Gender Bias Evaluation in Machine Translation for Amharic, Tigrinya, and Afaan Oromoo

Walelign Tewabe Sewunetie[1,2], Atnafu Lambebo Tonja[3,4,5], Tadesse Destaw Belay[4], Hellina Hailu Nigatu[6], Gashaw Kidanu[7], Zewdie Mossie[1], Hussien Seid[7], Seid Muhie Yimam[8]

[∀]Ethio NLP, [1]Debre Markos University, Ethiopia, [2]University of Miskolc, Hungary, [3]Lelapa AI, [4]Instituto Politécnico Nacional, Mexico, [5]Mohamed bin Zayed University of Artificial Intelligence, UAE, [6]University of California, Berkeley, USA, [7]Addis Ababa Science and Technology University, Ethiopia [8]Universität Hamburg, Germany

Abstract

While Machine Translation (MT) research has progressed over the years, translation systems still suffer from biases, including gender bias. While an active line of research studies the existence and mitigation strategies of gender bias in machine translation systems, there is limited research exploring this phenomenon for low-resource languages. The limited availability of linguistic and computational resources confounded with the lack of benchmark datasets makes studying bias for low-resourced languages that much more difficult. In this paper, we construct benchmark datasets to evaluate gender bias in machine translation for three low-resource languages: Afaan Oromoo (Orm), Amharic (Amh), and Tigrinya (Tir). Building on prior work, we collected 2400 gender-balanced sentences parallelly translated into the three languages. From human evaluations of the dataset we collected, we found that about 93% of Afaan Oromoo, 80% of Tigrinya, and 72% of Amharic sentences exhibited gender bias. In addition to providing benchmarks for improving gender bias mitigation research in the three languages, we hope the careful documentation of our work will help other low-resourced language researchers extend our approach to their languages.[1]

[1]Our dataset is available at https://huggingface.co/datasets/EthioNLP/Gender-Bias-Evaluation-Dataset

1 Introduction

Machine Translation (MT) systems play a pivotal role in breaking down language barriers and facilitating cross-cultural communication. Gender bias poses a significant challenge, particularly in languages with limited linguistic resources. The imbalance within datasets used for MT training often results in gender-related disparities. In low-resource languages like Amharic, Tigrinya, and Afaan Oromoo, and in morphologically rich languages like Arabic (Habash et al., 2019; Alhafni et al., 2022) professional names such as doctor, pilot, professor, etc., are mostly translated using the masculine gender.

Machine Translation services often default to masculine forms for professions like "doctor" and "nurse," for feminine forms potentially reflecting and reinforcing gender stereotypes. Figure 1 and Figure 2 demonstrate this for the Amharic language[2]. These types of bias can lead to misunderstandings and reinforce gender roles, influencing how people perceive different professions based on gender. Understanding and addressing gender bias in MT systems is vital for ensuring equitable and accurate communication across diverse linguistic communities.

Addressing the issue of gender bias in MT systems requires adequate datasets for evaluation; a challenging task in the context of low-resource languages. This work contributes to building equitable MT systems for low-resource languages by constructing a gold-test dataset for three languages: Amharic,

[2]In the screenshots provided, Google Translate transliterated the word "doctor" instead of translating it to the Amharic word for 'doctor' ሐኪም

Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies, pages 1–11
June 27, 2024 ©2024 European Association for Machine Translation

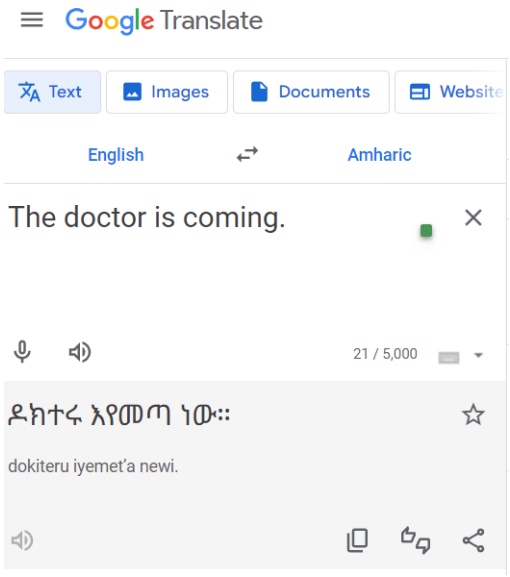

Figure 1: Translating the sentence "The doctor is coming" Google Translate translates the word "doctor" into masculine gender for the Amharic language. The word "doctor," translated in Amharic as "**ዶክተር**" (dokter), is gender-neutral. However, when translating "The doctor is coming," Google Translate translates the sentence to "**ዶክተሩ እየመጣ ነው።**" (dokteru eyemet'a new). Here the phrase "The doctor" becomes "**ዶክተሩ**" (dokteru); the prefix "u" indicates masculine gender in the Amharic language. In addition, the word "coming" translates into "**እየመጣ**" (eyemet'a); which also indicates masculine gender.

