

Closing the Resource Gap for Tigre: A Diaspora-Sourced English–Tigre Parallel Dataset for SMOL

Beshir Ibrahim
BeitTigreAI
bbeshir@gmail.com

Abstract

Tigre, an Afro-Semitic language maintaining strong linguistic continuity with Classical Ge’ez, holds a significant presence in Eritrea as well as across native and intergenerational refugee communities in eastern Sudan. It remains absent from major machine translation benchmarks and resources such as FLORES-200, despite an established orthography and active use in education. As a submission to the WMT26 Open Language Data shared task, we present a community-driven contribution of 11,998 English–Tigre pairs (word-, phrase-, and sentence-level) to the SMOL parallel corpus, written in the standardized Ge’ez-script form of Tigre used in Eritrean education and media. The data were produced through a structured, multi-stage post-editing and cross-validation pipeline involving two diaspora native-speaker translator groups. We document the collection methodology and translator workflow, and validate the data’s usefulness by fine-tuning NLLB-200-3.3B on the contributed pairs, observing about a 4.4× improvement in chrF++ over an untrained baseline for English-to-Tigre translation (4.92 to 21.63), with gains in both translation directions. We release the dataset under a Creative Commons Attribution (CC-BY 4.0) license to support future NLP research on this severely under-resourced language.

1 Introduction

Tigre, an Afro-Semitic language exhibiting deep archaic continuity with Classical Ge’ez, holds a significant presence in Eritrea as well as across native and intergenerational refugee communities in eastern Sudan. Despite possessing an established orthography, active use in education, and a growing community presence on digital platforms, Tigre remains

severely under-served in natural language processing (NLP). Notably, while closely related languages within the Afro-Semitic branch—such as Tigrinya (tir_Ethi)—have been integrated into massively multilingual translation models like Meta’s NLLB-200, Tigre (tig_Ethi) itself is entirely absent from major benchmarks such as FLORES-200 and standard commercial translation systems. This absence creates a persistent digital resource gap for Tigre speakers, hindering the automated creation of educational materials, cross-lingual retrieval, and localized technology tools.

To directly address this resource divide, we present a community-driven, parallel translation initiative that constructs and evaluates the English–Tigre dataset within the SMOL parallel corpus (Anagnostopoulou et al., 2025). Our effort produced 11,998 high-quality English–Tigre pairs through a structured, multi-stage post-editing and cross-validation pipeline involving two distinct native-speaker diaspora organizations. This work documents the dataset’s collection methodology, establishes baseline experimental validations, and is submitted to the WMT26 Open Language Data shared task.

The main contributions of this paper are three-fold:

- 1. Parallel Dataset Contribution:** We release a human-annotated and multi-stage peer-reviewed English–Tigre parallel dataset under a permissive Creative Commons Attribution (CC-BY 4.0) license.
- 2. Quality & Experimental Validation:** We establish benchmark translation experiments in both directions by fine-tuning NLLB-200-3.3B with expanded Tigre vocabulary, demonstrating a signifi-

cant $4.4\times$ increase in chrF++ (English to Tigre) and substantial BLEU gains over non-fine-tuned baselines.

3. **Documentation of Community Workflow:** We outline a reproducible, standard-conformant post-editing and cross-validation methodology tailored for diaspora-led, low-resource language preservation efforts.

2 Language Overview

2.1 Speaker Population and Geographic Distribution

Tigre is commonly estimated to be the second most widely spoken language in Eritrea¹. No census has been conducted in Eritrea since its independence, so speaker numbers are estimates rather than official figures. Speakers are concentrated in the northern, western, and coastal lowlands of Eritrea, as well as adjacent regions of eastern Sudan (Idris, 2015), which also host Eritrean refugee communities settled since the 1960s (Kibreab, 1987). In Eritrea, the language enjoys formal institutional recognition: since the early 1990s, national education policy has allowed children to attend elementary school in their mother tongue, including Tigre, which is taught in the Ge'ez script (Woldemikael, 2003; Asfaha et al., 2009). Implementation has been uneven, with Tigrinya, Arabic, and English dominating many schools (Woldemikael, 2003), but the policy gave Tigre a classroom and literacy tradition in the Ge'ez script, providing a base for digital adoption.²

2.2 Linguistic Classification and Typology

Tigre belongs to the Afro-Semitic (South Semitic) sub-branch of the Afro-Asiatic language family. It is genetically distinct from, though frequently confused with, its geographical neighbor Tigrinya. While both languages descend from Classical Ge'ez and share common structural affinities, they remain mutually unintelligible separate languages. Like Classical Ge'ez, Tigre exhibits

non-concatenative root-and-pattern morphology, synthetic verbal paradigms, and conservative Semitic features such as broken plurals.

