Expanding FLORES+ Benchmark for more Low-Resource Settings: Portuguese-Emakhuwa Machine Translation Evaluation

Felermino D. M. A. Ali[1,2,3,5], Henrique Lopes Cardoso[1,2], Rui Sousa-Silva[3,4]
[1]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC / LASI)
[2]Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[3]Centro de Linguística da Universidade do Porto (CLUP)
[4]Faculdade de Letras da Universidade do Porto, Via Panorâmica, 4150-564 Porto, Portugal
[5]Faculdade de Engenharia da Universidade Lúrio, Pemba 3203, Mozambique
{up202100778, hlc}@fe.up.pt, rssilva@letras.up.pt

Abstract

As part of the Open Language Data Initiative shared tasks, we have expanded the FLORES+ evaluation set to include Emakhuwa, a low-resource language widely spoken in Mozambique. We translated the *dev* and *devtest* sets from Portuguese into Emakhuwa, and we detail the translation process and quality assurance measures used. Our methodology involved various quality checks, including post-editing and adequacy assessments. The resulting datasets consist of multiple reference sentences for each source. We present baseline results from training a Neural Machine Translation system and fine-tuning existing multilingual translation models. Our findings suggest that spelling inconsistencies remain a challenge in Emakhuwa. Additionally, the baseline models underperformed on this evaluation set, underscoring the necessity for further research to enhance machine translation quality for Emakhuwa. The data is publicly available at https://huggingface.co/datasets/LIACC/Emakhuwa-FLORES

1 Introduction

Evaluation datasets and benchmarks are essential for advancing Natural Language Processing (NLP) models. They provide the necessary tools for assessing model performance and guiding further enhancements. However, the scarcity of evaluation datasets and benchmarks for low-resource languages has significantly hindered the progress of NLP technologies in these languages. Recognizing this challenge, the FLORES+ evaluation set has emerged as a critical tool for the Machine Translation (MT) community, especially in low-resource languages. It promotes a more inclusive approach to language technology development across diverse linguistic landscapes. This work focuses on expanding the FLORES+ (NLLB Team et al., 2022) evaluation set to include Emakhuwa, a low-resource language spoken in Mozambique by approximately 9 million people. Our dataset consists of the *dev* and *devtest* sets managed by the Open Language Data Initiative[1] (OLDI), which contain 997 sentences and 1012 sentences, respectively. Throughout our data collection process, we implemented robust quality assurance mechanisms, including thorough post-editing. The resulting dataset features multiple reference translations derived from these post-editing efforts.

2 Related Works

The Flores v1.0 MT evaluation set was introduced by Guzmán et al. (2019). This initial version focused on two language pairs: Nepali–English and Sinhala–English, with the data divided into *dev*, *test*, and *devtest* splits. After its release, the dataset was gradually expanded to include more languages. A significant expansion came with the work of Goyal et al. (2021), who introduced Flores-101, extending the evaluation set to support 101 languages. Further expansion was done with the release of Flores-200 by the NLLB team (NLLB Team et al., 2022) in 2022, which increased the language coverage to 204 languages. Additional contributions include Doumbouya et al. (2023), who added the Nko language, as well as AI4Bharat et al. (2023), who incorporated Bodo, Dogri, Meitei, Sindhi, and Goan Konkani into the dataset. These contributions have significantly broadened the opportunities for low-resource languages in MT, allowing researchers to track the progress of MT systems on these expanded evaluations. However, the coverage remains limited, especially considering that there are over 7,000 languages worldwide. One such language that remains underserved is Emakhuwa, which still lacks datasets for MT.

[1]https://oldi.org/

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 579–592
November 15–16, 2024.

3 Emakhuwa

Emakhuwa, alternatively referred to as Makua, Macua, or Makhuwa, belongs to the Bantu language family and is predominantly spoken in the northern and central regions of Mozambique, specifically in the Nampula, Niassa, Cabo Delgado, and Zambezia provinces. There are eight variants of Emakhuwa, with Emakhuwa-Central (ISO 639-3 code *vmw*) being the standard variety (Ngunga and Faquir, 2014).

Emakhuwa follows the Subject-Verb-Object (SVO) structure, use a Latin scripts (ISO 15924 *Latn*), and is gender-neutral. Furthermore, similarly to other languages in the Bantu family, it is linguistically rich, with complex morphology featuring agglutinative and tonal attributes.

3.1 Challenges in Emakhuwa

Emakhuwa digital resources are scarce, and the spelling standards are still under development. While a fully standardized system is not yet in place, the existing guidelines (Ngunga and Faquir, 2014) offer a critical framework for contemporary written communication in Emakhuwa. One problem stressed in official standardization (Ngunga and Faquir, 2014) is the lack of guidance on tonal marking. Consequently, existing materials exhibit inconsistent spelling, particularly when marking tone, which is essential in Emakhuwa for disambiguation. To give an example, let us consider two words carrying distinct meanings: *omala* and *omaala* / *omàla*; *omala* means “to finish", while *omaala* / *omàla* means “to silence" or "to hush." In this case, the tonal marker *aa* / *à* clarifies the intended meaning.

Spelling variations are largely evident in existing Emkahuwa text corpora, where some use diacritics (e.g., *à*, *è*, *ì*, *ò*, *ù*) and consonantal sounds (e.g, *kh*, *nn*) for tonal marking, others use vowel lengthening (e.g., *aa*, *ee*, *ii*, *oo*, *uu*), and some even use a combination of methods. Emakhuwa’s agglutinative nature with complex morphology further amplifies spelling discrepancies. Since tonal variations often occur at the morpheme level, different combinations of morphemes result in varied spellings of the same word.

