Yeswa-Stories: A Three-Way Parallel Dataset of Female African Figures in Low-Web Data Languages

Bethelhem Mamo
Addis Ababa University / Addis Ababa
bethelhem.y.mamo@gmail.com

Hellina Hailu Nigatu
DAIR Institute / Addis Ababa
hellina@dair-institute.org

Abstract

Language technologies used in everyday settings such as machine translation systems risk perpetuating societal bias. Prior work in creating benchmarks for gender bias in machine translation systems 1) focus primarily on high-resourced language pairs or a low-resourced language paired with a high-resource language, 2) use template based benchmarks that usually focus on occupational biases and stereotypes, and 3) translate high-resource benchmarks which may lack cultural significance to low-resourced languages. In this paper, we introduce *Yeswa-Stories*[1], a three-way parallel dataset comprising 1,300 aligned sentences in Amharic, Afaan Oromo, and Tigrinya. We constructed the dataset in two ways: first, we collected English sentences from Wikipedia articles about notable African women and translated them into the three target languages using human translators. To improve cultural representativeness, we further augment the dataset with locally sourced content reflecting the cultural context where the languages are spoken. Our dataset contributes a new resource for studying gender-inclusive translation in low-resourced settings.[2]

[1]"Yeswa" means "her" in Amharic.

[2]https://github.com/hhnigatu/Yeswa-Stories

1 Introduction

Gender bias in machine translation systems has received increasing attention, particularly in how models misrepresent or underrepresent women across languages (Savoldi et al., 2021; Savoldi et al., 2024).

Several works have introduced measurement and mitigation methods, especially for high-resource languages (Stanovsky et al., 2019; Prates et al., 2020). However, existing resources for gender bias evaluation often rely on template-based constructions or narrowly defined domains. These approaches fail to capture the richness of natural language and the diversity of cultural contexts in which gender is expressed. Further, these resources are limited to a handful of high-resourced languages (Joshi et al., 2020).

Creating resources for low-resourced languages is challenging due to the scarcity of parallel corpora, limited digital presence, and the lack of culturally grounded data that reflects real-world usage (Talat et al., 2022). Further, the term "low-resourced language" encompasses a wide array of languages with varying levels and types of resources (Nigatu et al., 2024). Following the recommendation in Nigatu et al., (2024), we specifically use the term "low-web data languages" in this paper to indicate that the resource that the languages of focus are lacking for the context of this paper is data.

In this paper, we introduce *Yeswa-Stories*, a three-way parallel dataset in Amharic, Afaan Oromo, and Tigrinya. Amharic and Tigrinya are Afro-Semitic languages that use the Ge'ez writing script. Afan Oromo is a Cushitic language that uses the Qubee alphabet and is written with Latin characters. While Tigrinya and Amharic are

grammatically gendered languages–meaning every noun is assigned a gender–Afan Oromo has a neutral pronoun. Further, Amharic has a neutral respectful pronoun while Tigrinya has a gendered respectful pronoun. Our dataset focuses on female-centered narratives and combines globally sourced content with locally relevant materials. By doing so, we provide a resource that supports research on gender-inclusive translation and highlight the importance of cultural diversity in low-resource settings.

The paper is structured as follows: We first review related literature, with a focus on gender bias and existing datasets for low-resource machine translation in Section 2. We then present the steps we took and challenges we faced in creating this dataset in Section 3. Finally, we present an analysis of the dataset and evaluate the performance of two widely used machine translation systems in Section 4. Our work contributes to the broader gap in lack of resources in these three low-web data languages and specifically to the gap in the representation of women in multilingual datasets.

2 Related Work

Gender bias evaluation in machine translation has traditionally been conducted using controlled benchmarks, often constructed with template-based sentences that explicitly encode gender variations. These benchmarks typically focus on occupational stereotypes (e.g., associating certain professions with a specific gender) and pronoun resolution tasks (Stanovsky et al., 2019; Cho et al., 2019). However, recent work has shown that such benchmarks often fail to capture real-world distributions of gender in datasets and may overlook biases already embedded in training data (Nigatu et al., 2025). While effective for isolating specific biases, such approaches are predominantly developed for high-resource languages and are often directly translated or adapted for low-resource settings. This adaptation assumes linguistic and cultural equivalence, which may not hold in practice, leading to evaluations that overlook culturally specific expressions of gender (Talat et al., 2022).

