

Measuring the Cost of Variety Conflation in Multilingual MT Evaluation: Adding Mozambican Xichangana, Nyanja and Sena to FLORES+

Felermينو D. M. A. Ali¹, Delfina Lázaro Mateus², Manuel Valente Manguê²

¹Faculdade de Engenharia da Universidade Lúrio,
Bairro Eduardo Mondlane, Pemba, Mozambique

²Universidade Eduardo Mondlane - Escola de Comunicação e Artes,
Av. Samora Machel, Maputo, Mozambique
felermينو.ali@unilurio.ac.mz

Abstract

In this paper, we extend FLORES+ with Portuguese-source evaluation sets for three Mozambican Bantu varieties: Xichangana, Mozambican Nyanja, and Sena. We compare Xichangana with the existing Tsonga reference and Mozambican Nyanja with Chichewa, and evaluate NLLB-200, Google Translate, GPT, and a variant-aware NLLB model. Holding system output fixed reveals substantial reference sensitivity. On *devtest*, changing only the reference from Tsonga to Xichangana reduces spBLEU by 13.10 points for NLLB-200 and 15.30 for Google. On matched Nyanja subsets, replacing Chichewa with Mozambican Nyanja produces smaller but consistent reductions of 3.03 and 6.10 spBLEU, respectively. Variant-aware fine-tuning reverses this pattern on the intended targets: relative to NLLB-200, it improves Xichangana by 7.04 spBLEU and Mozambican Nyanja by 5.33 on *devtest*, while losing performance on the sibling references. GPT is competitive on Tsonga and Chichewa but substantially weaker on the Mozambican varieties. For Sena, the finetuned model reaches 12.64 spBLEU and 36.21 chrF++ on *devtest*. These findings motivate variety-aware language identifiers, references, and reporting for cross-border languages or language dialects/variants. The data is publicly available on Hugging Face at https://huggingface.co/datasets/MOZNLP/FLORES_MOZ.

1 Introduction

Bantu languages are spoken across wide, contiguous regions of Africa, and colonial-era borders rarely coincide with linguistic ones. A single language is frequently spoken in several countries, developing varieties that differ in lexicon, orthography and, above all, in the languages they borrow from. These differences determine whether a speaker can actually use a translation system, and whether an evaluation score is a meaningful estimate of that usability.

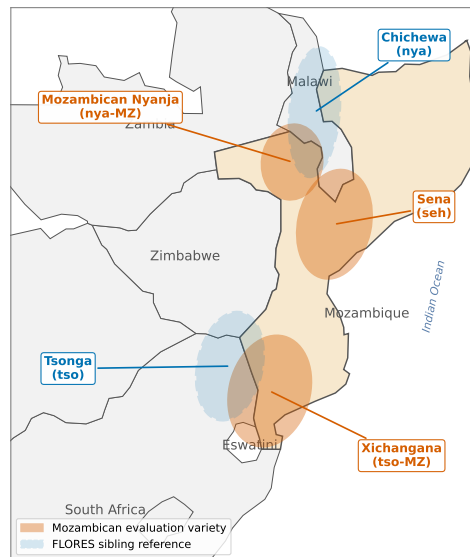


Figure 1: Locations of the evaluated Mozambican varieties and their FLORES sibling references.

Lexical borrowing is the clearest illustration. Malawian Chichewa and the Nyanja of Tete and Niassa in Mozambique share the code *nya*, but the former borrows overwhelmingly from English and the latter borrow from both Portuguese and English. A speaker in Tete reading *matilikula*, *munisipiyo* or *dhowutori* encounters forms that are at best unfamiliar, where *regisita*, *tawoni chipi* and *dokota* would be expected. The same asymmetry holds between South African Tsonga and Mozambican Xichangana, both tagged *tso*. Because FLORES+ contains only the South African and Malawian varieties, systems evaluated on those sets are credited for output a Mozambican user would not accept — the argument recently made for Sudanese Arabic by Samil and Adelani (2026).

We contribute (i) FLORES+ dev and devtest sets for **Xichangana** and **Mozambican Nyanja**, translated from Portuguese by native speakers of those varieties; (ii) the first FLORES+ set for **Sena** (*seh_Latn*), a two-million-speaker language of the

Zambezi valley almost absent from public corpora; (iii) a controlled comparison showing the existing tso_Latn and nya_Latn sets are not valid proxies for the Mozambican varieties; and (iv) a fine-tuning experiment showing the gap is a variety effect and is cheaply addressable. With Emakhuwa (Ali et al., 2024), these sets bring Mozambican coverage in FLORES+ from one language to four.