These spelling inconsistencies create significant obstacles for language technology processes. They lead to data sparsity, as some spelling variants appear less frequently, which impairs the model’s ability to learn the language’s nuances effectively. This sparsity inflates the vocabulary size and can result in reduced performance of language technologies.

An additional challenge in Emakhuwa that contributes to inconsistencies is the adaptation of loanwords. Emakhuwa text corpora frequently contain Portuguese loanwords with inconsistent adaptations due to the absence of standardized guidelines for integrating borrowed terms (Ali et al., 2024). These loanwords are adapted in one of three ways: phonetically to match Portuguese pronunciation, in alignment with Emakhuwa phonotactics, or retained unchanged from Portuguese.

4 Methodology

We chose to translate the *devtest* and *dev* sets from Portuguese (*pt*) into Emakhuwa (*vmw*) because our translators were only proficient in these two languages. We focus specifically on the central variant of Emakhuwa, as it is the standard and established language variant.

The translators were selected based on their proficiency in these languages and their proven experience in Portuguese-Emakhuwa translation. In total, we collaborated with five experts: two were assigned the tasks of translation and revision, while the remaining collaborators were responsible for evaluating the translations (refer to Table 6 in the appendix for more details).

In general, we implemented the workflow as a peer review process, divided into three main steps: Data Preparation, Translation, and Validation. Below is a detailed description of each step (refer to Figure 1).

4.1 Data Preparation

We compile the sentences in *devtest* and *dev* sets as segments and then load them to the Matecat[2] CAT (Computer-Assisted Translation) tool. Before assigning translation tasks, we prepare a guideline and glossary. The guidelines were adapted from the Open Language Data Initiative guidelines[3], written in Portuguese and suggesting that the translated text should adhere to the latest orthography standards of the central variant of Emakhuwa. On the other hand, the glossary was built by digitizing existing bilingual dictionaries and the glossary of Political, Sports, and Social Concepts from Radio of Mozambique (Moçambique E.P., 2016). We con-

[2]https://www.matecat.com/
[3]https://oldi.org/guidelines

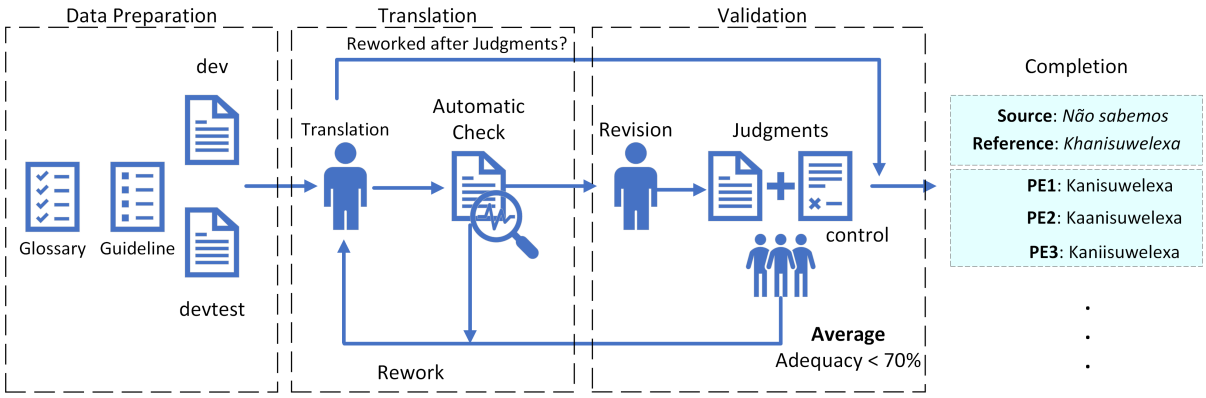

Figure 1: Workflow

ducted a small workshop to familiarize the team with the guidelines and gather feedback to improve them. The translation team found the glossary helpful, as it prevented using loanwords for existing Emakhuwa terms and ensured consistency in translations.

4.2 Translation

Translation tasks were subdivided between two translators: one worked on the *devtest* segments, and the other on the *dev* segments. Once all segments were translated, they were submitted to our spell checker system for an automatic check to identify potential misspellings (refer to Figure 10 in the appendix). We then provided feedback to the translators, asking them to review and refine their work if necessary.

4.3 Validation

The validation corresponds to two steps: revision and Judgments.

4.3.1 Revision

Following the translation step, we swapped the translated works between the two translators, asking them to post-edit each other's translations on the Matecat platform. Table 1 provides the Quality Report generated by Matecat, which includes various metrics used to evaluate the translation based on the revisions made. The report indicates that the reviewer working on *devtest* made more suggestions. A closer examination of the error categories on *devtest* (refer to Figure 9 in the appendix) reveals that most of the issues identified in the translation fell under the category of "Language Quality", meaning grammar, punctuation, and spelling errors. On the other hand, the reviewer of the *dev* set identified mostly errors related to "terminology and language consistency", suggesting that the translator was not consistently using the proper terms and maintaining uniformity throughout the text.

	dev	**devtest**
Post-Editing Effort	99%	95%
Time to edit	02m38s	05m42s
Quality score	23.31	54.22
Avg. Edit Distance	0.23 ± 1.77	7.09 ± 11.94

Table 1: Matecat's quality report post revision.