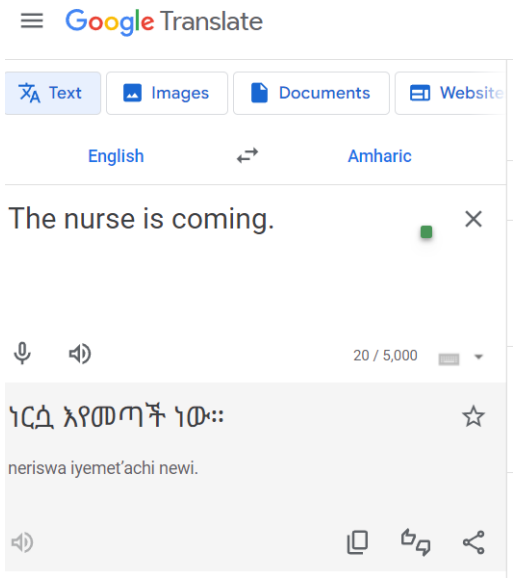

Figure 2: In the sentence "The nurse is coming", the word "nurse," translated in Amharic as "**ነርስ**" (ners), is gender-neutral. However, when translating "The nurse is coming," Google Translate translates the sentence to "**ነርሷ እየመጣች ነው።**" (nerswa eyemet'ach new). Here the phrase "The nurse" becomes "**ነርሷ**" (ners-wa); the prefix "wa" indicates feminine gender in the Amharic language. In addition, the word "coming" translates into "**እየመጣች**" (eyemet'ach); which also indicates feminine gender.

Tigrinya, and Afaan Oromoo. The methodologies developed in this research can subsequently be applied and scaled up to assess gender bias in other low-resource languages. We collected 2400 gender-balanced sentences, which can be used as a benchmark for gender bias evaluation in low-resource language translation.

In addition, this study investigates users' perceptions of gender bias in commercial MT systems and evaluates Google Translate as a use case in the three languages of study. Our analysis shows interesting differences in respondents' perceptions of gender bias across these language communities. These findings underscore the detailed relationship between language, culture, and gender bias perception in MT systems, highlighting the need for adapted approaches to mitigate bias and enhance translation accuracy within specific linguistic contexts. Furthermore, this study investigates the performance of one open-source MT model and one commercial model, namely, NLLB (Team et al., 2022), and Google MT using automatic evaluation metrics, such as SacreBleu (Post, 2018), and Chrf++ (Popović, 2017). The outcomes of this evaluation across various language pairs shed light on the efficacy and accuracy of MT systems in translating between English and the target languages. The evaluation shows diverse performance metrics across language pairs, with distinct variations in translation quality and effectiveness. These results underscore the importance of robust evaluation methodologies and metrics in assessing MT system performance and informing strategies for enhancing translation accuracy and efficiency across diverse linguistic contexts.

2 Related work

Investigating bias in MT systems is an active body of work in the NLP community. We use the taxonomy from (Blodgett et al., 2020a) and focus on representational harms due to stereotyping: sustaining stereotypical gender connotations for occupations during translation, thereby limiting the variety

of occupations a specific gender may or may not engage in[3]. Previous works in this space have relied on (1) curating benchmark datasets (e.g.(Wairagala et al., 2022; Cho et al., 2019)), (2) human evaluation schemes (e.g. (Stanovsky et al., 2019)), and (3) automatic evaluation schemes(e.g. (Savoldi et al., 2021)). In curating benchmark datasets, (Prates et al., 2020) prepared a gender-balanced dataset for evaluating gender bias in translation systems pertaining to occupation. Since different languages represent gender in various ways (Savoldi et al., 2021), evaluation and mitigation strategies might also have to account for such variation. For instance, (Cho et al., 2019) prepared test sets with gender natural pronouns used in the Korean language for investigating bias in Korean-English translation pairs.

In evaluating gender bias in MT, several works rely on automatic metrics. (Prates et al., 2020) found that Google Translate defaults to the masculine pronoun when translating job descriptions, particularly in relation to science, technology, engineering, and mathematics (STEM) professions. (Cho et al., 2019) introduces a new evaluation index, the Translation Gender Bias Index (TGBI), for measuring gender neutrality and evaluating Korean-English translation pairs. (Stanovsky et al., 2019) introduce an evaluation protocol that relies on co-reference resolution datasets and morphological analysis to automatically evaluate gender bias across eight target languages that use grammatical gender. (Wairagala et al., 2022) used the Word Embeddings Fairness Evaluation Framework (WEFE) to measure gender bias in MT systems built for Luganda-English translation. While automated measures allow us to capture a broader understanding of the phenomenon, they may limit the detail and depth of our analysis. The study by (Stanovsky et al., 2019) uses automatic and human evaluations in tandem, exploiting both the versatility of automated evaluation and the nuance and detail captured by human evaluation.

