Correcting the Tamazight Portions of FLORES+ and OLDI Seed Datasets

Alp Öktem
Col·lectivaT
alp@collectivat.cat

Mohamed Aymane Farhi
Tamazight NLP
aymenfarhi.25@gmail.com

Brahim Essaidi
Freelance Translator
essaidib2@gmail.com

Naceur Jabouja
Freelance Translator
contact@njtranslations.com

Farida Boudichat
Awal Team
awal@collectivat.cat

Abstract

We present the manual correction of the Tamazight portions of the FLORES+ and OLDI Seed datasets to improve the quality of open machine translation resources for the language. These widely used reference corpora contained numerous issues, including mistranslations, orthographic inconsistencies, overuse of loanwords, and non-standard transliterations. Overall, 36% of FLORES+ and 40% of Seed sentences were corrected by expert linguists, with average token divergence of 19% and 25% among changed items. Evaluation of multiple MT systems, including NLLB models and commercial LLM services, showed consistent gains in automated evaluation metrics when using the corrected data. Fine-tuning NLLB-600M on the revised Seed corpus yielded improvements of +6.05 chrF (en→zgh) and +2.32 (zgh→en), outperforming larger parameter models and LLM providers in en→zgh direction.

1 Introduction

High-quality parallel data is a cornerstone for the development of robust machine translation (MT) systems, particularly for low-resource languages where each sentence can significantly impact model performance. For Tamazight—spoken by over 40 million people across North Africa and the diaspora—parallel corpora are scarce, and widely used datasets such as FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024) and OLDI Seed (Maillard et al., 2023) play a significant role in enabling MT research and evaluation.

The Tamazight portions of FLORES+[1] and OLDI Seed[2] contain Standard Moroccan Tamazight (ISO 639-3: *zgh*), the standardized variety developed by the Royal Institute of Amazigh Culture (IRCAM) (Boukous, 2014) for education and official use since 2001. However, our initial inspection revealed substantial issues: mistranslations, orthographic inconsistencies, malformed or unnecessary loanwords, non-standard transliterations, and occasional semantic inaccuracies. Some of these errors echo broader findings about low-resource language datasets, where insufficient quality control leads to degraded MT outputs and downstream application failures (Kreutzer et al., 2022). Similar issues have been observed for other languages in FLORES+, prompting targeted correction efforts such as those for Hausa, isiZulu, Northern Sotho, and Xitsonga (Abdulmumin et al., 2024). In fact, when FLORES-200 was first released, the Tamazight data was mislabeled as Central Atlas Tamazight (*tzm*) and only corrected following community feedback to its correct code *zgh*.

As part of the Awal project (Öktem and Boudichat, 2025), which develops open-source MT and speech technologies for Tamazight and coordinates community data creation, we undertook a systematic manual correction of the FLORES+ dev and devtest sets (997/1,012 sentences) and the OLDI Seed corpus (6,193 sentences). Corrections were performed by expert linguists using authoritative IRCAM lexicographic and grammatical resources.

The motivation for this work is straightforward: to ensure these widely used benchmark and seed datasets truly reflect Standard Moroccan Tamazight, so that MT systems trained or evaluated on them can produce translations that are both accurate and culturally appropriate. We follow the approach of Abdulmumin et al. (2024) for quantifying the extent of changes. We then evaluate multiple state-of-the-art open-source and commercial MT systems (including LLMs) on both the original and corrected FLORES+ devtest data for English↔Tamazight, and fine-tune existing NLLB

[1] https://huggingface.co/datasets/openlanguagedata/flores_plus

[2] https://huggingface.co/datasets/openlanguagedata/oldi_seed

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 1072–1080
November 8–9, 2025.

models with the updated Seed data to assess downstream impact.

2 Background

2.1 Tamazight and NLP

Tamazight, also known as Amazigh or Berber[3], is part of the Afro-Asiatic language family and is spoken by over 40 million people across a vast area of North Africa and diaspora communities worldwide (Lafkioui, 2018). It is an official language in Morocco since 2011, Algeria since 2016, Libya since 2017, and Mali since 2023. The standardized form, Standard Moroccan Tamazight (ISO 639-3: *zgh*), was developed by IRCAM from 2001 onwards, drawing on features of the main Moroccan varieties—Tachelhit (*shi*), Tarifit (*rif*), and Central Atlas Tamazight (*tzm*)—and other variants such as Touareg (*tmh*) (Boukous, 2014). As part of this process, IRCAM adopted Tifinaghe-IRCAM (also known as Neo-Tifinagh) in 2003 as the official script, replacing earlier informal use of Latin and Arabic scripts, and providing a phonologically accurate and standardized writing system (Soulaimani, 2016; Ataa-Allah and Boulaknadel, 2012).