4.3.2 Judgments

Once all segments have been revised, we perform a second translation quality assessment using a Direct Assessment (DA) pipeline similar to the one described by Guzmán et al. 2019. Judgments were collected using our annotation tool (see Figure 7 in the appendix), and involve the following aspects.

Direct Assessment Three different raters evaluate the translation adequacy (i.e., the perceived translation quality) on a scale from 0 to 100. A score of 0 means that "no meaning was preserved in the translation". Scores from 1 to 34 - "the translation preserves some of the source meaning but loses significant parts", 35 and 67 - "the translation retains most of the source meaning", 68 to 99 - "the translation is consistent with the source text", and a score of 100 means "the translation is perfect". These quality intervals are inspired by the study of Wang et al. 2024.

Control To ensure raters' attentiveness and improve consistency during the evaluation, we included control instances with incorrect translation pairs. These incorrect pairs were generated using

the Madland-400-3bt[4] model (Kudugunta et al., 2024), a multilingual MT system that supports the Emetto variant of Emakhuwa (ISO 639-3 *mgh*). While this model typically performs poorly when translating from Portuguese to Emakhuwa, it produces similar words that can mislead inattentive annotators. Based on these control translations, we provided feedback to the evaluators as they progressed in their tasks. We used emojis to give the feedback in our annotation tool: a ☺️ appeared if less than 25% of control translations were incorrectly rated (i.e. scores above 34 points), 😠 if 25%-50% are incorrectly rated, 😡 if 50%-75% are incorrectly rated, and 😭 if more than 75% of control translations are rated too highly.

Post Editing During the validation phase, we asked evaluators to post-edit translations with lower scores to enhance fluency and better align them with the source sentence's meaning. However, this task was made optional to prevent evaluators from inflating scores to avoid additional post-editing work.

Standard orthography To assess the perceived usage of standard orthography, raters also judged whether the translated text used standard orthography on a scale from 1 (not using standard orthography) to 5 (entirely written in standard orthography).

Finally, we calculate the average score for each segment. We then returned segments scoring below 70 to the translator for reworking. Figure 2 shows the histogram of the average translation scores.

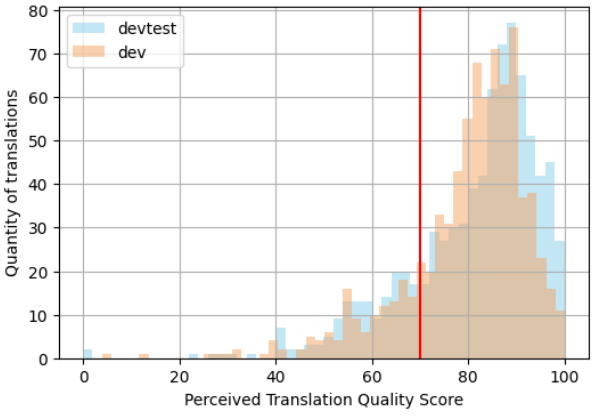

Figure 2: Averaged Translation Quality Score Histogram on both *dev* and *devtest* sets. Translations with an average score below 70 (indicated by the red line) were returned to the translator for rework.

[4]https://huggingface.co/google/madlad400-3b-mt

4.4 Analysis

Figures 3 and 4 show the raw scores per annotator for Direct Assessments. Given the mean scores, in both the *test* and *devtest* sets, Annotator 1 and Annotator 2 gave higher quality scores, while Annotator 3 was more critical but still within the spectrum of acceptable translations. This suggests a generally positive perception of the translations produced. Figure 5 displays the Direct Assessment scores on the control set. Annotator 1 and Annotator 3 have median scores below the threshold of 34 points, suggesting that, as expected, they have generally assessed the control translations as low quality. Annotator 2, however, has a median score above the threshold, suggesting a trend to a more positive assessment compared to the other two annotators and was less attentive among the annotators.

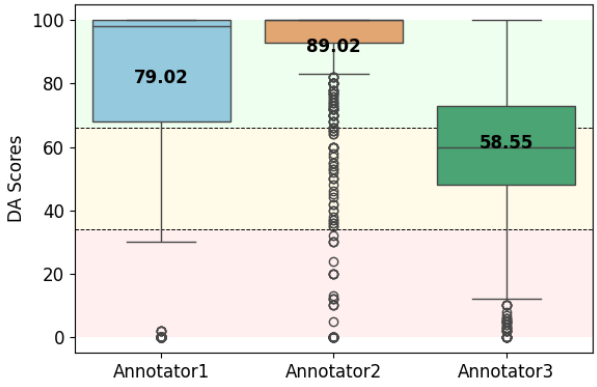

Figure 3: Direct Assessment adequacy scores per annotator on *dev* set

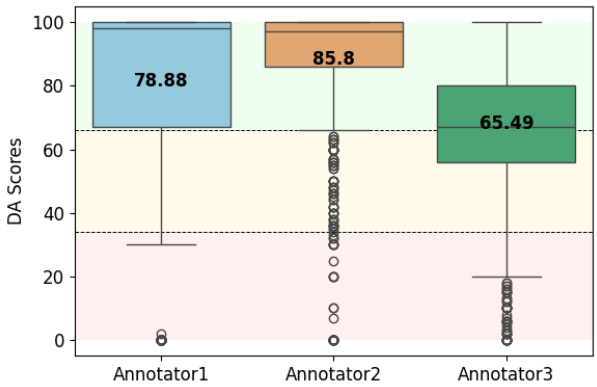

Figure 4: Direct Assessment adequacy scores per annotator on *devtest* set

Table 2 provides the reliability results for adequacy and standard orthography usage assessments. The inter-class correlation for adequacy is 0.67 for