As the work by (Blodgett et al., 2020b) argues, it is important first to articulate how bias in such systems can be harmful. Relying on the taxonomy of harms from prior work (Crawford, 2017; Barocas et al., 2017), we posit that understanding gender bias exhibited by MT systems would allow us to (1) uncover the representational harms the systems exhibit thereby understanding what power structures they uphold and (2) mitigate allocational harms that might result from deploying such systems in downstream applications (e.g. employment and job search).

One challenge in studying bias in machine-translated text is the diverse socio-cultural aspects that shape how gender is articulated among different groups and how stereotypes propagate in this diverse context. Talat et al. (2022) have shed light on the difficulty of studying and mitigating bias across multicultural, multilingual groups. Such contexts require community-rooted efforts that thoroughly investigate how the culture and language are structured. In this work, we curate benchmark datasets for three low-resource languages through collaborations among native speakers. Based on previous works, (Renduchintala et al., 2021; Stanovsky et al., 2019), we conduct an automatic evaluation of the translation quality overall and human evaluations of gender bias in popular machine translation systems to understand the current landscape of translation systems for these languages.

3 Background: Linguistic Gender Representation

Amharic, a Semitic language, uses grammatical gender. Most nouns and pronouns have distinct masculine and feminine forms. Gender-specific pronouns are used (e.g., **እሱ** (əssu) for "he" and **እሷ** (əsswa) for "she"), and job titles can also have gendered forms.

Like Amharic, Tigrinya, another Semitic language, has grammatical gender. Gender distinctions are marked in nouns and pronouns. There are specific pronouns for different genders (e.g., **ንሱ** (nəssu) for "he" and **ንሳ** (nəssa) for "she"), and job titles may vary depending on gender.

Afaan Oromoo, a Cushitic language, does not have grammatical gender in the same way as the other two languages. Gender-neutral pronouns are often used, but context can some-

[3]We note in this work, we are considering a binary gender system of men and women

times specify gender. Gender is less likely to be marked in job titles compared to the other two languages of study.

To illustrate more about the issues in translating a sentence and a professional word, we can see the following example of Gender Bias in English to Amharic Translation.

The English to Amharic Google Translate (accessed January 20, 2024) output of the sentence "The nurse helped the doctor" is "**ነርሷ ሐኪሙን ረድታለች።**" (nerswa hakimun redtalech). Here, "**ነርሷ**" (nerswa) 'the nurse' is female, and "**ሐኪሙን** " (hakimun) "the doctor" is male. The word "helped", while it has a translation issue[4], is translated to "**ረድታለች**" (redtalech), which is indicative of a feminine subject.

In Amharic, the source sentence "The nurse helped the doctor" can be translated in eight different ways as follows:

1. "**ነርሷ ሐኪሙን ረድታዋለች።**" (nerswa hakimun redtawalech). Here "**ነርሷ**" (neriswa) 'the nurse' is female, "**ሐኪሙን**" (hakimun) 'the doctor' is male, and "**ረድታዋለች**" (reditawalechi) 'she helped him'.

2. "**ነርሷ ሐኪሟን ረድታታለች።**" (neriswa ḥākīmwan reditatalechi). Here "**ነርሷ**" (neriswa) 'the nurse' is female, "**ሐኪሟን**" (ḥākīmwan) 'the doctor' is female, and "**ረድታታለች**" (reditatalechi) 'she helped her'.

3. "**ነርሷ ሐኪሙን ረድታቸዋለች።**" (neriswa ḥākīmun reditachewalechi) (for respect or plural). Here "**ነርሷ**" (neriswa) 'the nurse' is female, "**ሐኪሙን**" (hakimun) 'the doctor' is male, and "**ረድታቸዋለች**" (reditachewalechi) 'she helped him.'

4. "**ነርሷ ሐኪሟን ረድታቸዋለች።**" (neriswa ḥākīmwan reditachewalechi). Here "**ነርሷ**" (neriswa) 'the nurse' is female, "**ሐኪሟን**" (ḥākīmwan) 'the doctor' is female, and "**ረድታቸዋለች**" (reditachewalechi) 'she helped her' (for respect or plural).

5. "**ነርሱ ሐኪሙን ረድቶታል።**" (nersu ḥākīmun reditotale). Here "**ነርሱ**" (nersu) 'the nurse' is male, "**ሐኪሙን**" (hakimun) 'the doctor' is male, and "**ረድቶታል**" (reditotale) 'he helped him'.

6. "**ነርሱ ሐኪሟን ረድቷታል።**" (nersu ḥākīmwan reditwatale). Here "**ነርሱ**" (nersu) 'the nurse' is male, "**ሐኪሟን**" (hakimun) 'the doctor' is female, and "**ረድቷታል**" (reditwatale) 'he helped her'.