Both languages also retain substantial but differing degrees of lexical similarity to their common ancestor, Ge'ez, the liturgical language of the ancient Aksumite Empire. Tigre shows 71% lexical similarity to Ge'ez, reflecting a highly conservative vocabulary, and is considered the closest living language to the extinct Ge'ez (Bender, 1971). Tigrinya shows 68%, having diverged slightly more in its core vocabulary (Bender, 1971).

2.3 Script and Orthography

Tigre is written in the Ge'ez abugida. While the writing system is standardized for formal education and broadcasting in Eritrea, minor orthographic variations persist in digital text. Three sources of variation are particularly relevant to this contribution. First, gemination (consonant length) is phonemic in Tigre—e.g. *kəbər* 'pride' vs. *kəbbər* 'proud, bossy'—but the Ge'ez script has no dedicated diacritic for it, so writers rely on context or informal doubling conventions that vary across texts (Idris, 2015). Second, labialized velar consonants (k^w , q^w , g^w), attested mainly in the 'Algaden and Beni-Āmər dialect areas,³ use a distinct, less common glyph that some digital keyboard layouts do not support, leading typists to substitute the plain velar form instead (Idris, 2015). Third, the mid-central vowel /ə/ is realized with several allophones depending on dialect and phonetic environment, which can lead to inconsistent vowel-sign choices across writers (Idris, 2015).

2.4 Existing Resources and Under-Resourced Status

Historically, digital resources for Tigre have been severely fragmented. Early documentation was limited to descriptive grammars and localized dictionaries. While recent participatory movements have seen Tigre content emerge on platforms like Tatoeba, Wikipedia⁴, and Mozilla Common Voice, the language lacks large-scale, standardized NLP corpora,

¹Readers unfamiliar with Tigre may find useful background and ongoing digitization efforts at <https://beittigreai-resources.static.hf.space/>.

²For the history and implementation of the policy, see Hailemariam et al. (1999) and Woldemikael (2003).

³Transliteration follows the Semitistic convention: ' marks alef and ' marks ayin.

⁴<https://tig.wikipedia.org/>

pre-trained language models, or public evaluation benchmarks.

The Tatoeba platform, a volunteer-contributed collection of mostly short, everyday sentences, hosts a larger parallel Tigre resource than the present dataset: its Tigre portion comprises 51,994 parallel sentence pairs, of which about 59% are paired with Arabic, 20% with English (10,546 pairs), 18% with Swedish, and 3% with German. The Tatoeba sentences are very short, averaging about four words, and typical examples are commands, questions, exclamations, and simple statements. By comparison, the SMOL contribution is English-centered, its sentence-level items average 14–15 English words, and it adds word-level and document-level items (Section 3.3), multi-stage review by two native-speaker groups, and the SMOL format used for massively multilingual training. The two resources are therefore complementary, and because they differ in data type and evaluation setup, we compare them descriptively rather than by translation quality.

2.5 Target Variety: Standardized Written Tigre

The dataset is written in a single variety: standardized written Tigre in the Ge’ez script. As part of its 1990s primary mother-tongue education initiative, the Eritrean government standardized a Ge’ez orthography and grammar baseline for Tigre, providing the language with its first formal state textbooks and classroom instruction. This standard is the variety used in primary school education, state journalism, and national media across Eritrea. The dataset does not encode dialect-specific forms: all text is in the standard orthography, and no per-example dialect labels are provided. Dialectal variation in Tigre is documented in detail by [Idris \(2015\)](#). The language is identified by ISO 639-3 code `tig` with script code `Ethi (tig_Ethi)`; Glottolog code `tigr1270`.

3 Data Collection

The creation of the English–Tigre parallel dataset (SMOL) was organized as a collaborative translation and post-editing effort. Initiated by **BeitTigreAI** (huggingface.co/BeitTigreAI)—an open-source portal dedicated to Tigre language

data and AI models—the initiative brought together diaspora communities to construct high-quality, human-annotated parallel text for downstream machine translation and language technology development.

3.1 Translator Background and Expertise

The translation and quality assurance process involved two distinct diaspora native-speaker groups comprising over 25 contributors in total:

- **Mahber Wlad Hayghet:** A group of more than ten individuals residing in the United States. Its members include many native speakers whose home area is Mansa (also spelled Mensa).
- **Tigre Language Digital Advancement Group (TLDAG):** A team of approximately 15 members based across Europe and the Middle East. Most members received their primary formal education in Eritrea, bringing native fluency in the standardized written Tigre used in official communication and contemporary media.