The language exhibits considerable dialectal diversity across Morocco, Algeria, and other regions, a reflection of its vast geographic spread. Orthographic variation persists despite standardization, with Latin and Arabic scripts still used informally alongside Neo-Tifinagh. Historical marginalization—shaped by colonization and Arabization—has reduced intergenerational transmission in some areas and within diaspora communities, though Amazigh cultural movements such as the Amazigh Spring played a key role in securing recognition and institutional support (Roque, 2009; CIEMEN and Casa Amaziga de Catalunya, 2019).

From an NLP perspective, Tamazight poses challenges due to its rich morphology, complex orthography, and high dialectal variation. Inflectional and derivational processes, coupled with script diversity, make computational processing for tasks like tokenization, POS tagging, and MT more difficult (Ataa-Allah and Boulaknadel, 2012). Until recently, it remained underrepresented online, with limited user-generated content in the standard form. Notably, the Tamazight Wikipedia was only launched in 2023, marking a significant milestone in its digital presence.

Several important language resources and tools have been released in recent years. IRCAM has developed a Standard Tamazight Corpus (Boulaknadel and Ataa-Allah, 2013), a morphosyntactically annotated corpus (Amri et al., 2017), and tools such as an Amazigh verb conjugator (Ataa-Allah and Boulaknadel, 2014), a concordancer, and a Tifinagh-adapted search engine (Ataa-Allah and Boulaknadel, 2012). More recent research includes Tamazight word embeddings trained on web-collected corpora (Faouzi et al., 2023).

In the MT domain, Tamazight is one of the languages listed within the 200 languages of the *No Language Left Behind (NLLB)* project (NLLB Team et al., 2024), which relies on the FLORES training and evaluation sets (Goyal et al., 2022). Other relevant multilingual datasets that include Tamazight are SIB-200 (Adelani et al., 2024), MADLAD (Kudugunta et al., 2023), and GlotCC/GlotLID (Kargaran et al., 2023, 2025). However, as has been noted in recent research, the development of language technologies for underrepresented languages often follows a top-down approach led by large institutions or technology companies with little input from speaker communities (Moshagen et al., 2024; Bird, 2020; Schwartz, 2022). This dynamic risks misrepresenting languages through technologies built without meaningful community participation, and can perpetuate harm by commodifying indigenous knowledge and sidelining local authorities (Bird, 2020). For marginalized languages, the quality of training data is especially critical: inaccuracies can distort their digital representation and propagate errors through AI systems (Kreutzer et al., 2022; Lau et al., 2025). These issues are not unique to Tamazight—audits of widely used multilingual resources have shown that, for many of the languages they "cover," data quality and representativeness remain poor, creating an illusion of coverage while delivering limited practical usability (Lau et al., 2025).

2.2 Awal initiative

Crowdsourcing initiatives have emerged as a viable alternative to the top-down approaches that dominate language technology development for underrepresented languages. Grassroots efforts such as *Masakhane* (Nekoto et al., 2020), *NaijaVoices*

[3]We include the term "Berber" as it is commonly known in the global north but avoid its usage as it is often considered a pejorative term by the Imazighen (Amazigh people).

(Emezue et al., 2025), and *PARME* (Ahmadi et al., 2025) demonstrate that participatory models can produce data that is both higher quality and better aligned with community norms.

The Awal project (Öktem and Boudichat, 2025) follows this participatory model to create open-source MT and speech resources for Tamazight. Launched in 2024, its main platform, `https://awaldigital.org`, facilitates the translation of sentences from or into Tamazight, covering multiple dialects and scripts. Contributions come from volunteers' own translations or from Creative Commons–licensed material, which can also be post-edited from automatic translations produced by the NLLB engine. The platform integrates a peer-validation feature, where each sentence requires two independent approvals before entering the validated corpus. In addition to text, Awal collects Tamazight speech data through Mozilla Common Voice (Ardila et al., 2020), which is likewise validated via community review. All parallel and monolingual text is openly shared through the project's Hugging Face repository[4], while the voice data is released via Common Voice[5].

Awal's approach combines community-driven data collection with curated dataset creation by professional linguists, as in the present work correcting the Tamazight portions of FLORES+ and OLDI Seed.

3 Methodology

3.1 Correction Workflow

We corrected the Tamazight side of the FLORES+ dev (997 sentences), FLORES+ devtest (1,012 sentences), and OLDI Seed (6,193 sentences) datasets, obtained from the official OLDI Hugging Face repositories.