7. "**ነርሱ ሐኪሟን ረድቷቸዋል።**" (nersu ḥākīmwan reditwachewal). Here "**ነርሱ**" (nersu) 'the nurse' is male, "**ሐኪሟን**" (ḥākīmwan) 'the doctor' is female, and "**ረድቷቸዋል**" (reditwachewal) 'he helped her' (for respect or plural).

8. "**ነርሱ ሐኪሙን ረድቷቸዋል።**" (nersu ḥākīmun reditwachewal). Here "**ነርሱ**" (nersu) 'the nurse' is male, "**ሐኪሙን**" (ḥākīmun) 'the doctor' is female, and "**ረድቷቸዋል**" (reditwachewal) 'he helped him' (for respect or plural).

This range of translations reflects the potential for gender bias in translation when assumptions are made about the gender of individuals based on their professional names.

4 Gold Gender Bias Test Dataset Preparation

4.1 Dataset Collection and Composition

The gold gender bias test dataset was crafted by combining sentences from public repositories (Sharma et al., 2022), with a thorough examination of gender biases across these selected target languages. We first collected an English-centric dataset from a variety of publicly available sources such as SimpleGEN, [5] and winomt,[6] focusing on relevance and diversity. To maintain balance, for every gender-specific sentence, we ensured there was an equivalent counterpart. For example, if a sentence says, "He is a doctor," a corresponding sentence like "She is a doctor" is included for gender parity.

[4]It should be translated in this context as "**ረድታዋለች**" (redtawalech) or "**ረድታቸዋለች**" (redtachewalech) for respect (she helped him) instead of "**ረድታለች**" (redtalech).

[5]SimpleGEN:https://github.com/arendu-zz/SimpleGEN

[6]winomt: https://github.com/manandey/bias_machine_translation/tree/main/data/base/winomt

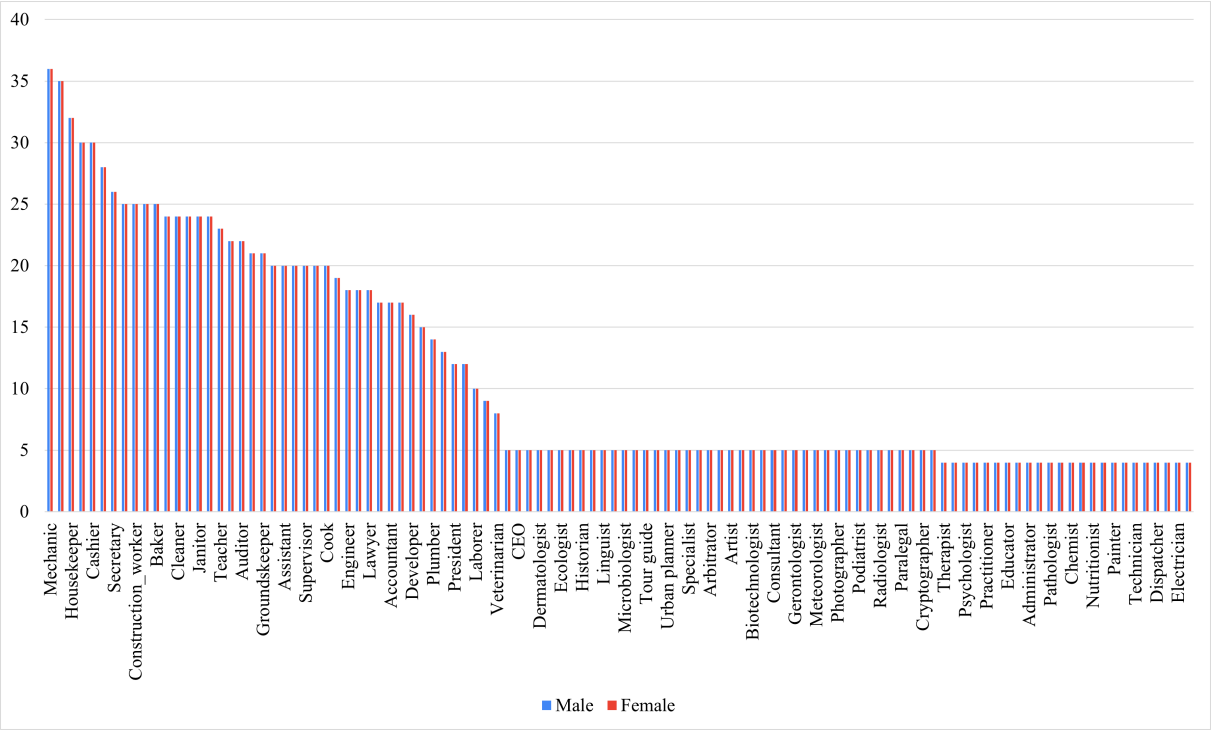

Figure 3: An example from our test dataset used out of 108 professional names.