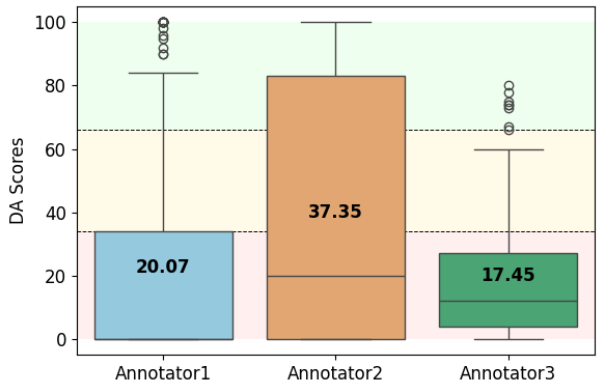

Figure 5: Direct Assessment adequacy scores per annotator on *control* set

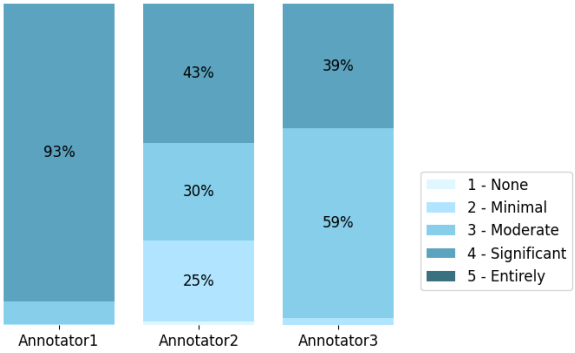

Figure 6: Assessment of standard orthography usage on the control set.

dev and 0.66 for *devtest*, suggesting moderate reliability. However, the inter-class correlations for standard orthography usage are lower, with values of 0.35 for *dev* and 0.27 for *devtest*, indicating considerable disagreement among annotators. This discrepancy highlights the ongoing lack of clarity regarding Emakhuwa spelling standards, as further illustrated in Figure 6, which depicts the varying assessments of standard orthography.

	Adequacy		**Orthography**	
	dev	devtest	dev	devtest
ICC	0.67	0.66	0.35	0.27
CI	[0.63, 0.71]	[0.62, 0.7]	[0.27, 0.42]	[0.18, 0.35]

Table 2: Intraclass Correlation Coefficient (ICC) and Confidence Interval (CI) Results for Adequacy and Orthography usage annotation.

4.5 Dataset Collected

Table 3 presents the statistics for the *devtest* and *dev* sets resulting from the completion of the translation tasks. The *dev* set comprises 997 sentence pairs, while the *devtest* set contains 1,012 sentence pairs. A sample of the dataset is displayed in Table 7 in the appendix.

	dev	**devtest**
#ref.	997	1,012
#ref. words	18,673	21,011
#post-edited refs	1,848	1,889

Table 3: Statitics for the resulting dataset sets

5 Experiments

This section describes the experiment involving training neural MT models using the training sets described below. Then, we performed a comprehensive benchmark evaluation using the evaluation sets introduced in this study.

5.1 Training Data

To train the models, we used the data outlined below:

- Ali et al. (2021) dataset: This subset comprises parallel data in Portuguese and Emakhuwa from different sources, including online texts from the Jehovah's Witness, the African Story Book websites, and Optical Character Recognition (OCR) extracted texts. The corpus contains diverse writing styles, spelling styles, and genres.

- Parallel News: This subset consists of news articles translated from Portuguese into Emakhuwa.

The dataset includes around 63k training parallel sentences and 964 validation parallel sentences, spanning a range of topics (see Table 4), where a significant portion of the data comes from the religious domain, mainly consisting of translations of biblical texts.

	Sentences		**Tokens**	
Source	Train	Dev	*pt*	*vmw*
Ali et al. (2021)	46,454	399	1,104,279	951,520
News	17,403	565	596,066	541,598
Total	63,857	964	1,700,345	1,493,118