2 Related Work

FLORES progressed from two low-resource language pairs (Guzmán et al., 2019) to 101 languages (Goyal et al., 2022) and subsequently 200 languages (NLLB Team, 2022), while establishing a multi-stage translation–review–adjudication protocol. FLORES+ is now maintained through community governance under the Open Language Data Initiative (Open Language Data Initiative, 2024).

Recent contributions have expanded coverage, introduced language varieties, and audited existing references. Kalejaiye et al. (2025) add four minority Nigerian languages and demonstrate their poor coverage by current models. Ali et al. (2024) contribute Emakhuwa, previously the only Mozambican language in FLORES+, using Portuguese as the source and documenting orthographic instability. We follow a similar protocol while adding Sena and two Mozambican varieties.

Our variety-focused contribution is closest to the Sudanese Arabic extension of Samil and Adelani (2026). However, because FLORES+ already contains sibling references for Tsonga and Chichewa, we can directly measure how evaluation changes when the reference represents the intended Mozambican variety and whether targeted fine-tuning reduces the mismatch.

3 The Three Languages

Portuguese is the official language of Mozambique but the first language of a minority; most of the population speaks one of roughly twenty Bantu languages natively. **Xichangana** (tso) is spoken by approximately 1.8M people in Gaza, Maputo and Inhambane, and belongs to the Tswa–Ronga group with South African Xitsonga (see Figure 1), from which it diverges in borrowing source, orthographic convention, and administrative register. **Nyanja** (nya) is spoken by approximately 1.2M people in Tete and Niassa, continuous with Chichewa in Malawi and Chinyanja in Zambia. **Sena** (seh) is spoken by approximately 2M people along the

lower Zambezi; unlike the others it has its own code and no FLORES+ entry at all, and it is the least standardised of the three in writing, with religious translation the dominant written genre.

4 Approach

We follow the FLORES+ contribution guidelines (Open Language Data Initiative, 2024) so that the resulting sets are directly comparable to existing entries.

4.1 Source data

We translate the *dev* (997 sentences) and *devtest* (1,012 sentences) splits of FLORES+ from Portuguese (*pt*) into each of the three target varieties. Portuguese was chosen as source rather than English for two reasons. Practically, our translators are proficient in Portuguese and their native language but not uniformly in English; Portuguese is the language of schooling and administration in Mozambique and thus the language in which bilingual professional translators for these varieties actually exist. Following Ali et al. (2024), who translated Emakhuwa from Portuguese for the same reason, we treat translator proficiency as a stronger determinant of reference quality than source-language canonicity.

Because FLORES+ is multi-parallel, the Portuguese side is itself a professional translation of the same English source used by every other entry, so our sets remain sentence-aligned with tso_Latn and nya_Latn. This is what makes the controlled comparison in Section 7 possible: any score difference is attributable to the target variety alone. Pivoting risks propagating Portuguese-side artefacts, which we revisit in Limitations. Translators saw full document context and worked exclusively from the Portuguese side.

4.2 Target variety and translators

For each language, we collaborated with two language experts, all of whom were trained linguists by the Bantu Linguistics Department at Eduardo Mondlane University. Translators had to be native speakers of the targeted variety, resident where it is spoken, professionally competent in Portuguese, and experienced in Portuguese-to-target translation. All contributors were paid at or above local professional rates.

4.3 Translation platform

We developed a custom web-based computer-assisted translation (CAT)¹ platform to support translation, quality control, and review (see Figure 2). The platform was mobile-friendly and allowed translators to work offline, with completed work and local changes synchronized automatically when an Internet connection became available. This was particularly important for participants working in settings with intermittent connectivity. Translators were assigned individual segments and given access to bilingual Portuguese–target-language dictionaries. For each source segment, the CAT platform automatically identified and highlighted potentially relevant glossary entries, helping translators select appropriate terminology and maintain consistency. The platform also performed automatic quality-assurance checks and displayed warnings for potential misspellings, mismatched punctuation or numbers, possible untranslated text, disproportionate source–target lengths, and inconsistent capitalization. Translators could inspect each warning, revise the translation where necessary, or confirm that the flagged difference was intentional. Completed segments were assigned to reviewers for independent verification. Reviewers could correct an issue directly or return the segment to the translator with feedback. Returned segments remained open until the translator submitted a revision and the reviewer approved it. A segment was considered complete only after this review cycle had been successfully concluded.