All sentences were exported into spreadsheets to enable structured, sentence-by-sentence review. The FLORES+ dev and devtest sets were revised in full in two iterations by two linguists. The OLDI Seed dataset was divided into batches of 1,000 sentences and distributed among three professional Tamazight translators, with allocation based on their availability and delivery capacity. To ensure quality control, each linguist's work was spot-checked through random sampling by another linguist.

[4] `https://hf.co/datasets/collectivat/amazic`
[5] `https://commonvoice.mozilla.org`

For the FLORES+ splits, the spreadsheet included the English source sentence, the original Tamazight sentence, and a column for the corrected version. For the OLDI Seed dataset, the spreadsheet additionally included the Arabic translation from the original resource, allowing the translators to use both English and Arabic as context. In all cases, if a sentence required no changes, the original Tamazight sentence was copied into the "corrected" column; if corrections were needed, translators directly post-edited the original, making orthographic, lexical, and syntactic adjustments as appropriate.

Corrections were guided by authoritative lexicographic and grammatical resources from IRCAM and other reference works, including dictionaries (Ameur et al., 2016; Chafik, 1996; Akioud et al., 2022), phonology (Boukous, 2009), and grammar books (Boukhris et al., 2008; El Moujahid, 2022; Laabdelaoui et al., 2012).

3.2 Challenges of Standardization and Dialect Representation

The review process also highlighted the limitations of relying solely on such references for a language that is still undergoing standardisation and must represent multiple dialects. Certain inconsistencies are found even in official resources: for example, masculine nouns are described in the *New Grammar of Amazigh* (Boukhris et al., 2008) as generally beginning with ⴰ, ⵓ, or ⵉ and feminine nouns with ⵜ, yet forms like ⵙⵙⵉⵏⵎⴰ (*ssinima*, 'cinema') or ⵍⴱⴰⵏⴽ (*lbank*, 'bank') appear in IRCAM dictionaries without the expected nominal prefix. Similarly, the rule that a word containing an emphatic consonant should be fully emphatic is not applied consistently across dictionary entries and published texts.

Lexical variation across Tamazight dialects further complicates correction. The verb "to give," for example, appears in IRCAM dictionaries with multiple forms—ⵀⴽ (*fk*), ⴽⵀ (*kf*), ⵓⵛ (*uc*), ⵡⵛ (*wc*), ⵙⵙⵉⵢ (*ssiy*)—without clear indication of which is considered standard. For highly standardised languages, such variation is usually resolved in reference materials, but in Tamazight, the standard form is still being shaped through a combination of institutional policy, community use, and corpus-based practice. This reality means that "correction" work often involves navigating legitimate competing forms rather than simply enforcing a single prescriptive norm.

In addition to these challenges, standardization in Tamazight also involves the ongoing creation of new terms, particularly in scientific and technical domains. This process is not a matter of direct translation but requires applying the language's own morphological strategies for derivation, compounding, and adaptation. As such, parallel data creation and curation go beyond translating sentences: they must be informed by a deep understanding of Tamazight's morphological rules and involve specialists in terminology and linguistic planning. Without this, datasets risk importing external forms rather than contributing to the gradual, community-driven development of a functional standard.

3.3 Error Taxonomy

During the review, we identified and categorized recurrent error types in the original datasets. Below we summarize each type and show representative examples.

3.3.1 Spelling Mistakes. These included violations of standardized Tamazight orthography. Key issues were the improper use of specific characters such as (ⵯ), (ⵯ), and emphatic consonants (ⵅ, ⵇ, ⵜ, ⵚ, ⵄ), which were corrected to align with IRCAM rules. For example, "Open" was corrected from ⵓⵅⵯⵔ to ⵇⵅⵔ, and "Three" from ⴽⵓⵄⵜ to ⴽⵇⵄⵜ. Proper nouns were often transliterated inconsistently, such as "Louis Mayer" which needed to be adjusted from ⵍⵯⵙⵙ ⵔⵄⵛⵓ to ⵍⵍⵙⵙ ⵔⵄⵛⵓ, and "Maria Feodorovna" from ⵔⵄⵓⵙⵄ ⵀⵛⵯⴰⵯⵓⵯⵀⵍⵄ to ⵔⵄⵓⵛⵛⵄ ⵀⵛⵯⴰⵯⵓⵯⵀⵍⵄ. We also found erroneous attachment of pronouns to nouns (e.g. "His film" from ⵍⵀⵙⵍⵎⵙⵙ ⵏⵙ to ⵄⵀⵙⵍⵎ ⵏⵙ), which were separated for clarity and grammatical correctness.