Beyond measurement, the impact of gender bias in machine translation systems is significant. Biased outputs can reinforce harmful stereotypes and contribute to representational harm by systematically underrepresenting or misrepresenting women (Blodgett et al., 2020). These issues extend beyond academic concerns, affecting real-world applications such as hiring, education, and access to information. Moreover, correcting such biases often requires substantial resources, including human post-editing and system retraining, which introduces economic disparities (Savoldi et al., 2024). As a result, already marginalized communities bear a disproportionate burden in addressing these shortcomings, highlighting the importance of developing more inclusive systems.

Prior work on gender bias in low-resource languages has begun to reveal the depth of the problem, though research remains sparse compared to high-resource settings (Wairagala et al., 2022; Ramesh et al., 2023). Most related to our work, Sewunetie et al., (2024) constructed a benchmark of 2,400 gender-balanced sentences for Amharic, Afaan Oromo, and Tigrinya and evaluated Google Translate and the NLLB model against it using both automatic metrics and human annotation. Their human evaluators find that approximately 93% of English-to-Afaan-Oromo translations, 81% of English-to-Tigrinya translations, and 73% of English-to-Amharic translations exhibit gender bias, making theirs the most comprehensive gender bias benchmark for these three languages to date. Crucially, Sewunetie et al., (2024) show that the bias is rooted in the linguistic structure of each language: Amharic and Tigrinya encode grammatical gender in pronouns, verb agreement, and many occupational titles, while Afaan Oromo conveys gender primarily through context rather than morphology, meaning that bias manifests differently across the three languages. They also demonstrate that professional names in standard datasets are overwhelmingly presented in their masculine form, directly linking dataset-level imbalance to the observed translation errors. This finding motivates our own approach: rather than constructing another template-based occupational benchmark, *Yeswa-Stories* introduces naturally occurring, female-centred narratives that expose translation systems to a broader and more culturally grounded distribution of gendered language. We treat the benchmark of Sewunetie et al., (2024) as a reference point for interpreting our evaluation results and for situating our contribution within the emerging literature on gender bias in Ethiopian language MT.

3 Data Collection

The dataset was constructed through a multi-stage and iterative process. Initially, we aimed to collect stories and narratives specifically centered on Ethiopian women directly from local sources. However, we found that such content was either scarce, not digitized, or not available in a structured format suitable for dataset creation. Similarly, our attempts to locate parallel or comparable articles focusing on women across Amharic, Afaan Oromo, and Tigrinya were largely unsuccessful.

Due to these limitations, we shifted our approach to use English Wikipedia as a starting point. We collected sentences from articles about notable African women, including scientists, public figures, and other well-documented individuals. These sentences served as a base for constructing parallel data.

We then translated the collected sentences into Amharic, Afaan Oromo, and Tigrinya using human translators. This process posed significant challenges due to the limited availability of skilled translators and budget constraints, requiring careful selection of sentences and prioritization of quality over quantity.

To address the lack of cultural representation in Wikipedia-derived content, we further incorporated locally sourced materials written in the target languages. These included articles from Ethiopian news outlets and cultural platforms that reflect traditional roles, festivals, and everyday experiences of women. These texts were then translated into the remaining languages to ensure full three-way parallel alignment.

The final dataset is the result of combining these two sources, balancing globally available narratives with culturally grounded content. This design choice aligns with recent findings that emphasize the importance of incorporating bias considerations during dataset creation, rather than relying solely on post-hoc evaluation (Nigatu et al., 2025).

3.1 Data Description

The final dataset consists of 1,300 parallel sentences aligned across Amharic, Afaan Oromo, and Tigrinya. The dataset includes a mix of globally sourced and locally grounded content. Topics covered include biographies of notable women, descriptions of cultural practices, traditional roles, and narratives related to Ethiopian festivals and community life. This diversity allows the dataset to capture different dimensions of gender representation, moving beyond narrow professional domains and incorporating culturally specific perspectives. All sentences are manually translated, ensuring high-quality alignment and consistency across the three languages. The dataset is intended to support research on gender representation, bias analysis, and inclusive translation.

4 Results

We evaluate the dataset by analysing the performance of two widely used machine translation systems—NLLB (NLLB Team et al., 2022) and Google Translate—on *Yeswa-Stories*. We report BLEU (Post, 2018) and chrF (Popović, 2017) scores computed against human reference translations across all six translation directions. Because Amharic, Tigrinya, and Afaan Oromo are morphologically rich languages in which a single concept can be realised through many valid surface forms, chrF—which operates at the character level and is therefore more tolerant of valid paraphrases—is our primary metric. BLEU scores are reported for completeness and comparability with prior work. Table 1 presents the results. Three patterns emerge clearly from the scores.