Table 4: Training and Validation data statistics

		dev				*devtest*			
		Single Ref.		**Multi Ref.**		**Single Ref.**		**Multi Ref.**	
		BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
transformer-base									
Baseline	pt→vmw	3.7	30.67	$3.95_{(+0.25)}$	$31.32_{(+0.65)}$	3.27	29.23	$3.57_{(+0.3)}$	$29.84_{(+0.61)}$
	vmw→pt	4.36	25.48	-	-	2.93	23.96	-	-
Multilingual Language Models									
afri-byT5	pt→vmw	10.32	41.88	$10.81_{(+0.49)}$	$42.64_{(+0.76)}$	7.03	35.87	$7.73_{(+0.7)}$	$36.72_{(+0.85)}$
	vmw→pt	**22.45**	**47.31**	-	-	13.74	**37.78**	-	-
afri-mT5	pt→vmw	5.66	35.37	$5.96_{(+0.3)}$	$36.01_{(+0.64)}$	4.7	32.7	$5.06_{(+0.36)}$	$33.25_{(+0.55)}$
	vmw→pt	12.12	38.18	-	-	7.39	32.92	-	-
byT5	pt→vmw	**10.66**	**42.37**	$\mathbf{11.2}_{(+0.54)}$	$\mathbf{43.16}_{(+0.79)}$	**7.49**	**36.33**	$\mathbf{8.13}_{(+0.64)}$	$\mathbf{37.15}_{(+0.82)}$
	vmw→pt	22.24	47.01	-	-	**14.1**	37.75	-	-
mT0	pt→vmw	5.52	30.33	$5.76_{(+0.24)}$	$30.9_{(+0.57)}$	4.69	27.89	$5.02_{(+0.33)}$	$28.36_{(+0.47)}$
	vmw→pt	17.46	38.92	-	-	10.63	32.69	-	-
mT5	pt→vmw	6.76	34.09	$7.18_{(+0.42)}$	$34.8_{(+0.71)}$	5.67	31.67	$6.06_{(+0.39)}$	$32.23_{(+0.56)}$
	vmw→pt	15.42	37.58	-	-	9.65	32.22	-	-
Many-to-Many Multilingual Translation Language Models									
M2M100	pt→vmw	8.25	39.22	$8.79_{(+0.54)}$	$40.14_{(+0.92)}$	6.92	36.33	$7.57_{(+0.65)}$	$37.19_{(+0.86)}$
	vmw→pt	21.08	45.31	-	-	13.67	37.46	-	-
NLLB	pt → vmw	8.19	41.44	$8.74_{(+0.54)}$	$42.32_{(+0.88)}$	5.88	36.13	$6.34_{(+0.46)}$	$37.01_{(+0.88)}$
	vmw→pt	17.41	42.88	-	-	10.35	35.05	-	-

Table 5: BLEU and chrF scores for various models on *dev* and *devtest* splits, for single and multiple references

5.2 Setup

We trained MT models in both directions, *pt-vmw* (Portuguese to Emakhuwa) and *vmw-pt* (Emakhuwa to Portuguese), using two approaches: training a vanilla transformer model and fine-tuning existing multilingual language models.

Training We adopt the transformer architecture (Vaswani et al., 2017), implemented through the OpenNMT toolkit (Klein et al., 2017). The model consists of an encoder and decoder comprising 6 layers, 8 heads, and 512 hidden units in the feed-forward network. We used an embedding size of 512 dimensions for both source and target words and a batch size of 32. We applied layer normalization and added dropout with a 0.1 probability to the embedding and transformer layers. Additionally, the Adam optimizer (Kingma and Ba, 2014) was used, and a learning rate of 0.0002. The checkpoints were saved every 1000 updates. We preprocess the input, applying the Byte Pair Encoding subword segmentation.

Fine-tuning Multilingual Models Multilingual language models are one of the most prominent approaches to low-resource languages nowadays since it enables knowledge transfer among related languages, making cross-lingual transfer and zero-shot learning possible.

In our experiments, we fine-tuned various multilingual language models that are well-established in the literature, namely: mT5 (Xue et al., 2021), byT5 (Xue et al., 2022), and the multilingual translation models M2M-100 (Fan et al., 2021) and NLLB (NLLBTeam et al., 2024). Specifically, we use mT5-base (580M parameters), byT5-base (580M parameters), M2M-100 (418M parameters), and NLLB-200's distilled variant (600M parameters). Additionally, we also fine-tuned the African-centric language models, namely, AfribyT5 (580M parameters) and AfrimT5 (580M parameters) by Adelani et al., 2022.

5.3 Evaluation

To assess the systems' performance, we used the SacreBLEU toolkit (Post, 2018) to compute the BLEU (Papineni et al., 2002) and ChrF scores (Popović, 2015).

6 Results and Discussion

Results are presented in Table 5. Our baseline results, derived from a vanilla transformer-base model, set a foundational performance benchmark. On the *devtest* set, the baseline model achieved a BLEU score of 2.93 and a ChrF score of 23.96 for the *vmw* → *pt* translation direction. These modest scores underscore the limitations of the vanilla

transformer-base model in handling the complexities of translation tasks involving low-resource languages like Emakhuwa.

However, introducing multilingual language models enhanced translation performance, particularly in the *vmw* → *pt* direction. Among these, models based on byT5 demonstrated superior performance. For instance, the fine-tuned byT5 model achieved a BLEU score of 14.1 and a ChrF score of 37.75 on the *devtest* set, which marks a substantial improvement over the baseline. This highlights the advantage of leveraging tokenization-free approaches, which are better suited for handling the morphological richness and orthographic variations characteristic of Emakhuwa.

Across Table 5, our results show that while BLEU scores remained relatively low in the *pt* → *vmw* translation direction, ChrF were consistently higher. This discrepancy between BLEU and ChrF scores suggests that BLEU may be disproportionately penalizing spelling variations and minor orthographic differences, which are more prevalent in Emakhuwa translations. ChrF, on the other hand, being more sensitive to character-level n-grams, captures better the quality of translations. Nevertheless, further studies need to be done to assess the correlation of these automatic metrics with human evaluations.

Using multiple references Notably, using multiple references improved scores for both BLEU and ChrF across all models. Specifically, BLEU scores increased by +0.24 to +0.54 on the *dev* set and by +0.3 on the *devtest* set.