3.3.2 Transliteration Errors. Particular attention was paid to the accurate transliteration of foreign words and names. The most frequent problems involved the inconsistent mapping of the letter *V* (correctly rendered as ⵀ) and *P* (correctly rendered as ⵜ) in foreign words. For example, "Nova Scotia" was corrected from ⵍⵯⵜⵄ ⵙⴽⵯⵜⵛⵄ to ⵍⵯⵀⵄ ⵙⴽⵯⵜⵛⵄ, and "Sveriges radio" from ⵄⵍⵣⵔⵄⵛ ⵍ ⵙⵜⵙⵙⵓⵙⵠⵙ to ⵄⵍⵣⵔⵄⵛ ⵍ ⵙⵀⵙⵙⵓⵙⵠⵙ. Other non-Tamazight sounds were adjusted to their closest Tamazight equivalents following standardized transliteration practices.

Unnecessary Loanwords		
English	**Original**	**Corrected**
Work	ⵍⵅⵏⵔⵜ	ⵜⵄⵔⵯⵓⵙ
Theatre	ⵍⵔⵙⵇⵄⵗ	ⵄⵔⵣⵠⵯⵍ
Candy	ⵍⵀⵄⵍⵙⵏ	ⵜⵄⵔⵣⵙⵜ
Start	ⵜⵏⵯ	ⵙⵙⵍⵜⵙ
Anemia	ⵀⵄⵡⵓ ⵏⵏⵄⵔⵔ	ⵯⵣⵣⵯⵀ

Malformed Loanwords		
English	**Original**	**Corrected**
Cinema	ⵙⵙⵙⵍⵙⵔⵄ	ⵄⵙⵙⵍⵙⵔⵄ
Film	ⵍⵀⵙⵍⵎ	ⵄⵀⵙⵍⵎ
Television	ⵜⵙⵍⵙⵀⵙⵣⵠⵯⵍ	ⵄⵜⵙⵍⵙⵀⵙⵣⵠⵯⵍ
Oxygen	ⵍⵯⴽⵙⵓⵍⵙⵍ	ⵯⴽⵙⵓⵙⵍⵙⵍ
Nitrogen	ⵍⵏⵙⵟⵇⵯⵍⵙⵍ	ⵄⵍⵙⵟⵇⵯⵍⵙⵍ

Table 1: Unnecessary and malformed loanwords detected in the corpus and their corrections.

3.3.3 Unnecessary Loanwords. Many sentences used Arabic or French loanwords where Tamazight equivalents exist. Following IRCAM's prioritization guidelines, we replaced these with native terms when available, giving preference first to Moroccan Tamazight variants, then to other Tamazight varieties (e.g. Touareg, Kabyle), and retaining loanwords only when unavoidable. Examples are shown in the upper portion of Table 1.

3.3.4 Malformed Loanwords. In cases where loanwords were retained, they often failed to follow Tamazight morphological patterns. Corrections ensured that such words received appropriate prefixes (e.g., ⵄ- for masculine nouns) and were adapted phonologically to fit Tamazight norms. Examples are presented in the lower portion of Table 1.

3.3.5 Mistranslations. Some translations did not accurately reflect the source meaning, introducing semantic drift or mistranslated idiomatic expressions. These were corrected to capture the intended sense and to align with Tamazight syntax. Representative examples are presented in Table 2.

4 Change analysis

We assessed the extent of changes made to the Tamazight portions of the FLORES+ dev/devtest and OLDI Seed datasets using token divergence (Abdulmumin et al., 2024), BLEU (Papineni et al.,