However, these open-source datasets do not contain all professional names relevant to our communities of interest, even though they contain enough test datasets. For this reason, we used a crowdsourcing approach to collect additional data that reflects various professions. For this approach, we first incorporated the major professional names from (1) the Ethiopian Civil Service Commission list of job titles and (2) querying GPT 3.5 for recent technological professional words. Through this process, we collected 108 unique professional names. Figure 3 demonstrates a sample gender-balanced dataset of each professional name.

Then, we used paid freelancers for crowdsourcing and prepared a Google form containing clear and short instructions about the task. The goal of the crowdsourcing task was to create gender-balanced translation pairs from English-centric data from various sources. One of the key considerations was to include both pronouns and occupations in the dataset. This ensured that each profession is associated with different pronouns, such as "he," "his," and "him" for the masculine, and "she" and "her" for the feminine gender. For this task, ten freelancers were involved and signed an incentive agreement first. Then, we collected the English dataset from SimpleGEN (n=130), winomt (n=192), and crowd-sourced (n=2078), a total of 2400 sentences.

4.2 Dataset Translation

The next task is to translate this collected dataset into three Ethiopian languages: Amharic, Afaan Oromoo, and Tigrinya. Likewise, we have used paid linguistic experts who were proficient in one of our target languages, then undertook the translation process to preserve linguistic accuracy and capture cultural differences specific to each target language.

To prevent boredom and errors, we engaged six language experts and fluent speakers per language pair, totaling eighteen individuals from various universities. We assigned 600 sentence pairs per individual to keep the task manageable. After the translation, we recruited two paid professional linguists and editors for each language pair for quality checking.

The dataset used in this research, referred to as the Gold Gender Bias Test Dataset (GG-BTD), comprises 2400 sentence pairs for each language pair, specifically English-Amharic, English-Afaan Oromoo, and English-Tigrinya, resulting in a total of 7200 sentence pairs.

Within each language pair, the dataset maintains a comprehensive gender balance. Specifically, for each language pair, 1200 sentences represent masculine gender expressions, while the remaining 1200 sentences capture feminine gender expressions.

5 Evaluation Techniques

5.1 Automatic Evaluation

Different evaluation metrics are usually employed to automatically evaluate MT systems. These metrics are often based on word overlap and/or context similarity between references and model outputs. In our work, we employ both types of metrics to evaluate the quality of NLLB and Google MT that we consider in our study. Namely, we used SacreBleu (Post, 2018) and Chrf++ (Popović, 2017) machine translation evaluation metrics. We chose these MT evaluation metrics for several reasons. Firstly, these metrics are widely recognized and utilized in the field of MT research, ensuring compatibility and comparability with existing literature (Kadaoui et al., 2023).

Additionally, SacreBleu and Chrf++ are known for their robustness and effectiveness (Puduppully et al., 2023) in assessing translation quality across different languages and translation systems. Their ability to capture detailed aspects of translation quality, such as fluency, adequacy, and fidelity to the source text, makes them suitable choices for our evaluation framework. Furthermore, both metrics are supported by well-established methodologies and have demonstrated consistent performance in benchmarking studies, giving us confidence in their reliability. However, these metrics evaluate only the overall translation accuracy.

5.2 Human Evaluation

In this work, we relied solely on human-level evaluation techniques for evaluating gender bias. We assessed the gender bias in two MT systems: (1) open source NLLB model and (2) commercially available Google Translate. We chose these models since they support all three languages (Amharic, Tigrinya, Afaan Oromoo).

Given the high cost of human-level evaluation, we only evaluated the gender bias of Google Translate. For the human-level evaluation, first, we developed the evaluation guidelines shown in the appendix 10.1, and used the Potato annotation tool (Pei et al., 2022). Figure 4 shows the Potato annotation tool GUI for human-label evaluation, which supports all modern browsers and can be accessed both from computers and mobile phones for manual annotation. Criteria included gender biases, translation quality, and the accuracy of professional name translations. For evaluation, eighteen paid linguistic experts per language were selected. To avoid subjectivity, we divided evaluators into three groups and made the evaluation into three phases; this implies each sentence is evaluated three times. This is good for taking the majority vote for result analysis.

After each sentence in each of the three languages is evaluated by three evaluators, the annotation tool decides whether the sentence is biased or not by taking the majority vote of the three evaluators.

6 Result and Analysis

Figure 5 provides a clear comparison of responses across three language categories, allowing for insights into the distribution of responses within each language. It presents the gender bias across various language groups, delineating respondents' perceptions regarding the presence or absence of gender bias within each language category.