5 Reference-Variety Analysis

Before evaluating MT systems, we compare each Mozambican reference with the corresponding FLORES sibling variety. The analysis covers all 2,009 aligned segments: 997 from *dev* and 1,012 from *devtest*. Because the references were translated independently, dissimilarity may reflect variety, orthographic convention, lexical choice, paraphrase, or data quality.

5.1 Mozambican Nyanja and Chichewa

Mozambican Nyanja and Chichewa are related but far from textually interchangeable. Mean character similarity is 54.39%, median character similarity is 54.98%, mean token-set overlap is 14.33%, and vocabulary Jaccard overlap is 17.37%. No sentence pair is an exact match. The corpora remain

Reference pair	Segments	Char. sim.	Token overlap	Length ratio	< 0.5 length
Nyanja MZ / Chichewa	2,008	54.39	14.33	1.05	0
Xichangana / Tsonga	2,008	48.97	9.47	0.80	40

Table 1: Reference-pair diagnostics. The Nyanja MZ/Chichewa row uses only trusted assignee-B segments; the Xichangana/Tsonga row uses the complete splits. Similarity and overlap are percentages; length ratio is the sentence-level ratio of Mozambican-reference tokens to sibling-reference tokens. The final column counts Mozambican segments with fewer than half as many tokens as the sibling reference.

broadly balanced in length with a mean sentence-level token-length ratio of 1.05. No Mozambican segment contains fewer than half as many tokens as its Chichewa counterpart. The observed divergence is therefore not explained by systematic truncation.

The strongest recurring distinction is the treatment of *c* and *ch*. A bare *c* not followed by *h* occurs 1,610 times in the Mozambican reference but 367 times in Chichewa; *ch* occurs 281 and 1,849 times, respectively. The most frequent aligned orthographic correspondences include *calcha* and *nchito/ntchito*, each observed 12 times, followed by *awili/awiri* and *nzindalmzinda*. Additional examples include *cifukwa/chifukwa*, *cakalchaka*, and *mwacitsanzolmwachitsanzo*.

The difference is not reducible to spelling. Recurrent substitutions such as *zimenel/zomwe*, *amenel/omwe*, and *monga/ngati* indicate differences in lexical and grammatical selection. Names and borrowed forms also vary, as illustrated by *afil-ikalafrika*.

5.2 Xichangana and Tsonga

Xichangana and Tsonga show a stronger separation. Mean character similarity is 48.94%, median character similarity is 49.47%, mean token-set overlap is 9.47%, and vocabulary Jaccard overlap is 9.87%. Again, no sentence pair is an exact match. The Xichangana reference contains 40,893 tokens, compared with 52,443 in Tsonga, yielding a mean token-length ratio of 0.80.

The most prominent orthographic distinction is nearly complementary: *sv* occurs 3,681 times in Xichangana and only 5 times in Tsonga, whereas *sw* occurs 42 and 4,191 times, respectively. Examples include *svalswa*, *lesvaku/leswaku*, *lesvilleswi*, and *nasvona/naswona*. Other recurring correspondences include *lakra*, *doropal/doroba*, *ndzhaku/endzhaku*, *nkama/nkarhi*, and *mpimu/mpimo*. Function words and agreement

¹CATfrica: <https://mozcat.vercel.app>

forms also vary, including alternations among *nalni*, *kalekalku*, and several noun-class agreement markers.

6 Experimental Setup

We evaluate three systems for Portuguese-to-target translation. The **NLLB-200** baseline is facebook/nllb-200-distilled-600M. **Google** denotes the Google Cloud Translation general NMT model², with Portuguese specified as the source and ny or ts as the target. Google provides no separate identifiers for the Mozambican varieties, so each Google output is scored against both the sibling and Mozambican reference. No Google Sena output was collected.

GPT³ denotes openAI API endpoints. A zero-shot translation prompt (see Appendix B) specified Portuguese as the source and separately described standard Chichewa, Mozambican Nyanja, standard Xitsonga/Tsonga, Xichangana, and Sena as the five targets. Each response was constrained to structured JSON and validated for complete segment and target coverage. GPT therefore generates a distinct output for every target variety, including Sena, rather than sharing one hypothesis across each sibling–Mozambican pair.