English Source	Original Translation	Back-translation	Correct Translation
Even if you're driving through the subtropical rainforest, a few seconds with the doors open while you get inside the vehicle is enough time for mosquitoes to get in the vehicle with you.	ⵍⵓⵅⵅⵓ ⴰⵓ ⵜⵍⴰⴰⵀⵜ ⵅ ⵜⵓⵅⵓⵍⵉⵍ ⵍ ⵑⵍⵣⵓⵇ ⵉⵇⵍⵍⵓⵍ ⵓⵇⵀⵅⵇ, ⴰⵇⵑⵇ ⵍ ⵜⵑⵇⴰⵉⴰⵉⵍ ⴰ ⵜⵣⵏⵍⵍⵉⵍ ⵉⵣⵡⵇⵇⵍ ⵜⵉⵏⵉⵍ ⵓⵅⵍⵇⵑ ⵍ ⵜⵇⵇⵇⵑⵑⵉⵜ ⵉⵅⵓ ⵓⵇ ⵉ ⵍⵓⵀⵉⵀⵓ ⵓⴰ ⵉⴰⵅ ⵉⵇⵇⵑⵇ ⵜⵓⵇⵇⵑⵑⵉⵜ.	Although you are driving in rainy forests like "asbgs", few minutes and the doors open and you are inside the vehicle is fine for mosquitoes to share the vehicle with you.	ⵍⵓⵅⵅⵓ ⴰⵓ ⵜⵍⴰⴰⵀⴰ ⵅ ⵜⵓⵅⵓⵍⵉⵍ ⵍ ⵑⵍⵣⵓⵇ ⵜⵉⵏⵅⵍⵓⵜⵑⵇⵉⵍ, ⵅⵑⵏⵜ ⵅⵇⵓ ⵍ ⵜⵇⵉⵍⵉⵍ ⴰⵇⵑⵇⵍⵉⵍ ⵓⵅⵑⵓ ⵜⵡⵣⵏⴰ ⵜⵍⵏⵉⵍ ⵓⴰ ⵜⵅⵏⵏⴰ ⵜⵓⵏⵇⵑⵑⵉⵜ ⵉ ⵍⵓⵀⵉⵀⵓ ⵅⵏⵓ ⵓⴰ ⵉⴰⵅ ⵉⵇⵇⵑⵇ ⵜⵓⵏⵇⵑⵑⵉⵜ.
The walls and roofs of ice caves can collapse and cracks can get closed.	ⵉⵅⵑⴰⵓⵇ ⴰ ⵉⵀⵇⵓⵍ ⵍ ⵑⵅⵇⵉⵇ ⵉⵡⵉⵎⵓⵡ ⵓⴰ ⵉⵜⵜⵑⵅⵍ ⴰ ⵉⵀⵑⵇⵓⵍⵍ ⵓⴰ ⵡⵏⵍ.	The walls and ice caves can be high and skin fissure will close.	ⵡⵎⵓⵡⵍ ⵉⵅⵑⴰⵓⵇ ⴰ ⵉⵅⵅⵉⵜⵍ ⵍ ⵉⵀⵇⵓⵍ ⵍ ⵑⵅⵇⵉⵇ ⵓⴰ ⵇⴰⵍⵍ ⴰ ⵉⵇⵜⵜⵉⵏⵍ ⵓⴰ ⵡⵏⵍ.
For a few pennies some children will tell you the story.	ⵏⵓⵇ ⵅⵇⵓ ⵍ ⵜⵇⵑⵇⵑⵀⵉⵍ ⴰⵇⵑⵇⵍ, ⵓⴰ ⵓⵅ ⵓⵎⵇⵍ ⵉⵡⵉⵇⵇⵓⵍ ⵜⵓⵍⵀⵑⵇⵜ.	Just some few steps children will tell you the story.	ⵏⵓⵇ ⵅⵇⵓ ⵍ ⵜⵓⵇⵇⵏⵓⵎⵉⵍ ⴰⵇⵑⵇⵍⵉⵍ, ⵇⵓⴰ ⵓⵅ ⵓⵎⵇⵍ ⵉⵡⵉⵇⵇⵓⵍ ⵜⵓⵍⵀⵑⵇⵜ.

Table 2: Examples of mistranslations and their corrections. For each case, we show the original English source sentence, the problematic Tamazight translation from the dataset, a back-translation of that Tamazight into English, and the corrected Tamazight translation.

dataset	#rows	#corr. (%)	% token div.	$BLEU_c$	TER_c	WER_c	CER_c
dev	997	384 (39%)	19.36	81.06	12.22	13.35	5.70
devtest	1012	339 (34%)	19.46	79.85	12.35	13.50	5.48
FLORES+	2009	723 (36%)	19.41	78.26	14.84	17.73	8.83
seed	6193	2490 (40%)	25.01	80.49	12.28	13.42	5.60
ALL	8202	3213 (39%)	23.75	78.72	14.29	16.76	8.10

Table 3: Correction statistics and edit-distance metrics for each dataset. Shows the number of sentences corrected (as percentage of total) and average change metrics computed only on modified sentences. FLORES+ aggregates dev and devtest splits, while COMBINED combines FLORES+ and OLDI Seed datasets.

dataset	>50% (%)	>80% (%)
dev	20 (5.2%)	6 (1.6%)
devtest	26 (7.7%)	3 (0.9%)
seed	316 (12.7%)	62 (2.5%)
FLORES+	46 (6.4%)	9 (1.2%)
ALL	362 (11.3%)	71 (2.2%)

Table 4: Number and percentage of corrected sentences with token divergence greater than 50% and 80%, indicating substantial rewrites.