The data in Table 1 underscores the disparate perceptions of gender bias among respondents across different linguistic backgrounds. Particularly notable is the significantly higher percentage (92.96%) of Afaan Oromoo respondents who indicated observing gender bias compared to other language groups, with only 7.04% indicating otherwise. Similarly, in the Amharic group, approximately 72.50% of respondents indicated observing gender bias, contrasting with 27.50% who did not. Likewise, in the Tigrinya group, the majority (80.96%) indicated observing gender bias, while 19.04% expressed no bias. These findings reveal distinct patterns regarding whether speakers observe gender bias across language groups, suggesting potential implications for addressing and understanding

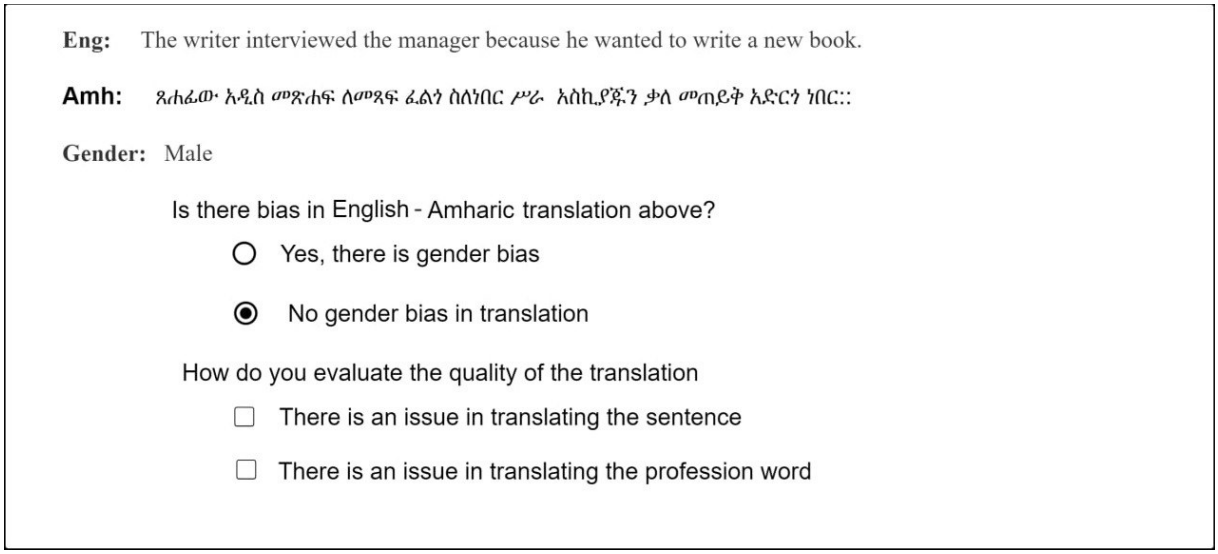

Figure 4: The Potato annotation GUI for the evaluation annotation.

Table 1: Translation Issues by Language

	Amharic	Tigrinya	Afaan Oromoo
There is an issue in translating the sentence	1429	936	918
There is an issue in translating the profession	258	475	612
No issue	510	619	421
Both issues	203	370	449
Total	2400	2400	2400

gender bias in MT within these communities.

Table 1 outlines translation issues across languages, categorized into "Translating the sentence issue" and "Professional word translation issue." Amharic records the highest instances of sentence translation issues at 1429, followed by Tigrinya with 936, and Afaan Oromoo with 918. Regarding professional word translation, Afaan Oromoo leads with 612 instances, trailed by Tigrinya at 475, and Amharic at 258. Tigrinya exhibits the fewest reported issues overall, with 619 sentences indicating no translation issues, compared to 510 for Amharic and 421 for Afaan Oromoo. Conversely, Amharic shows the highest incidence of respondents facing both types of issues at 203, followed by Afaan Oromoo at 449, and Tigrinya at 370. This data underscores the diverse challenges faced in translation across languages and provides valuable insights for enhancing translation quality and addressing language-specific obstacles.

Table 2 presents the evaluation results for NLLB and Google Translate models in the selected language pairs. The table is divided into rows representing different language pairs and columns representing the specific evaluation metrics. Each language pair is evaluated in both translation directions (e.g., Eng-Amh and Amh-Eng), providing insights into machine translation systems' translation quality and performance across various linguistic contexts.

The result shows that the Google MT system outperformed the NLLB model when using English as the source language in both evaluation metrics. This shows that translating English sentences into the target Ethiopian language is challenging for the model. On the other hand, the Google MT system showed better results compared to the NLLB model when translating English sentences into target Ethiopian languages. We observed better performance results when using English as the target language than when using it as the source language in the NLLB model. From this, we can see that for low-resource languages, publicly available MT models like NLLB are strug-