7 Conclusion

In conclusion, this study expanded the FLORES+ evaluation set to include Emakhuwa, a low-resource language spoken in Mozambique. By translating the *dev* and *devtest* sets from Portuguese to Emakhuwa. We discussed key challenges such as spelling inconsistencies and loanword adaptations, which are prevalent due to Emakhuwa's underdeveloped spelling standards. Our rigorous methodology, involving translation, post-editing, and validation, ensured high-quality datasets used to benchmark neural MT models. The results indicate that incorporating multiple reference translations can enhance translation quality, particularly in languages with underdeveloped orthographies such as Emakhuwa. The dataset is publicly available, providing a valuable resource for future research in low-resource language MT.

Acknowledgements

This work was financially supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC) as well as supported by the Base (UIDB/00022/2020) and Programmatic (UIDP/00022/2020) projects of the Centre for Linguistics of the University of Porto. Felermino Ali is supported by a PhD studentship (with reference SFRH/BD/151435/2021), funded by Fundação para a Ciência e a Tecnologia (FCT).

We sincerely thank the Lacuna Fund for their generous sponsorship, which made the creation of this dataset possible. Our gratitude also goes to the Translation Team for their dedication and hard work on this project.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. Towards a parallel corpus

of portuguese and the bantu language emakhuwa of mozambique.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Detecting loanwords in emakhuwa: An extremely low-resource Bantu language exhibiting significant borrowing from Portuguese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4750–4759, Torino, Italia. ELRA and ICCL.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory Conde, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine translation for nko: Tools, corpora, and baseline results. In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

R. de Moçambique E.P. 2016. Glossários de conceitos políticos, desportivos e sociais (português-línguas moçambicanas). Retrieved from `http://197.249.65.29/moodle/file.php/1/Glosario_RMe.pdf`.

Armindo Ngunga and Osvaldo Faquir. 2014. *Padronização da Ortografia de Línguas Moçambicanas: Relatório do VI Seminário*. Centro de Estudos das Línguas Moçambicanas.

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

NLLBTeam, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, Yao Lu, and Pontus Stenetorp. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Name	Role	Tasks	Expertise	Alias
Araibo Suhamihe	Translator	Translate *devtest*, Revise *dev*	Professional experience	Translator1
Salustiano Eurico Ramos	Translator	Translate *dev*, Revise *devtest*	Professional experience	Translator2
Gito Anastácio Anastácio	Evaluator	Evaluate and post-edit *devtest* / *dev*	Professional experience	Annotator1
Júlio José Paulo	Evaluator	Evaluate and post-edit *devtest* / *dev*	Professional experience	Annotator2
Vasco André António	Evaluator	Evaluate and post-edit *devtest* / *dev*	Professional experience	Annotator3

Table 6: Translation Team

Figure 7: Annotation Tool User Interface.

pt → vmw		
Source	*pt*	**A camada é mais fina debaixo dos mares e mais espessa abaixo das montanhas**.
Translation	*en*	*It is thinner under the maria and thicker under the highlands.*
References (*vmw*)	A	Mpattapatthaaya tiwoyeva vathi wa mphareya ni yowoneya vathi wa miyaako.
	B	Mpattapatthaaya ti'yottettheeya othi wa iphareya ni yookhoomala vathi wa miyaako.
Systems (*vmw*)	**baseline**	Khalai atthu yahikhotta vathi-va, khukelela vasulu vaya.
	afri-byT5	Okhala wira okathi wa okathi ole ti wootepexa ottuli wa iphareya ni otepexa ottuli wa miyako.
	afri-mT5	Nthowa nenlo ninkhala ntoko nsuwa ntoko nsuwa ni ninkhala ntoko nsuwa ni ninkhala ntoko nsuwa ni ninkhala ntoko nsuwa.
	byT5	Ekamada eyo yootepa omalela vathi va iphareya ni yootepa omalela vathi va miyaako.
	mT0	Okhala wira ematta eyo enniphwanyaneya ottuli wa maasi, nto ematta eyo enniphwanyaneya ottuli wa maasi.
	mT5	Ekatana eyo ti yootepa otthuneya ovikana maasi ni yootepa otthuneya ovikana maasi.
	M2M100	Ekaaxa ele ti yootepa otthuneya vathi va ephareya ni yootepa otthuneya vathi va mwaako.
	NLLB	Mukattelo ti woorekama vathi vathi wa ophareya ni wootepa maasi vathi wa miyaako.

vmw → pt		
Source	*vmw*	**Mpattapatthaaya tiwoyeva vathi wa mphareya ni yowoneya vathi wa miyaako.**
References	*pt*	A camada é mais fina debaixo dos mares e mais espessa abaixo das montanhas.
Systems (*pt*)	**baseline**	A sua <unk> ainda é a propriedade que existe no <unk> sistema de coisas <unk>.
	afri-byT5	A sua aliança é pequena sobre o mar e visível das montanhas.
	afri-mT5	A sua vantagem é pequena sobre o mar e pequena sobre os oceanos.
	byT5	O amigo é pequeno sobre o mar e visível sobre os montes.
	mT0	O companheiro é pequeno na água e pequeno na água.
	mT5	O seu amigo é pequeno na água e pequeno na água.
	M2M100	A arca é pequena debaixo do mar e visível debaixo das montanhas.
	NLLB	A bacia é barata no fundo do mar e muito clara no fundo das margens.