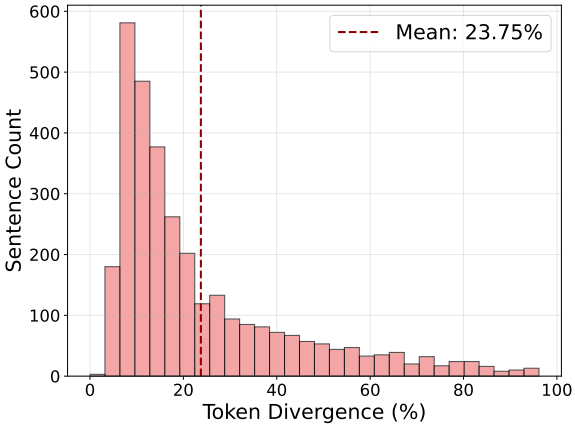

Figure 1: Token divergence distribution across all corrected sentences in the combined FLORES+ and OLDI Seed datasets.

2002), Translation Edit Rate (TER) (Snover et al., 2006), Word Error Rate (WER), and Character Error Rate (CER). Following Abdulmumin et al. (2024), all metrics were computed only on sentences that were modified, avoiding dilution by unchanged items.

Token divergence was calculated as:

$$\text{divergence} = \frac{|T_o - T_c| + |T_c - T_o|}{|T_o \cup T_c|} \quad (1)$$

where T_o is the set of tokens in the original sentence and T_c is the set of tokens in the corrected sentence. This metric measures the proportion of unique tokens that differ between the two versions of a sentence, with 0 indicating identical token sets and 1 indicating no overlap. Reported values are averages over all changed sentences and expressed as percentages.

Table 3 reports, for each dataset, the proportion of sentences corrected and the mean metric scores. In addition to individual datasets, we also report aggregated results for FLORES+ overall

(dev+devtest) and for the combined FLORES+ and Seed datasets. Figure 1 presents the token divergence distribution across all corrected sentences in the combined dataset.

The results show that the proportion of sentences corrected is 36% for FLORES+ overall and 40% for the Seed dataset. Average token divergence values are 19% for FLORES and 25% for Seed, indicating that corrections typically involved localised changes to a minority of tokens in a sentence rather than complete rewrites. This is consistent with the high BLEU scores (≈78–81) and the relatively modest WER (13–18%) and CER (5–8%) values. While substantial rewrites are relatively rare, they are still present in both datasets. Table 4 shows the proportion of corrected sentences whose token divergence exceeded 50% and 80%. For example, 6.4% of corrected sentences in FLORES overall and 12.7% in the Seed dataset had over 50% divergence, indicating a reworking of more than half of their token content.

5 Machine Translation Evaluation

We evaluated English↔Tamazight (zgh) translation quality on the FLORES+ devtest split, comparing results obtained with the original and corrected versions of the dataset. The evaluation covered:

- **Baseline models**: NLLB checkpoints 600M, 1.3B, 1.3B Distilled, 3.3B
- **Fine-tuned models**:
  i) NLLB-600M fine-tuned on the corrected OLDI Seed dataset for 3 epochs.
  ii) NLLB-600M fine-tuned for 1.25 epochs on the corrected OLDI Seed dataset and other parallel data of approximately 50,000 segments sourced from Awal, Tatoeba[6], and Tamazight NLP[7] initiatives.
- **Commercial LLMs**: Gemini Pro 2.5, Claude 3.5, and Claude 3.7[8].

Scores were computed using BLEU, chrF++ (Popović, 2015), and TER (Snover et al., 2006) for both translation directions. Table 5 reports the results, with each cell showing the original dataset score / corrected dataset score format.

Across all models tested, corrections led to consistent improvements in BLEU and chrF++, with corresponding reductions in TER. We see an average chrF increase of 0.14 points in en→zgh, and 0.23 in zgh→en direction among all off-the-shelf models and LLM services. This confirms that noise in the original FLORES+ Tamazight references measurably affected evaluation quality.

Fine-tuning on the corrected OLDI Seed dataset (NLLB-0.6B-FT) yielded substantial gains compared to the same 0.6B baseline: +6.05 chrF and +3.05 BLEU in the en→zgh direction. The fine-tuned model achieved 31.69 chrF, performing better than all other larger models and LLM providers, showing that carefully curated data can yield substantial improvements.