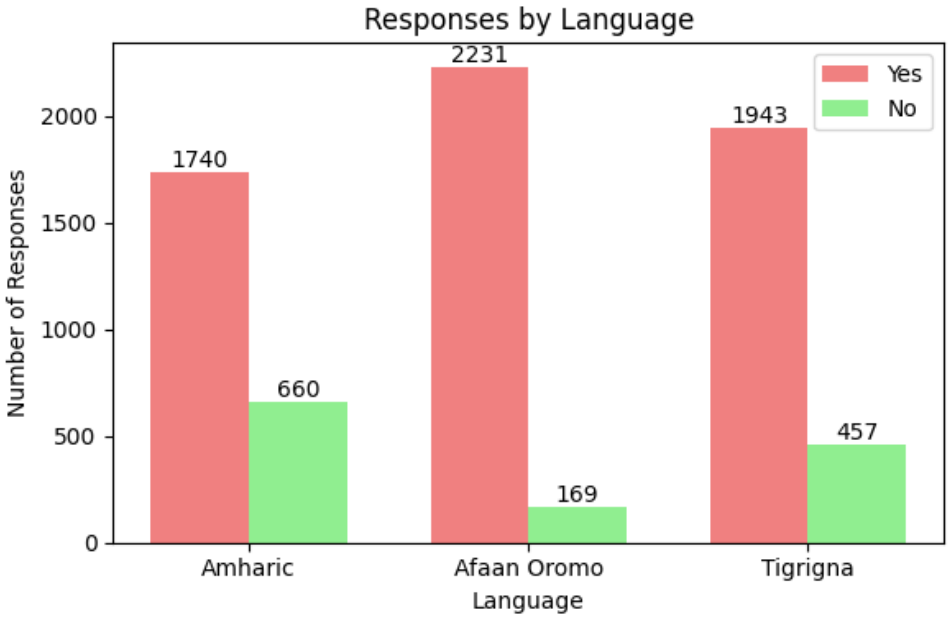

Figure 5: Illustration of the Google Translation Gender Bias test dataset human evaluation result. "Yes" and "No" are the answers to the question, "Is there bias in the translation?". "Yes" means the sentence contains gender bias when translated to a specified language. "No" is no gender bias observed in the translated sentence; the sentence is correctly translated.

Table 2: Automatic Evaluation Results

Language	NLLB		Google MT	
	SacreBleu ↑	Chrf++ ↑	SacreBleu ↑	Chrf++ ↑
Eng- Amh	3.48	23.73	16.13	47.97
Amh- Eng	21.87	50.76	-	
Eng- Orm	4.85	34.85	22.96	56.71
Orm- Eng	17.80	41.63	-	
Eng- Tir	3.89	18.52	16.00	38.00
Tir- Eng	20.01	43.91	-	

gling to predict the correct translation when using English as the source language.

7 Conclusion and Future Work

In this paper, we curated a benchmark dataset for evaluating gender bias in machine translation systems in three low-resource languages. With this test dataset, we conducted a human-level gender bias evaluation of Google Translate and NLLB MT models for the given language pairs. The evaluation result shows that 92.96% of Eng-Orm, 80.96% of Eng-Tir, and 72.50% of Eng-Amh language pairs translations have a gender bias. In addition, we used the automatic evaluation to measure the translation quality of the currently available translation tools that support Amharic, Tigrinya, and Afaan Oromoo languages.

Our findings highlight the need for further research and development efforts to mitigate gender bias and promote gender-inclusive language translation. We observed that this work can be scaled up and used as a benchmark for other low-resource languages. In future work, we will use automatic gender bias evaluation metrics in addition to human evaluation. In addition, we will prepare a gender-balanced dataset for the given language, and we will fine-tune the currently available MT tools.

8 Limitations

The cost and time constraints limit our work to only three language pairs. The sources of gender biases in NLP are different such as the

nature of the language gender, unbalanced professional names in the dataset, and gender unbalanced pronouns in the dataset. This work only focuses on unbalanced professional names.

9 Acknowledgments

We thank GRAIN for funding this work. We also thank the linguistic experts who participated in the annotation and translation of the gender bias test dataset. We thank Hailay Teklehaymanot for his feedback on our manuscript. Finally, we thank the reviewers for their extremely helpful remarks and feedback.

References

Alhafni, Bashar, Nizar Habash, and Houda Bouamor. 2022. The arabic parallel gender corpus 2.0: Extensions and analyses. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In 9th Annual conference of the special interest group for computing, information and society.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020a. Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050.

Blodgett, Su Lin, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020b. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. ACL Anthology, pages 5454–5476, July.

Cho, Won Ik, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. arXiv preprint arXiv:1905.11684.

Crawford, Kate. 2017. The trouble with bias.

Habash, Nizar, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 155–165.

Kadaoui, Karima, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. arXiv preprint arXiv:2308.03051.

Pei, Jiaxin, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In Che, Wanxiang and Ekaterina Shutova, editors, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 327–337, Abu Dhabi, UAE, December. Association for Computational Linguistics.

Popović, Maja. 2017. chrf++: words helping character n-grams. In Proceedings of the second conference on machine translation, pages 612–618.

Post, Matt. 2018. A call for clarity in reporting bleu scores. arXiv preprint arXiv:1804.08771.

Prates, Marcelo OR, 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.

Puduppully, Ratish, Raj Dabre, Ai Ti Aw, and Nancy F Chen. 2023. Decomposed prompting for machine translation between related languages using large language models. arXiv preprint arXiv:2305.13085.

