0.202	0.215	0.027 (0.0000)	2293026	7.3
uzn	0.119	0.077	**0.280**	0.042 (0.0000)	2973725	9.5
avg.	0.212	0.260	0.251	0.136 (0.0003)	3088848	12.0

Table 5: % Gender Distribution in a Common Crawl sample. We **bold** the most represented gender class, and underline the second most represented gender class. We report Δ the gender gap defined as the absolute difference between the Feminine and Masculine scores along with the standard error (ste.). % matched documents refers to the coverage of our pipeline (cf. § 5.1).

Language Code	**Language**
arb_Arab	Modern Standard Arabic
asm_Beng	Assamese
bel_Cyrl	Belarusian
ben_Beng	Bengali
bul_Cyrl	Bulgarian
cat_Latn	Catalan
ces_Latn	Czech
ckb_Arab	Central Kurdish
cmn_Hans	Mandarin Chinese (simplified script)
cym_Latn	Welsh
dan_Latn	Danish
deu_Latn	German
ell_Grek	Greek
eng_Latn	English
est_Latn	Estonian
fin_Latn	Finnish
fra_Latn	French
gle_Latn	Irish
hin_Deva	Hindi
hun_Latn	Hungarian
ind_Latn	Indonesian
ita_Latn	Italian
jpn_Jpan	Japanese
kat_Geor	Georgian
khk_Cyrl	Halh Mongolian
kir_Cyrl	Kyrgyz
lit_Latn	Lithuanian
lug_Latn	Ganda
lvs_Latn	Standard Latvian
mar_Deva	Marathi
mlt_Latn	Maltese
nld_Latn	Dutch
pan_Guru	Eastern Panjabi
pes_Arab	Western Persian
pol_Latn	Polish
por_Latn	Portuguese
ron_Latn	Romanian
rus_Cyrl	Russian
slk_Latn	Slovak
slv_Latn	Slovenian
spa_Latn	Spanish
swe_Latn	Swedish
swh_Latn	Swahili
tam_Taml	Tamil
tha_Thai	Thai
tur_Latn	Turkish
ukr_Cyrl	Ukrainian
urd_Arab	Urdu
uzn_Latn	Northern Uzbek
vie_Latn	Vietnamese
yue_Hant	Yue Chinese (traditional script)
kan_Knda	Kannada
tel_Telu	Telugu
tgl_Latn	Tagalog
zul_Latn	Zulu

Table 6: The 55 languages analyzed in this work, subselected from the 200 NLLB languages (NLLB Team et al., 2022).