NLLB+finetuned starts from the same distilled 600M NLLB checkpoint. We retain the standard *nya_Latn*, *tso_Latn*, and *seh_Latn* tokens for Chichewa, Tsonga, and Sena, and introduce *nya_MZ_Latn* and *tso_MZ_Latn* for Mozambican Nyanja and Xichangana. The two new token embeddings are initialized when the tokenizer vocabulary is extended and learned during fine-tuning. We also prepend an explicit target marker such as <2nya-MZ> or <2tso-MZ> to each source sentence. The fine-tuned model can therefore generate different hypotheses for the sibling and Mozambican targets.

6.1 Fine-tuning data

Table 2 reports the amount of direct Portuguese data and the total number of eligible rows involving each target variety, together with the corpus or domain coverage represented in the training data.

6.2 Optimization

Training uses four NVIDIA RTX A6000 GPUs with DistributedDataParallel.

The model is trained for 15,000 optimizer steps with maximum sequence length 200. We use Adafactor with learning rate 1×10^{-4} , weight decay 1×10^{-3} , clip threshold 1.0, and a constant schedule after 250 warmup steps. Random seeds are initialized from 222 with a rank-specific offset. Model, tokenizer, optimizer, and scheduler states are saved every 500 steps.

6.3 Evaluation

The aligned evaluation data contain 997 *dev* and 1,012 *devtest* sentences per direction. We translate the complete Portuguese source into *nya*, *nya-MZ*, *tso*, *tso-MZ*, and, where supported, *seh*.

For the off-the-shelf NLLB baseline, the sibling and Mozambican labels map to the same standard NLLB tokens, so its hypotheses are identical within each pair and only the reference changes. Google similarly produces one *ny* output and one *ts* output per Portuguese sentence. NLLB+finetuned instead uses the distinct regional tokens described above.

We recompute all scores from the saved hypotheses using SacreBLEU 2.6.0. spBLEU uses the flores200 tokenizer⁴; chrF++ uses character order 6 and word order 2⁵. Results are reported separately for *dev* and *devtest*.

7 Results and Discussion

Table 3 reports spBLEU and chrF++ on the Portuguese-source evaluation data. The Tsonga, Xichangana, Chichewa, and Sena columns use the complete splits: 997 *dev* and 1,012 *devtest* sentences.

On the complete sibling-language sets, Google is the strongest system. On *devtest*, it obtains 22.42 spBLEU / 47.89 chrF++ for Tsonga and 18.16 / 45.65 for Chichewa. GPT is close on Tsonga at 21.39 / 47.68 and reaches 15.14 / 44.75 on Chichewa, while NLLB-200 obtains 18.91 / 44.82 and 14.70 / 42.12, respectively. The ranking changes on the Mozambican targets. NLLB+finetuned is strongest for Xichangana at 12.85 / 38.95 and for Mozambican Nyanja at 17.89 / 44.98. It also provides the baseline Sena result on *devtest*, reaching 12.64 / 36.21.

²Queried on September, 2026

³gpt-5.6-sol, queried on September, 2026

⁴nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp

⁵nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no

Target variety	Corpus/domain coverage	Train (pt↔XX)
Chichewa (nya)	Jehovah Witness	56,902
Nyanja (nya-MZ)	Portuguese news: culture, sport, economy	8,164
Tsonga (tso)	Jehovah Witness	54,678
Xichangana (tso-MZ)	Jehovah Witness; UDHR; Legal; Portuguese news: culture, sport, economy	43,662
Sena (seh)	Jehovah Witness; Portuguese news: culture, sport, economy	64,532

Table 2: Data coverage. Counts are directed sentence-pair rows for Portuguese–variety translation directions.

System	Dev					Devtest				
	pt→tso	pt→tso-MZ	pt→nya	pt→nya-MZ	pt→seh	pt→tso	pt→tso-MZ	pt→nya	pt→nya-MZ	pt→seh
NLLB-200	20.67 / 46.12	5.15 / 30.74	15.21 / 42.60	10.41 / 39.08	–	18.91 / 44.82	5.81 / 32.98	14.70 / 42.12	12.56 / 39.70	–
Google	23.94 / 48.84	5.01 / 31.57	18.08 / 45.81	10.11 / 39.54	–	22.42 / 47.89	7.11 / 36.03	18.16 / 45.65	12.37 / 39.96	–
GPT	22.17 / 48.30	7.05 / 33.77	15.07 / 44.84	7.12 / 33.21	9.26 / 34.46	21.39 / 47.68	7.46 / 34.55	15.14 / 44.75	8.71 / 33.52	9.92 / 35.30
NLLB+finetuned	17.42 / 42.30	10.34 / 36.45	10.76 / 36.81	15.87 / 44.13	10.56 / 34.07	17.50 / 42.05	12.85 / 38.95	11.06 / 37.35	17.89 / 44.98	12.64 / 36.21

Table 3: Portuguese-to-XX results on dev and devtest. Each cell reports spBLEU / chrF++.