Overall scores are low, confirming dataset difficulty. Even Google Translate—a well-resourced commercial system—achieves chrF scores ranging from 25.78 to 49.61, and BLEU scores between 4.79 and 15.80 across the six directions. NLLB performs considerably worse, with BLEU scores as low as 0.87 for Afaan Oromo→Tigrinya. These figures are substantially below scores typically reported on standard benchmarks such as FLORES for the same language pairs, indicating that the female-centred narratives and culturally grounded content in *Yeswa-Stories* present a harder translation challenge than the news and Wikipedia-domain text on which current systems are predominantly trained and evaluated. This supports our central argument that template-based and domain-narrow benchmarks are insufficient for capturing the full range of gendered language use in these communities.

Google Translate consistently outperforms NLLB. Google Translate surpasses NLLB in every direction and on both metrics, with the gap most pronounced for Afaan Oromo as the source language: chrF differences of

		NLLB		Google Translate	
Direction	**#Sents**	**BLEU**	**chrF**	**BLEU**	**chrF**
Afaan Oromo → Amharic	1,287	1.31	14.78	12.43	36.76
Afaan Oromo → Tigrinya	1,299	0.87	12.18	4.79	25.78
Amharic → Afaan Oromo	1,300	4.53	40.29	13.65	49.61
Amharic → Tigrinya	1,300	5.71	29.01	6.99	30.44
Tigrinya → Amharic	1,300	8.80	36.64	15.80	42.27
Tigrinya → Afaan Oromo	1,300	3.26	36.86	9.22	45.52

Table 1: BLEU and chrF scores for NLLB and Google Translate on *Yeswa-Stories* across all six translation directions, evaluated against human reference translations. Higher is better for both metrics.

22.0 points (Oromo→Amharic) and 13.6 points (Oromo→Tigrinya). This pattern replicates and extends the finding of Sewunetie et al. (2024), who similarly observed NLLB underperforming Google Translate on gender-sensitive content for these three languages—there using an occupational benchmark, and here using naturally occurring female-centred narratives. The consistency of this finding across two structurally different datasets strengthens the conclusion that NLLB's training data is particularly poorly suited to gender-inclusive content in Ethiopian languages.

Afaan Oromo as a source language is the hardest direction. Both systems perform worst when translating from Afaan Oromo into either Amharic or Tigrinya. This is consistent with the linguistic profile of the language: as a Cushitic language, Afaan Oromo is typologically distant from both Semitic targets, and it conveys gender primarily through context rather than overt morphological marking (Sewunetie et al., 2024). The absence of explicit gender morphology in the source makes it harder for MT systems to select the correct gendered form in the target, compounding the general data scarcity for this language pair. Conversely, both systems perform best when Amharic is the source language, reflecting its comparatively larger digital presence and the availability of more parallel training data.

4.1 Qualitative Analysis

Beyond aggregate scores, we manually examined a subset of annotations to characterise the types of errors produced by each system. Figure 1 presents four representative examples drawn from the annotation of Afaan Oromo → Amharic translations. The examples illustrate distinct failure modes that aggregate metrics alone cannot capture.

The examples in Figure 1 reveal three recurring error patterns that are particularly consequential for a female-centred dataset. First, NLLB produces **hallucinated content** wholly unrelated to the source: in Example 1, a sentence about Mary Anning's geological expertise is rendered as a description of art and literary work, erasing her scientific identity entirely. Second, both systems exhibit **agency removal**, where the grammatical subject—always a woman in *Yeswa-Stories*—is dropped or neutralised. In Examples 3 and 4, NLLB converts active constructions in which Anning is the explicit actor into passive or impersonal forms, so that her contribution is no longer expressed. This is a direct instance of the representational harm described by Blodgett et al. (2020), the woman is present in the source but invisible in the translation. Third, we observe **lexical substitution errors** that distort factual content, as in Example 2 where Google Translate replaces "cliffs" with "shelters" () and mistranslates the season, producing a sentence that is both factually incorrect and incoherent. These qualitative findings complement the low aggregate scores in Table 1 and illustrate why female-centred narratives such as those in *Yeswa-Stories* are essential for exposing failure modes that template-based occupational benchmarks do not surface.

5 Discussion and Conclusion

Our work highlights several important considerations for gender-inclusive translation in low-resource settings. First, relying solely on global sources such as Wikipedia can lead to skewed representations of women, often emphasizing specific professions while neglecting culturally grounded roles. Second, the inclusion of local data sources is crucial for capturing the diversity of gendered experiences and expressions. This is particularly important for languages and communities that are underrepresented in existing NLP resources.