Both groups worked toward the same written standard (Section 2.5); they differ in location and background, not in target variety.

3.2 Machine Translation Disclosure and Workflow

To seed the initial parallel text, a preliminary translation draft was automatically generated using `BeitTigreAI/tigre-nllb-200-3.3B`, a fine-tuned model hosted on Hugging Face. The base `NLLB-200-3.3B` model predates SMOL and did not support Tigre until its tokenizer and embeddings were extended for the language, and this seed model was not trained on any SMOL sentences. Native speakers then executed a multi-tier human post-editing and quality assurance pipeline:

1. **MT Post-Editing & Initial Translation:** An individual contributor from one of the two groups post-edited the draft output generated by the seed MT model, correcting syntactic errors and lexical inaccuracies.

2. **In-Group Review:** The draft underwent two independent rounds of peer review by other members within the primary author’s group to ensure natural phrasing and accurate semantic alignment.
3. **Cross-Group Quality Assurance (QA):** Once internal consensus was reached, the translated batch was transferred to the secondary group for cross-validation. The receiving team performed a joint review to check conformance to the standard orthography and usage, resolve remaining disagreements, and finalize the entry.

The two groups therefore acted as successive stages of a single pipeline, not as separate translators of parallel versions.

Stage	Process	Responsible Party
1. MT Draft	Automated translation generation	BeitTigreAI/tigre-nllb-200-3.3B
2. Initial Post-Editing	Syntax and lexical correction	Single primary translator
3. In-Group Review	Peer review, semantic alignment	2 internal reviewers (same group)
4. Cross-Group QA	Conformance to standard orthography and usage; final validation	Secondary group joint review

Table 1: Translation and QA pipeline stages.

3.3 Dataset Composition

The 11,998 pairs come from 7,997 source entries across the three SMOL components (Table 2), of which 7,988 are distinct English strings. GATITOS contains 4,000 word- and phrase-level entries. An entry can have several accepted Tigre renderings (1 to 11; 2,267 of the 4,000 entries have more than one), and each rendering is listed as its own pair, which is why 4,000 entries give 8,001 pairs. SmolSent contributes 863 sentences, and SmolDoc contributes 3,134 sentences from 260 documents. Only the 3,997 SmolSent and SmolDoc pairs (about one third of the total) are sentence-level.

The data is two-way parallel and contains a single written variety. Each source entry went

Component	Entries	Pairs	Words En/Tig
GATITOS	4,000	8,001	1.10 / 1.24
SmolSent	863	863	14.22 / 14.14
SmolDoc	3,134	3,134	15.19 / 13.62
Total	7,997	11,998	5.72 / 5.40

Table 2: Composition of the contributed data by SMOL component. Word counts are averages of whitespace-delimited tokens per pair; Tigre counts are computed on the normalized text used for training (Section 4.1).

through the pipeline described above once; for GATITOS entries, the pipeline output may list several accepted renderings.

3.4 Execution, Data Management, and Licensing

The end-to-end execution of the data collection effort spanned approximately nine months. The parallel data was distributed across three Google Sheets documents with restricted, role-based access control. Cross-group sync meetings and joint QA sessions were conducted weekly over Zoom, allowing the teams to discuss edge cases, standardize orthographic conventions in the Ge’ez script, and reach consensus on non-standard vocabulary.

To support future research and open-source NLP development for under-resourced Afro-Semitic languages, the finalized dataset is released under a **Creative Commons Attribution (CC-BY 4.0)** license, permitting free, unrestricted access, modification, and distribution. The data follows the SMOL file format, with one file per component (GATITOS, SmolSent, SmolDoc), and preserves the original Tigre text, including punctuation.

4 Experimental Validation

To validate the quality and usefulness of the contributed data, we fine-tuned the out-of-the-box Meta NLLB-200-3.3B checkpoint with its vocabulary extended to include Tigre (tig_Ethi)—distinct from the already Tigre-adapted BeitTigreAI/tigre-nllb-200-3.3B model used in Section 3.2 to seed the initial translation drafts.

4.1 Setup

Vocabulary and initialization. A SentencePiece model (13,618 pieces) trained on the Tigre side of the training split is merged into a frequency-trimmed NLLB vocabulary, adding 13,282 new tokens for a total of 57,151 pieces, and the embedding matrix is resized; after trimming, the model has about 2.94B parameters. Existing tokens keep their embeddings; new tokens are initialized as the mean embedding of their pieces under the original tokenizer; the new language code `tig_Ethi` is initialized from `tir_Ethi` (Tigrinya). The model has a single Tigre language code and no variety-selection control, since the data contains one variety.