7.1 Sibling references substantially overestimate Xichangana quality

Xichangana exhibits a large and consistent mismatch with the existing FLORES Tsonga reference. On *devtest*, the same NLLB-200 hypothesis scores 18.91 / 44.82 against Tsonga but only 5.81 / 32.98 against Xichangana, reductions of 13.10 spBLEU and 11.84 chrF++. Google shows the same pattern, falling from 22.42 / 47.89 to 7.11 / 36.03, with reductions of 15.30 spBLEU and 11.87 chrF++.

The mismatch is larger on *dev*. NLLB-200 falls from 20.67 / 46.12 on Tsonga to 5.15 / 30.74 on Xichangana, while Google falls from 23.94 / 48.84 to 5.01 / 31.57. The corresponding differences are 15.53 and 18.93 spBLEU and 15.38 and 17.27 chrF++, respectively. Because these comparisons change only the reference, the gaps cannot be attributed to decoding, prompting, or model selection. They demonstrate that performance measured on *tso_Latn* is not a reliable proxy for performance on Mozambican Xichangana.

This finding is consistent with the reference analysis in Section 5. Xichangana and Tsonga exhibit low token overlap, systematic orthographic differences such as *sv* versus *sw*, and substantial lexical and grammatical divergence. A system can consequently receive a strong Tsonga score while failing to produce forms appropriate for Xichangana.

7.2 Mozambican Nyanja evaluation

Google obtains 10.11 / 39.54 on *dev* and 12.37 / 39.96 on *devtest*, close to NLLB-200 at 10.41 / 39.08 and 12.56 / 39.70. GPT is weaker, reaching 7.12 / 33.21 and 8.71 / 33.52.

In *devtest*, NLLB-200 decreases from 15.59 / 42.80 in Chichewa to 12.56 / 39.70 in Mozambican

Nyanja, resulting in reductions of 3.03 spBLEU and 3.10 chrF++. Google decreases from 18.47 / 45.88 to 12.37 / 39.96, reductions of 6.10 spBLEU and 5.92 chrF++. On *dev*, NLLB-200 decreases from 14.38 / 42.61 to 10.41 / 39.08, while Google decreases from 17.19 / 45.62 to 10.11 / 39.54.

The matched data therefore reveal a systematic Chichewa–Mozambican Nyanja gap, but one smaller than the Xichangana–Tsonga gap.

7.3 Variant-aware fine-tuning

Variant-aware fine-tuning produces substantial gains on both Mozambican targets. On the complete Xichangana *devtest* set, NLLB+finetuned improves over NLLB-200 from 5.81 / 32.98 to 12.85 / 38.95, gains of 7.04 spBLEU and 5.97 chrF++. It also exceeds Google at 7.11 / 36.03 and GPT at 7.46 / 34.55. On *dev*, it reaches 10.34 / 36.45, compared with 5.15 / 30.74 for NLLB-200, 5.01 / 31.57 for Google, and 7.05 / 33.77 for GPT.

On the Mozambican Nyanja subsets, NLLB+finetuned improves over NLLB-200 from 12.56 / 39.70 to 17.89 / 44.98 on *devtest*, gains of 5.33 spBLEU and 5.28 chrF++. It exceeds Google by 5.52 spBLEU and 5.02 chrF++, and GPT by 9.18 spBLEU and 11.46 chrF++. On *dev*, it obtains 15.87 / 44.13, compared with 10.41 / 39.08 for NLLB-200, 10.11 / 39.54 for Google, and 7.12 / 33.21 for GPT. The results therefore show a clear fine-tuning advantage on both splits.