Finally, the creation of high-quality parallel corpora through human translation remains a resource-intensive process, but it is essential for ensuring linguistic and cultural accuracy. We hope

this resource will support future research on gender bias evaluation and culturally aware machine translation for low-resource languages.

Carbon Impact Statement

For our evaluation experiments, we used Google Colab platform to run inference for the NLLB model using a CPU runtime session. We used the NLLB-base model which has 1.3B parameters. During inference we used maximum token length of 200, and used beam search with a beam size of 4. We also enabled early stopping. Since Google Translate is a closed-source model, we do not have the model details or the runtime environment details.

References

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

Cho, Won Ik, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 173–181.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.

Nigatu, Hellina Hailu, Bethelhem Yemane Mamo, Bontu Fufa Balcha, Debora Taye Tesfaye, Elbethel Daniel Zewdie, Ikram Behiru Nesiru, Jitu Ewnetu Hailu, and Senait Mengesha Yayo. 2025. Evaluating machine translation datasets for low-web data languages: A gendered lens. *Proceedings of the Findings of the Association for Computational Linguistics*.

Nigatu, Hellina Hailu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's Paradox of Low-Resource Languages *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

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, and Al Youngblood et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Popović, Maja. 2017. chrF++: Words helping character n-grams. *Proceedings of the Second Conference on Machine Translation*, 612–618.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191.

Prates, Marcelo O. R., 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.

Ramesh, Krithika, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. *Findings of the Association for Computational Linguistics: EACL 2023*, 2106–2119.

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.

Savoldi, Beatrice, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the harm? Quantifying the tangible impact of gender bias in machine translation with a human-centered study. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18048–18076.

Sewunetie, Walelign, Atnafu Tonja, Tadesse Belay, Hellina Hailu Nigatu, Gashaw Gebremeskel, Zewdie Mossie, Hussien Seid, and Seid Yimam. 2024. Gender bias evaluation in machine translation for Amharic, Tigrigna, and Afaan Oromoo. *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, 1–11.

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

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. *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 26–41.

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. *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas*, 274–286.

Source (Afaan Oromo)	MT Output (Amharic)	System	Gender Bias	Adequacy	Annotator Comment
Anning became well known in geological circles in Britain, Europe, and America, and was consulted on issues of anatomy as well as fossil collecting	አኒንግ ኮራቱ በዩናይትድ ኪንግደም፣ በአውሮፓና በአሜሪካ ውስጥ ታዋቂነት ያተረፈ ሲሆን በሥነ-ጥበብና በሥነ-ጽሑፍ ሥራዎች እንዲሁም በሥነ-ጽሑፍ አዘጋጅነትም ይታወቃል	NLLB	Yes	No	Unrelated translation: content about art and literature substituted for geology and fossil collecting
Anning became well known in geological circles in Britain, Europe, and America, and was consulted on issues of anatomy as well as fossil collecting	አኒንግ በብሪታንያ፣ አውሮፓ እና አሜሪካ ታዋቂ ተመራማሪ እና የእንስሳት አፅም እና የታሪክ ቅሪቶች አማካሪ ናቸው።	Google	No	Yes	Accurate rendering of her role as researcher and advisor
The cliffs could be dangerously unstable, especially in winter when rain weakened them, causing landslides	እነዚያ መጠለያዎች በተለይ በበጋ ወቅት ወንዞች በሚጥሉበት ጊዜ ችግር እና አደጋ ሊያስከትሉ ይችሉ ነበር።	Google	No	No	"መጠለያዎች" (shelters) used instead of ``cliffs''; season also mistranslated ("winter" is translated as "summer")
Her investigations played a crucial role in several finds...coprolites...belemnite fossils	በወቅቱ "የዞር ድንጋዮች"...ከብዙዎች የበለጠ ተገልጧል	NLLB	Yes	No	Named entity mistranslation; action no longer attributed to her --- agency erased and sentence turned neutral
Henry De la Beche...sold prints for her benefit	ይሁን እንጂ...የተሰራጩት በአኒንግ የተገኙት ቅሪተ አካላት ላይ የተመሠረቱ ነበሩ።	NLLB	Yes	No	Subject (she/her) removed; sentence turned passive and neutral --- her agency in the discovery is no longer expressed

Figure 1: Representative annotation examples from the Afaan Oromo → Amharic subset of Yeswa-Stories. Underlined phrases highlight the specific mistranslated or problematic text identified by annotators. Gender Bias and Adequate Translation are majority-vote human annotation labels (Yes/No). We put a figure of the table since the conference guidelines do not support Ge'ez script fonts.

A Additional Result

In this section, we provide qualitative results. In Figure 1, we provide representative annotations to complement the quantitative results reported in Table 1.