Data and training. The pairs are split at random into 10,799 training and 1,199 validation pairs (seed 42) and used in both directions (English $\leftrightarrow$ Tigre; 21,598 training examples). Stage 1 trains only the shared embeddings (117.7M parameters) for one epoch; stage 2 trains all parameters for three epochs. Both stages use AdamW, learning rate 10^{-4} , batch size 16, 200 warm-up steps, weight decay 10^{-3} , and maximum length 128 tokens. We use the final checkpoint; the validation split is not used for model selection or early stopping. Results come from a single run.

Preprocessing. Tigre text is normalized to Ge’ez letters and whitespace for training and scoring: punctuation, digits, and Latin characters are removed. This changed 4,084 of the 11,998 targets (nearly all SmolSent and SmolDoc sentences). Because deletion inserts no space, in 518 targets (4.3%) punctuation sat directly between letters, so its removal fused neighboring words; Latin letters or digits were also removed from 272 targets. The released files keep the original text (Section 5).

Evaluation. We evaluate on the 1,199 held-out pairs in both directions using chrF++ (word order 2) and BLEU, both computed with sacreBLEU 2.6.0 (Papineni et al., 2002; Popović, 2015, 2017), with default metric and decoding settings. We compare against an untrained baseline: the same vocabulary-extended stock model prior to any Tigre fine-tuning, decoded with `tig_Ethi` as the forced target-language token, isolating the contribu-

Direction	Model	chrF++	BLEU
En→Tig	No fine-tuning	4.92	0.10
En→Tig	Fine-tuned	21.63	6.74
Tig→En	No fine-tuning	10.11	0.19
Tig→En	Fine-tuned	29.62	12.33

Table 3: Translation quality on the 1,199 held-out pairs before (vocabulary-extended NLLB-200-3.3B, no fine-tuning) and after fine-tuning on the contributed data.

tion of the training data itself from the tokenizer/embedding initialization.

4.2 Results

Fine-tuning on the contributed data yields about a $4.4\times$ improvement in English-to-Tigre chrF++ (4.92 to 21.63) and a BLEU gain from near-zero, with gains in the reverse direction as well (10.11 to 29.62 chrF++; Table 3), demonstrating that the dataset provides meaningful, learnable signal for English–Tigre translation despite its modest size. The baseline starts from a very low base, since the model has not seen Tigre text and the `tig_Ethi` embedding is initialized from Tigrinya, so the ratio should be read accordingly.

Table 4 breaks the scores down by SMOL component. Two thirds of the validation set (807 of 1,199 pairs) are single words or short phrases, which score lowest, so the overall figures mostly reflect the mix of word-level and sentence-level items. SmolDoc scores highest (25.50 chrF++ English to Tigre). SmolSent has only 101 validation pairs, so its BLEU is unstable. GATITOS items are scored against a single reference although several renderings may be acceptable, which lowers their scores.

4.3 Limitations

We note an important limitation of this validation: the held-out set (1,199 pairs) is drawn from the same distribution and collection process as the training data, rather than from an independent benchmark, since Tigre does not currently appear in FLORES-200/FLORES+ and no established external benchmark exists for this language pair. Evaluating on held-out portions of one’s own training data is a comparatively weak signal of true generalization and can often tell one relatively little beyond confirming the data is learnable; we

Component	Model	Pairs	En→Tig chrF++ / BLEU	Tig→En chrF++ / BLEU
GATITOS	No fine-tuning	807	2.44 / 2.38	5.53 / 0.03
	Fine-tuned		10.45 / 9.17	17.48 / 17.88
SmolSent	No fine-tuning	101	5.27 / 0.14	11.11 / 0.19
	Fine-tuned		17.16 / 1.62	21.53 / 1.91
SmolDoc	No fine-tuning	291	5.43 / 0.12	10.96 / 0.29
	Fine-tuned		25.50 / 8.62	35.49 / 16.02

Table 4: Scores by SMOL component on the held-out pairs.

present these results accordingly, as a lower-bound demonstration that the dataset carries meaningful signal rather than as a definitive measure of translation quality. The absolute BLEU score is modest, which is expected given the corpus size and limited fine-tuning epochs. The data also spans multiple sources and styles (SMOL combines GATITOS, SmolSent, and SmolDoc) rather than a single genre, which may further affect metric stability.