The Mozambican gains coincide with lower scores on the complete sibling sets. On *devtest*, fine-tuning changes Tsonga from 18.91 / 44.82 to 17.50 / 42.05 and Chichewa from 14.70 / 42.12 to 11.06 / 37.35. The same trade-off appears on *dev*, where Tsonga falls from 20.67 / 46.12 to 17.42 /

42.30 and Chichewa from 15.21 / 42.60 to 10.76 / 36.81. This pattern is consistent with the model learning variety-specific lexical and orthographic choices rather than receiving a uniform improvement across related targets. Because the adapted system generates separate hypotheses for each target identifier, the result demonstrates controllable specialization rather than a reference substitution effect.

7.4 Sena, and broader implications

GPT performs competitively on the established sibling languages but is much weaker on the Mozambican targets. On *devtest*, its Tsonga result (21.39 / 47.68) approaches Google, and its Chichewa result (15.14 / 44.75) exceeds NLLB-200 in chrF++. However, GPT reaches only 7.46 / 34.55 on Xichangana and 8.71 / 33.52 on the Mozambican Nyanja subset. This contrast suggests that specifying a variety in a zero-shot prompt does not provide sufficient control for these underrepresented targets.

For Sena, NLLB+finetuned obtains 10.56 / 34.07 on *dev* and 12.64 / 36.21 on *devtest*, compared with 9.26 / 34.46 and 9.92 / 35.30 for GPT. Fine-tuning yields higher Sena spBLEU on both splits and higher chrF++ on *devtest*, although GPT is marginally higher in chrF++ on *dev*. Since no NLLB-200 or Google Sena output is available, this comparison covers the two systems with complete Sena predictions but does not establish improvement over the original NLLB baseline.

Overall, the results support following conclusions: (1) sibling references can substantially overestimate translation quality for Mozambican varieties, especially Xichangana. (2) supervised variety-aware fine-tuning is substantially more effective than generic decoding or zero-shot target naming on the Mozambican targets.

We therefore report sibling and Mozambican references separately. Automatic metrics establish reference sensitivity and the value of target-specific adaptation, but they cannot determine whether an output is acceptable to the intended speakers.

8 Conclusion

We contribute Portuguese-source FLORES+ evaluation sets for Xichangana, Mozambican Nyanja, and Sena. The results show that an existing reference in a closely related national variety is not necessarily an adequate proxy for the community being

evaluated. Under identical hypotheses, replacing the Tsonga reference with Xichangana reduces *devtest* spBLEU by 13.10 points for NLLB-200 and 15.30 for Google. The corresponding Chichewa–Mozambican Nyanja reductions on matched evaluation subsets are smaller but remain systematic, at 3.03 and 6.10 spBLEU. These differences are consistent with the orthographic, lexical, and grammatical patterns observed in the reference analysis and demonstrate that aggregate language codes can conceal meaningful variety-level differences.

Variant-aware NLLB fine-tuning substantially improves performance on the intended Mozambican targets. On *devtest*, it gains 7.04 spBLEU and 5.97 chrF++ over NLLB-200 for Xichangana, and 5.33 spBLEU and 5.28 chrF++ for Mozambican Nyanja. These gains coincide with lower scores on Tsonga and Chichewa, indicating target-specific specialization rather than a uniform model improvement. GPT remains competitive on the established sibling languages but performs markedly worse on the Mozambican varieties, while the adapted model establishes a reproducible Sena result of 12.64 spBLEU / 36.21 chrF++ on *devtest*. We therefore recommend variety-specific references, target identifiers, and disaggregated reporting for cross-border Bantu languages so that benchmark scores correspond to the communities that MT systems are intended to serve.

9 Limitations

Our references were produced from Portuguese rather than the original English source. Although the Portuguese side is itself a professional translation and the sets remain sentence-aligned with all other entries, pivoting risks propagating Portuguese-specific phrasing and existing errors into our targets; an English-sourced control subset is left to future work.

10 Ethical Considerations

All translators were professionals compensated at or above prevailing local rates, gave informed consent for release.

Acknowledgements

This work was carried out with support from lacunafund.org and google.org. **Disclaimer:** The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, or CENIA.

References

- Felermينو Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. [Expanding FLORES+ benchmark for more low-resource settings: Portuguese-Emakhuwa machine translation evaluation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA. Association for Computational Linguistics.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Oluwadara Kalejaiye, Luel Hagos Beyene, David Ifeoluwa Adelani, Mmekut-Mfon Gabriel Edet, Aniefon Daniel Akpan, Eno-Abasi Urua, and Anietie Andy. 2025. [Ibom NLP: A step toward inclusive natural language processing for Nigeria’s minority languages](#). In *Proceedings of the 2025 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AAACL)*.
- NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Open Language Data Initiative. 2024. Open language data initiative: Contribution guidelines. <https://oldi.org>.
- Hadia Mohammedosman Ahmed Samil and David Ifeoluwa Adelani. 2026. [Sudanese-flores: Extending FLORES+ to Sudanese Arabic dialect](#). In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP)*, pages 243–247, Rabat, Morocco. Association for Computational Linguistics.