pt → vmw		
Source	*pt*	**Todos os cidadãos da cidade do Vaticano são católicos romanos.**
Translation	*en*	*All citizens of Vatican City are Roman Catholic.*
References (*vmw*)	A	Atthu ootheene opooma wo Vatikaanu anatiini a ekirixitawu ya katolika.
	B	Atthu ootheene opooma ya oVatikaanu anatiini a ekirixitawu katolika.
Systems (*vmw*)	**baseline**	Anammuttettheni otheene a epooma ya Vatoolika aari aRoma.
	afri-byT5	Atthu otheene a epooma ya oVaticano ti makatooliku a oRoma.
	afri-mT5	Otheene a epooma ya Vatikaano ti maKatoliko a oRoma. Otheene atthu otheene a epooma ya Vatikaano ti maKatoliko romano.
	byT5	Atthu otheene a epooma ya oVatikano ti makatooliku a oRoma.
	mT0	Atthu otheene a epooma ya oVaticano ti maKatolika a oRoma.
	mT5	Atthu otheene a epooma ya oVaticano ari maKristau a oRoma.
	M2M100	Atthu otheene a epooma ya oVaticano ari maKatoolika a oRoma.
	NLLB	Atthu ootheene anikhala epooma ya Vatikaano ti makatooliku a orooma.

vmw → pt		
Source	*vmw*	**Atthu ootheene opooma wo Vatikaanu anatiini a ekirixitawu ya katolika.**
References	*pt*	Todos os cidadãos da cidade do Vaticano são católicos romanos.
Systems (*pt*)	**baseline**	Todos na cidade do Vaticano apela a terra de <unk>.
	afri-byT5	Toda a população na cidade do Vaticano realiza a religião católica.
	afri-mT5	Todos os cidadãos em Vaticano são religiosos da igreja católica.
	byT5	Toda a população na cidade do Vaticano é religiosa da cristã católica.
	mT0	Todos os cidadãos da cidade de Vaticane são cristãos da igreja católica.
	mT5	Todos os cidadãos na cidade de Vaticano são cristãos católicos.
	M2M100	Todos os cidadãos do Vaticano são cristãos católicos.
	NLLB	Todos na cidade do Vaticano são religiosos católicos.

Table 7: Example of source-reference sentences pairs from *devtest* and outputs from translating source text using models discussed in Section 5.2