The split is random at pair level, so training and validation overlap. 605 of the 1,199 validation English sources (50.5%) also occur in training, almost all of them GATITOS words (604 of 807) whose other Tigre renderings are in training; only one validation pair is an exact duplicate of a training pair. All 164 SmolDoc documents in the validation set also contribute sentences to training. On the 594 validation pairs whose English source is not in training, the fine-tuned model scores 23.07 chrF++ and 6.98 BLEU English to Tigre, close to the overall 21.63 and 6.74, but this subset contains proportionally fewer GATITOS words, so the comparison reflects composition as much as overlap. Results are from a single run with one hyperparameter setting, and the normalization described above (punctuation and digits removed) means the reported scores do not measure punctuation or numerals. We see establishing an independent benchmark for Tigre, potentially via a FLORES+ contribution, as important future work enabled by this dataset.

5 Data Sample

Table 5 (last page) shows representative GATITOS and SmolSent pairs in their released form, with punctuation. Where a GATITOS entry has several renderings, only one is shown.

6 Conclusion

This SMOL contribution is one of several ongoing language-technology initiatives undertaken by the Tigre diaspora community, alongside a separately underway effort to translate the FLORES evaluation set into Tigre, which we expect will provide the independent benchmark this work’s evaluation currently lacks.

References

- Aris Anagnostopoulos, Naman Goyal, Jason Riesa, et al. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). *arXiv preprint arXiv:2502.12301*.
- Yonas Mesfun Asfaha, Jeanne Kurvers, and Sjaak Kroon. 2009. [Grain size in script and teaching: Literacy acquisition in Ge’ez and Latin](#). *Applied Psycholinguistics*, 30(4):709–724.
- M. Lionel Bender. 1971. The languages of Ethiopia: A new lexicostatistic classification and some problems of diffusion. *Anthropological Linguistics*, 13(5):165–288.
- Naman Goyal, Jason Riesa, et al. 2023. [Bilex Rx: Lexical data augmentation for massively multilingual machine translation](#). *arXiv preprint arXiv:2303.15265*.
- Chefena Hailemariam, Sjaak Kroon, and Joel Walters. 1999. [Multilingualism and nation building: Language and education in Eritrea](#). *Journal of Multilingual and Multicultural Development*, 20(6):475–493.
- Saleh Mahmud Idris. 2015. *A Comparative Study of the Tigre Dialects*, volume 18 of *Semitica et Semiohamitica Berolinensia*. Shaker Verlag, Aachen.
- Gaim Kibreab. 1987. *Refugees and Development in Africa: The Case of Eritrea*. Red Sea Press, Trenton, NJ.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, et al. 2022. [No language left behind: Scaling human-centered machine translation](#). *Nature*, 630:883–891.
- 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.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Tekle M. Woldemikael. 2003. [Language, education, and public policy in Eritrea](#). *African Studies Review*, 46(1):117–136.

English	Tigre	Source
“what are you doing”	“ሚ ትወዴ ህሌከ?”	GATITOS
“Dear Customer”	“ሕሹም ዐሚል”	GATITOS
“I miss you”	“ሰፍለልኮኪ”	GATITOS
“They accomplished their goals and created a legacy of self-determination to which others may aspire.”	“ሀደፎም እንዴ ሐቀቀው: እግል ብዕዳም ሐቅ ተቅረር መሲር ለትናይት አሰር ሐደገው::”	SmolSent
“Just to clarify, though, assertive is not the same as aggressive or arrogant.”	“በስ እብ አኪደት ወድሖት ክራህየት አው ክብር አክልሕድ ይኮነኒ::”	SmolSent
“Some might prefer early morning while others prefer late night.”	“ሰሮም ግውሐት ጽብሕ ምድር እግል ልሕረው ገድሮ እት ህለው ለብዕዳም ላኪን ወረብ ማሲ ምድር እግል ልሕረው ገድሮ::”	SmolSent
“Peace brings along its buddies joy, strength, hope, and faith.”	“ሰላም ምስለ መልህየት: ፈርሓት: ቅወት (ሐለት): አመል ወኤማን ተአመጽእ::”	SmolSent
“They should look for those things that inspire dedication, loyalty, honor and dignity.”	“እግል ሎሆም ሓጃት: መትባልሖት ነፍስ አማነት: ሸረፍ: ወከራመት ለክልቅ እግል ልሕዘው ወጀቦም::”	SmolSent

Table 5: Example English–Tigre pairs from the contributed dataset.