Language	Reference sentence
Example 1: “Photons are even smaller than the stuff that makes up atoms!”	
Portuguese source	Os fótons são ainda menores do que as partículas que formam os átomos!
Mozambican Nyanja	Fotoni tili tocepa kwambiri kuposera mbali zimene zimapanga matomo!
Chichewa	Ma photon ndi ocheperapo kuposa zinthu zomwe zimapanga ma atom.
Xichangana	Mafotoni mahali matsongo kutlula sviphemu lesvi sviwumbaka ma-atomo!
Tsonga	Tifothoni i titsongo eka swilo leswi endlaka tiathomu!
Example 2: “It’s worth half an hour to stroll about the intriguing village.”	
Portuguese source	Vale a pena caminhar meia hora pela intrigante vila.
Mozambican Nyanja	Ndi kosangalatsa kwambiri kuyenda hafu awala muvila yocititsa kasoyo.
Chichewa	Ndikofunikila theka la ola kuuyendela mudzi wopatsa chidwiwo .
Xichangana	Svayampsa kufambafamba hafu ya wara hi mugaga wohlamalisa .
Tsonga	Swi fanele awara na hafu ku famba famba eka tiko xikaya leri ro tsakisa .

Table 4: Aligned examples showing the English reference, Portuguese source, Mozambican varieties, and their FLORES sibling references. Boldface highlights selected orthographic and lexical contrasts;

A Reference-Variety Examples

Table 4 presents two aligned examples from the evaluation data. For each example, it shows the English reference, the Portuguese source used for translation, the Mozambican Nyanja and Chichewa references, and the Xichangana and Tsonga references. Selected forms are highlighted to make recurring orthographic and lexical contrasts easier to identify.

B GPT translation prompts

The following system and user prompts were used for GPT translation. In the user prompt, <INPUTS_JSON> represents the JSON serialization of a batch of Portuguese source segments, each containing its split, segment identifier, and source text. Long lines below are wrapped for typesetting; the wrapped portions are joined with spaces in the actual prompt.

System prompt.

You are a professional African-language translator. Translate faithfully from Portuguese. Preserve meaning, names, numbers, punctuation, and sentence boundaries. Produce natural target-language text, not explanations. Keep regional varieties distinct. Return valid JSON only.

User prompt.

Translate every input into all five targets.

Target definitions:

- "nya": standard Chichewa as used in Malawi; use standard Malawian Chichewa vocabulary and orthography
- "nya-MZ": Mozambican Nyanja; use vocabulary and orthographic conventions appropriate for Nyanja speakers in Mozambique
- "tso": standard Xitsonga/Tsonga as used in South Africa; use standard South African Xitsonga vocabulary and orthography
- "tso-MZ": Xichangana as used in Mozambique; use Mozambican vocabulary and orthography rather than South African Xitsonga
- "seh": Sena as used in Mozambique; use natural Sena vocabulary and orthography

Return exactly this JSON shape:

```
{"translations":[{"split":"dev","segment_id":1,"nya":"...",
"nya-MZ":"...", "tso":"...", "tso-MZ":"...", "seh":"..."}]}
```

Return one object per input, in the same order. Do not omit, merge, or add segments. Do not include Markdown.

Inputs:

<INPUTS_JSON>

Each input object in <INPUTS_JSON> has the following form:

```
{"split":"dev","segment_id":1,"portuguese":"..."}
```

C The CATfrica Translation Tool

CATfrica is a web-based Computer-Assisted Translation (CAT) tool built to support translation work for low-resource languages. It covers the whole cycle of a translation job: project preparation, task distribution, translation, quality control, revision, delivery and payment.

Roles and workflow. Five roles are supported: administrator, project manager, translator, revisor and annotator. A project passes through the stages *draft*, *in progress*, *under review*, *approved* and *paid*, and each sentence carries its own status, so different parts of a document can be assigned to different translators and progress can be monitored precisely. Large documents can be split automatically into sub-projects translated in parallel.