Renduchintala, Adithya, Denise Díaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during Speed-Quality optimization in Neural Machine Translation. ACL Anthology, pages 99–109, August.

Savoldi, Beatrice, 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.

Sharma, Shanya, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1968–1984, Abu Dhabi, United Arab Emirates, Dec. Association for Computational Linguistics.

Stanovsky, Gabriel, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. arXiv preprint arXiv:1906.00591.

Talat, Zeerak, Aurelie Neveol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. ACL Anthology, pages 26–41, May.

Team, NLLB, 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, Loïc Barrault, Gabriel Mejía González, 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. arXiv preprint arXiv:2207.04672.

Wairagala, Eric Peter, Jonathan Mukiibi, Jeremy Francis Tusubira, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, and Ivan Ssenkungu. 2022. Gender bias evaluation in luganda-english machine translation. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 274–286.

10 Appendix

10.1 Appendix: Human-level Evaluation Guideline

Hello everyone,

We are excited to invite you to participate in an important evaluation task aimed at assessing gender bias in Google Translation from English into Amharic, Afaan Oromoo, and Tigrinya. As well as, to evaluate the quality of the overall translation, you are asked to evaluate the translation issue of the whole sentence and whether there is an issue with professional name translation only. As an evaluator, your valuable insights will help us ensure that translations accurately reflect gender inclusivity and professionalism. By carefully reviewing each sentence pair and considering both gender specification and professional terminology, you will play a pivotal role in enhancing translation quality. Your diligent efforts in evaluating 400 sentences will contribute to creating more inclusive and accurate translations. Thank you for your time and cooperation in this endeavor. Let's work together to promote fairness and accuracy in translation.

Evaluation Task: Gender Bias in Google Translation from English into Amharic, Afaan Oromoo, and Tigrinya

1. Login Credentials: Use the provided username and password to access the evaluation platform.

2. Accessing the Task: Open the designated link on your preferred device, whether mobile or computer.

3. Evaluation Procedure:

   - Reviewing Sentences: Carefully examine each provided sentence in English alongside its translation into Amharic, Afaan Oromoo, or Tigrinya.
   - Identifying Gender Bias: Determine the presence of gender bias by considering two factors:
     – Gender Section: Assess whether the translated gender (feminine or masculine) aligns with the gender specified in the original sentence.
     – Professional Words: Check if professional terms are translated with the same gender as provided in the original sentence.
   - Selecting Response: Choose "Yes, there is gender bias" if bias is detected, or "No, gender bias in translation" if not.
   - Evaluate the quality of translation: Select the first check box "There is an issue in translating the sentence" if there is an issue in overall translation or/and select the second check box "There is an issue in translating the profession word".
   - Moving to Next Sentence: Click the "Next" button after making your assessment to proceed to the next set of sentences.

4. Total Sentences: The evaluation task consists of 400 sentences to be assessed.

5. Completion and Compensation: Upon completing the evaluation of all 400 sentences, compensation will be provided according to the prearranged agreement.

We appreciate your dedication and cooperation in contributing to this evaluation task. Your feedback is crucial for improving translation quality and mitigating gender bias.

10.2 Appendix: List of Pronouns in English, Amharic, Tigrinya, Afaan Oromoo

Table 3: Pronouns in English, Amharic, Tigrinya, and Afaan Oromoo.
Key: M=Masculine, F=Feminine, sg=singular, pl=plural, R=Respect

English	Amharic	Tigrinya	Afaan Oromoo
I	**እኔ** (əne)	**ኣነ** (anä)	ana, na
We	**እኛ** (əññа)	**ንሕና** (nəḥəna)	nu
You (M. sg.)	**አንተ** (antä)	**ንስኻ** (nəssəxa)	
You (F. sg.)	**አንቺ** (anči)	**ንስኺ** (nəssəxi)	
You (sg.)			si
You (R)	**እርስዎ** (ərswo)		
You (F, R)		**ንሰን/ንስኽን** (nsen/nskhn)	
You (M, R)		**ንሶም/ንስኹም** (nsom/nskhum)	
You (pl.)	**እናንተ** (ənnantä)		isin
You (M. pl.)		**ንስኻትኩም** (nəssəxatkum)	
You (F. pl.)		**ንስኻትክን** (nəssəxatkən)	
He	**እሱ** (əssu)	**ንሱ** (nəssu)	isa
She	**እሷ** (əsswa)	**ንሳ** (nəssa)	isii, ishii, isee, ishee
S/he (R)	**እሳቸው** (əssaččäw)		
She (R)		**ንሰን** (nsen)	
He (R)		**ንሶም** (nsom)	
They	**እነሱ** (ənnä̈ssu)		isaan
They (M.)		**ንሳቶም** (nəssatom)	
They (F.)		**ንሳተን** (nəssatän)	