In the zgh→en direction, the fine-tuned 0.6B model improved with +2.32 chrF and +1.89 BLEU from corrections, however it was still outperformed by other models, with Gemini Pro 2.5 achieving the highest chrF score of 44.29.

We also report results with additional parallel data (NLLB-0.6B-FT+): 32.71 chrF and 8.84 BLEU in en→zgh, and 40.66 chrF and 18.29 BLEU in zgh→en.

6 Conclusions

We carried out a systematic manual correction of the Tamazight portions of the FLORES+ dev/devtest and OLDI Seed datasets, resolving orthographic inconsistencies, mistranslations, and problematic loanwords. These corrections, performed by professional linguists with reference to authoritative resources, affected a substantial share of sentences and resulted in more accurate benchmarks for MT evaluation.

Across all off-the-shelf models and commercial LLMs tested, the corrected data yielded consistent improvements in automatic evaluation metrics. Fine-tuning on the corrected OLDI Seed dataset further demonstrated the impact of these revisions: the NLLB-600M model trained on the revised data outperformed larger parameter models and LLM providers in en→zgh direction, and recorded improvements in zgh→en direction.

Beyond the experiments, our work underlines broader challenges of dataset creation for a language still undergoing standardization, where legit-

[6] https://tatoeba.org

[7] https://huggingface.co/Tamazight-NLP

[8] We also intended to evaluate Google Translate; however, its API does not currently support Tamazight, making programmatic evaluation impractical, so it was excluded from this study.

Model	BLEU ↑		chrF++ ↑		TER ↓	
	en-zgh	zgh-en	en-zgh	zgh-en	en-zgh	zgh-en
NLLB-0.6B	4.86/4.96	14.84/15.3	25.52/25.64	37.1/37.38	94.97/94.77	77.27/76.06
NLLB-0.6B-FT	7.8/8.01	17.13/17.19	31.46/31.69	39.51/39.7	83.72/83.32	73/73.01
NLLB-0.6B-FT+	8.64/8.84	17.97/18.29	32.5/32.71	40.27/40.66	82.17/81.87	71.72/71.18
NLLB-0.6B-FT+, S=2	9.12/9.38	18.29/18.58	33.26/33.52	40.45/40.87	81.38/**80.97**	70.71/70.45
NLLB-0.6B-FT+, S=8	9.42/**9.69**	18.37/18.72	33.79/**34.05**	40.55/40.91	81.92/81.49	70.63/70.38
NLLB-1.3B Dist.	6.84/7.02	18.16/18.32	28.93/29.07	39.73/39.97	86.24/85.99	72.05/71.68
NLLB-1.3B	6.34/6.37	17.23/17.35	27.35/27.44	39.11/39.33	89.75/89.59	73.28/72.85
NLLB-3.3B	7.52/7.68	17.72/18.05	29.7/29.8	39.43/39.8	83.52/83.42	72.81/72.14
Gemini Pro 2.5	4.7/4.83	21.34/**21.31**	26.4/26.56	44.35/**44.29**	86.59/86.27	69.16/**69.23**
Claude 3.5	5.43/5.51	17.41/17.65	27.33/27.42	38.52/38.81	82.25/82.19	74.62/74.24
Claude 3.7	5.47/5.51	17.95/18.09	28.1/28.18	41.29/41.43	83.31/83.17	73.5/73.13

Table 5: Machine translation evaluation results on FLORES+ devtest for English↔Tamazight translation. Each cell shows original dataset score / corrected dataset score. NLLB-0.6B-FT refers to the model fine-tuned on the corrected OLDI Seed dataset, while NLLB-0.6B-FT+ includes additional parallel data. S refers to the beam size used when decoding using beam search; when not specified, greedy search is used. Bold values mark the best performing metric on corrected evaluation.

imate variation coexists with quality issues. It also illustrates how massively parallel datasets can be problematic for low-resource languages: without careful linguistic validation, they risk amplifying errors and misrepresenting the language in downstream systems.

Future work will extend fine-tuning experiments with additional parallel data and explore other language pairs involving Tamazight, ensuring that corrected datasets serve as a stronger foundation for both evaluation and system development.

Acknowledgments

This work was conducted as part of the *Som Part* project, led by CIEMEN and the Fundació pels Drets Col·lectius dels Pobles, with funding from the Catalan Agency for Development Cooperation (ACCD) and the Municipality of Barcelona.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

David Adelani, Hong Liu, Xu Shen, Nikita Vassilyev, Jesujoba Alabi, Yuxiang Mao, Hongyu Gao, and Eric Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245. Association for Computational Linguistics.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, and 14 others. 2025. PARME: Parallel corpora for low-resourced Middle Eastern languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30032–30053, Vienna, Austria. Association for Computational Linguistics.