Project preparation. Managers upload one or more text files and choose sentence- or paragraph-level segmentation. The original layout is preserved, so the translated document can be rebuilt with the same paragraphs, spacing and punctuation. Translators may split or merge segments when the automatic division is inadequate.

Translation editor. Source and target are shown side by side. For every sentence the tool automatically offers previous translations of identical or similar sentences from the project's translation memories, terminology suggestions from its glossaries, and machine translation. It further provides search and replace, concordance search, per-sentence comments, version history and automatic saving. If the connection is lost, work continues offline and is uploaded automatically when connectivity returns.

Quality assistance. A spell checker flags words absent from the target-language word list and proposes corrections. Automatic checks warn about unusual length differences, unbalanced brackets or quotes, missing or altered numbers, inconsistent final punctuation, double spaces, wrong apostrophes and stray line breaks. Both components were adapted to the orthographic conventions of African languages; word-internal apostrophes, for instance, are not treated as errors.

Memories, glossaries and annotation. Translation memories and glossaries can be created in the tool or imported from spreadsheets and attached to any project. Current version support biligual dictionary for Portuges— Nyanja, Xichangana, Emakhuwa and Sena, and English — Swahili, Xhosa. Matching is tolerant: suggestions appear for merely similar sentences, and terms are recognised in inflected forms. Terms may be labelled as ordinary terms, loanwords, code-switching or bilingual expressions. An alignment function additionally lets annotators link a source expression to its counterpart in the translation, so that annotated bilingual lexicons and code-switching resources are produced alongside the translation itself.

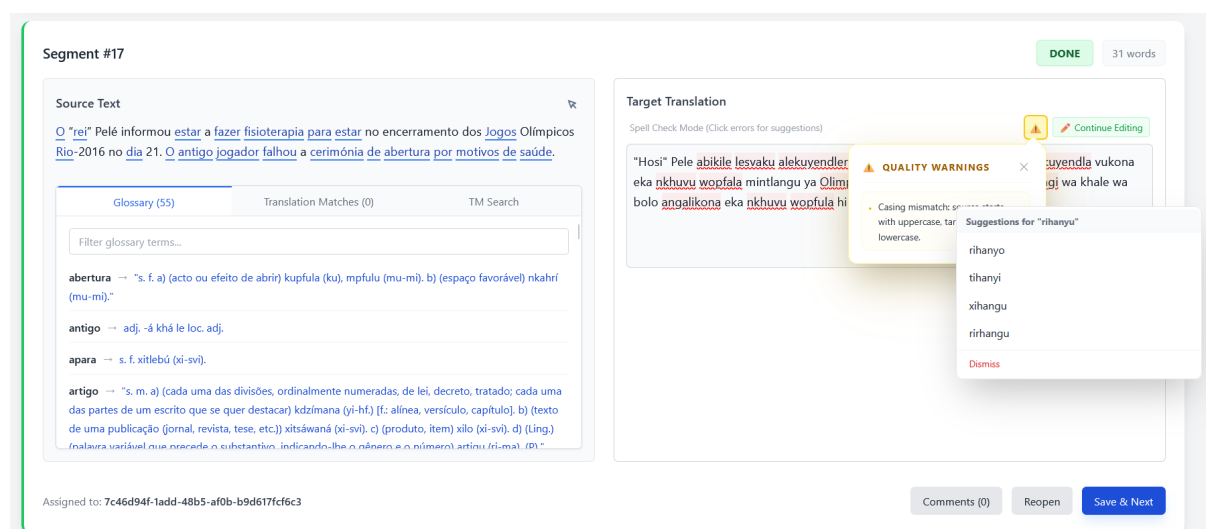


Figure 2: Screenshot of the custom web-based computer-assisted translation (CAT) platform to support translation.

Machine translation. Each project can be linked to an engine adapted to its language pair. Suggestions are reused when the same sentence recurs, and models can alternatively be downloaded and run on the translator's own computer, which matters where connectivity is limited.

Revision and delivery. Finished translations are reviewed sentence by sentence by a revisor, who cannot be the original translator; rejected sentences return to the translator for correction. Approved projects are exported as formatted text, as an archive with one file per uploaded document, as structured data including the alignments, or in a standard translation-memory format for reuse as a resource or as training data.

Payment and administration. Managers set a per-word rate; earnings are computed from completed words and settled through invoices. Users purchase word credits via PayPal or M-Pesa, with all transactions recorded in a transparent statement. An administration area manages accounts and invitations, supported languages, translation engines and backups.