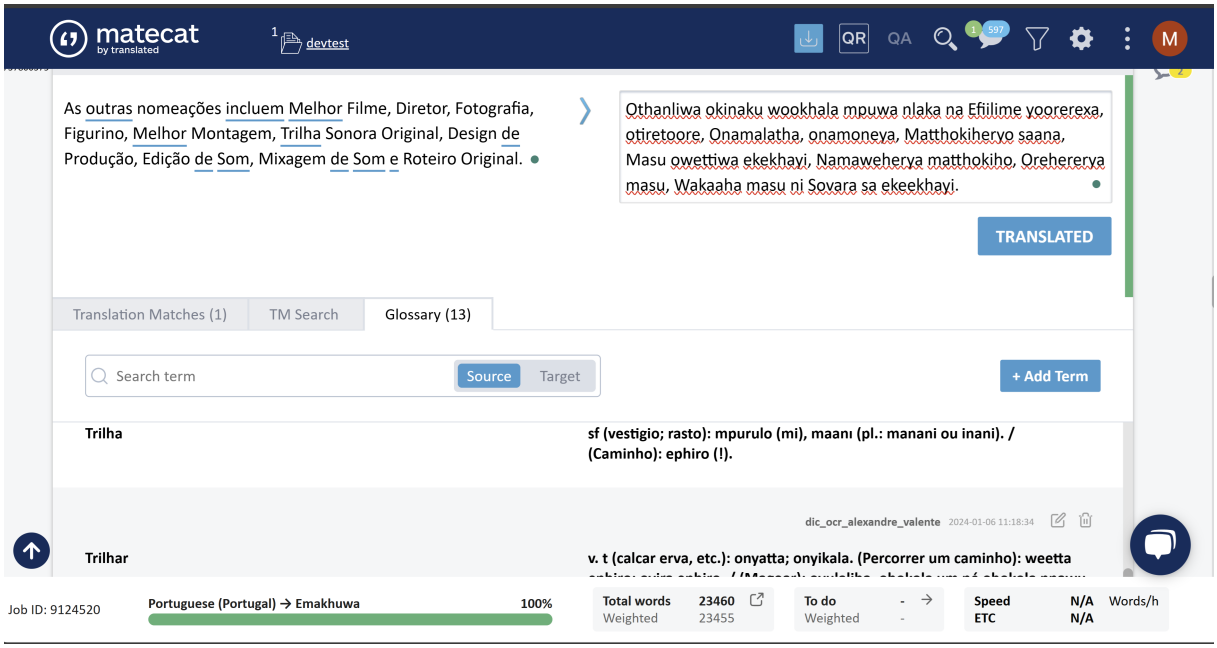

Figure 8: Matecat User Interface

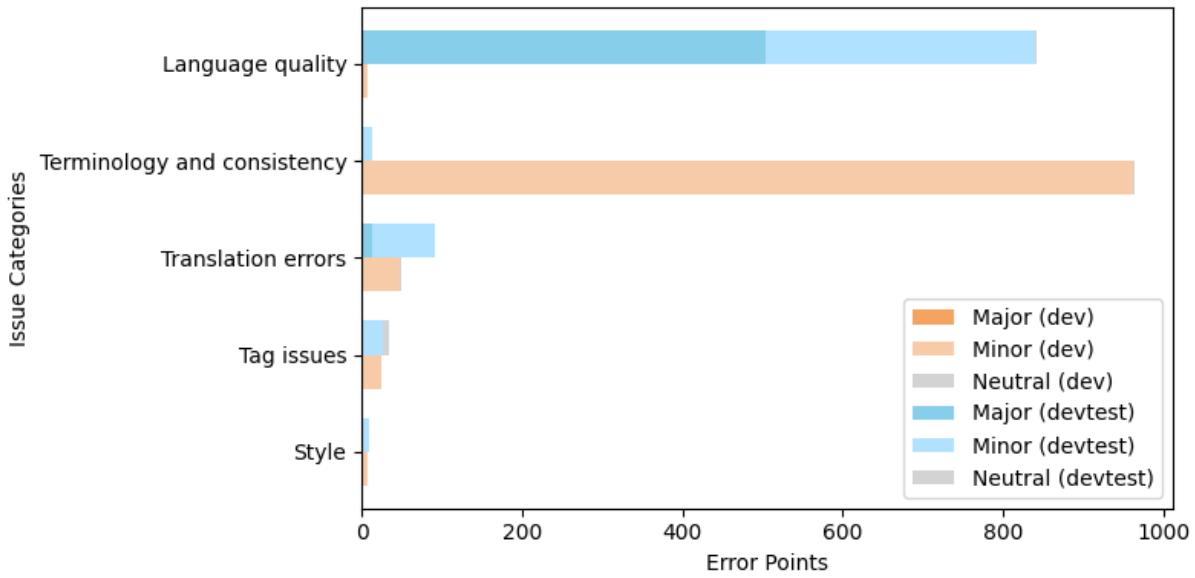

Figure 9: Matecat Quality Report after revision, categorized by the following translation issue typologies: 1) Style (readability, consistent style, and tone); 2) Tag issues (mismatches, whitespaces); 3) Translation errors (mistranslation, additions or omissions); 4) Terminology and translation consistency; 5) Language quality (grammar, punctuation, spelling). The error point count corresponds to the number of segments found with any of the issues described above.

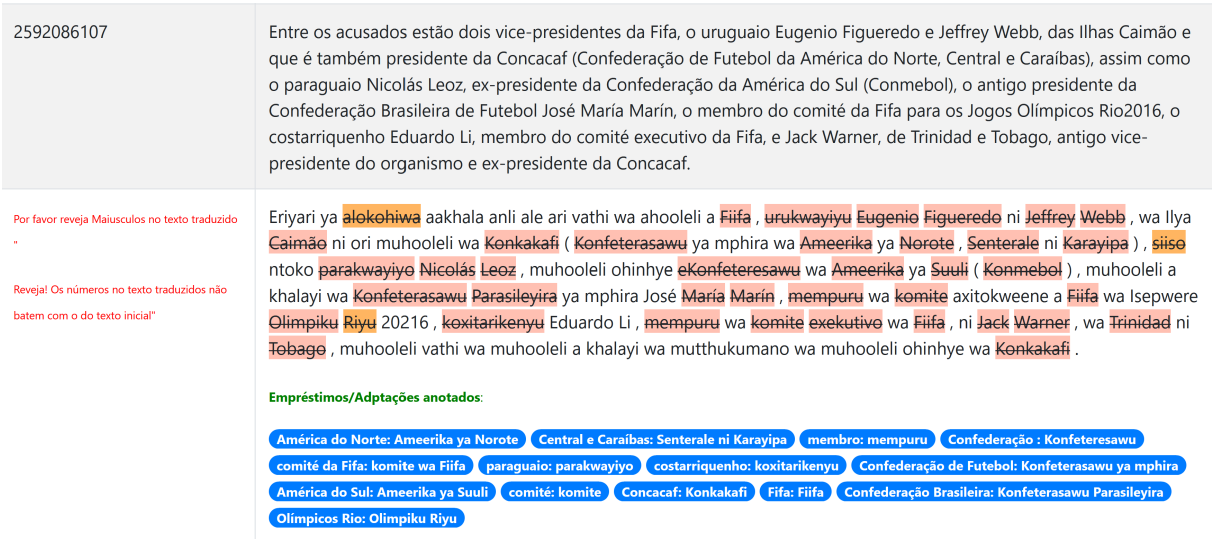

Figure 10: Screenshot of a spelling report. The report is organized into two columns: the first column lists the segment ID along with any potential translation issues (i.e., punctuations, source-target length ratio flag, number mismatch, loanwords not annotated, case mismatch, etc.). The second column displays the source text and its translation. Potential misspellings are highlighted within the translation. In the translation, potential misspellings are highlighted in yellow and red—yellow indicating that suggestions for corrections are available and red indicating that no suggestions exist. Additionally, the report lists all words that translators have annotated as loanwords from Portuguese, using the format *<donor sequence in Portuguese>:<recipient sequence in Emakhuwa>*

Model	Size	Hyperparameters
byT5-base / afri-byT5-base	580M	• Max source length: 200 • Max target length: 200 • Batch size: 8 • Beams: 4
mT5-base / afri-mT5-base	580M	• Max source length: 200 • Max target length: 200 • Batch size: 8 • Beams: 4
mT0	580M	• Max source length: 200 • Max target length: 200 • Batch size: 8 • Beams: 4
NLLB-200-distilled-600M	600M	• Max steps: 60000
M2M100	418M	• Max tokens: 1200 • Layers: 12 • Dropout: 0.3 • Attention dropout: 0.1 • Learning rate: 3e-05 • Max update: 40000 • Emakhuwa was mapped to Swahili (sw)

Table 8: MT Models Configurations