Hassan Akioud, Meftaha Ameur, Khalid Ansar, Abdessalam Boumasser, and Noura El Azrak. 2022. *Arabic-Amazigh-French Landforms Dictionary*. Royal Institute of Amazigh Culture.

Meftaha Ameur, Khalid Ansar, Abdellah Boumalk, Noura El Azrak, Rachid Laabdelaoui, and Hamid Souifi. 2016. *Dictionnaire Général de la Langue Amazighe*. Institut Royal de la Culture Amazighe.

Samira Amri, Lahcen Zenkouar, and Mustapha Outahajala. 2017. Build a morphosyntaxically annotated amazigh corpus. In *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*. Association for Computing Machinery.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Fadoua Ataa-Allah and Siham Boulaknadel. 2014. Amazigh verb conjugator. In *Proceedings of the*

Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1051–1055, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fouad Ataa-Allah and Sada Boulaknadel. 2012. Toward computational processing of less resourced languages: Primarily experiments for moroccan amazigh language. In S. Sakurai, editor, *Theory and Applications for Advanced Text Mining*, chapter 9. IntechOpen.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fatima Boukhris, Abdallah Boumalk, El Houssaïn El Moujahid, and Hamid Souifi. 2008. *La Nouvelle Grammaire de l'Amazighe*. Institut Royal de la Culture Amazighe.

Ahmed Boukous. 2009. *Phonologie de l'Amazighe*. Institut Royal de la Culture Amazighe.

Ahmed Boukous. 2014. The planning of standardizing amazigh language: The moroccan experience. *Îles d'Imesli*, 6(1):7–23.

Sada Boulaknadel and Fouad Ataa-Allah. 2013. Building a standard amazigh corpus. In *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011)*, pages 91–98, Prague, Czech Republic.

Mohammed Chafik. 1996. *Arabic-Amazigh Dictionary*. Academy of the Kingdom of Morocco.

CIEMEN and Casa Amaziga de Catalunya. 2019. *El poble amazic a Catalunya*. CIEMEN.

ElHoussain El Moujahid. 2022. *Grammaire Générative de l'Amazighe: Morphologie et Syntaxe du Nom*. Institut Royal de la Culture Amazighe.

Chris Emezue, NaijaVoices Community, B. Awobade, A. Owodunni, H. Emezue, G. M. T. Emezue, N. N. Emezue, S. Ogun, B. Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. In *Proceedings of Interspeech 2025*, pages August 17–21, Rotterdam, Netherlands. International Speech Communication Association (ISCA).

Hicham Faouzi, Mohammed El-Badaoui, Mohammed Boutalline, Abdelaziz Tannouche, and Hamza Ouanan. 2023. Towards amazigh word embedding: Corpus creation and word2vec models evaluations. *Revue d'Intelligence Artificielle*, 37(3):753–759.

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.

Amirhossein Kargaran, Amirhossein Imani, François Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4863–4875. Association for Computational Linguistics.

Amirhossein H. Kargaran, François Yvon, and Hinrich Schütze. 2025. Glotcc: an open broad-coverage commoncrawl corpus and pipeline for minority languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024)*, volume 37, pages 16983–17005, Red Hook, NY, USA. Curran Associates Inc.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. 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.

Rachid Laabdelaoui, Abdallah Boumalk, El Mehdi Iazzi, Hamid Souifi, and Khalid Ansar. 2012. *Manuel de Conjugaison de l'Amazighe*. Institut Royal de la Culture Amazighe.

Mena B. Lafkioui. 2018. *Berber Languages and Linguistics*. Oxford Bibliographies.

Michael Lau, Qi Chen, Yuchen Fang, Tian Xu, Tong Chen, and Pavel Golik. 2025. Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Svein N. Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2):102–116.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Alp Öktem and Farida Boudichat. 2025. Awal – community-powered language technology for tamazight. In *Proceedings of the Conférence Internationale sur les Technologies d'Information et de Communication pour l'Amazighe (TICAM)*, Rabat, Morocco. Institut Royal de la Culture Amazighe (IRCAM). Submitted.

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.

Mikel Roque. 2009. *Els amazigs avui, la cultura berber*. Pagès Editors / IEMed.

Lane Schwartz. 2022. Primum non nocere: Before working with indigenous data, the acl must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, USA.

Driss Soulaimani. 2016. Writing and rewriting amazigh/berber identity: Orthographies and language ideologies. *Writing Systems Research*, 8(1):1–16.