DATASHI: A Parallel English–Tashlhiyt Corpus for Orthography Normalization and Low-Resource Language Processing.

Nasser-Eddine Monir[1], **Zakaria Baou**[2]
[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]ISIMA, Université Clermont Auvergne, 63178 Aubière, France
nasser-eddine.monir@inria.fr, zakaria.baou@edu.uca.fr

Abstract

DATASHI is a new parallel English–Tashlhiyt corpus that fills a critical gap in computational resources for Amazigh languages. It contains 5,000 sentence pairs, including a 1,500-sentence subset with expert-standardized and non-standard user-generated versions, enabling systematic study of orthographic diversity and normalization. This dual design supports text-based NLP tasks—such as tokenization, translation, and normalization—and also serves as a foundation for read-speech data collection and multimodal alignment. Comprehensive evaluations with state-of-the-art Large Language Models (GPT-5, Claude-Sonnet-4.5, Gemini-2.5-Pro, Mistral, Qwen3-Max) show clear improvements from zero-shot to few-shot prompting, with Gemini-2.5-Pro achieving the lowest word and character-level error rates and exhibiting robust cross-lingual generalization. A fine-grained analysis of edit operations—deletions, substitutions, and insertions—across phonological classes (geminates, emphatics, uvulars, and pharyngeals) further highlights model-specific sensitivities to marked Tashlhiyt features and provides new diagnostic insights for low-resource Amazigh orthography normalization.

Keywords: Corpora, low-resource languages, orthography normalization, Tashlhiyt.

1. Introduction

Tashlhiyt—also known in the literature as Tashelhiyt or Shilha—is one of the three major Amazigh varieties spoken in Morocco, alongside Central Atlas Tamazight and Tarifit. It belongs to the Afroasiatic language family and is primarily used in the southern regions of Morocco, from the High Atlas to the Sous region. Although spoken by several million people, Tashlhiyt remains markedly underrepresented in digital resources and computational research. According to recent surveys of Amazigh language technologies (Akallouch et al., 2025), the overall distribution of natural language processing (NLP) datasets across dialects is highly uneven as Tashlhiyt receives only limited coverage, resulting in a 20–25% performance gap in multilingual or cross-dialectal systems.

The creation of dedicated computational resources for Tashlhiyt is essential for advancing text-based technologies. Most existing datasets are either too small, thematically narrow, or insufficiently standardized to support robust evaluation and reproducibility. The **DATASHI** corpus addresses this gap by providing parallel English–Tashlhiyt data capturing both non-standard user-generated writing and expert-standardized equivalents. This dual representation is valuable for developing normalization models and evaluating how orthographic variation affects downstream NLP tasks such as translation, tokenization, and language modeling.

At a broader level, corpora that explicitly encode orthographic diversity not only improve text-based processing, but also facilitate cross-domain transfer to speech-related technologies. Consistent text normalization enhances pronunciation modeling, forced alignment, and lexicon construction—key components for building speech recognition or synthesis systems in low-resource settings. In this sense, while **DATASHI** is primarily a textual resource, its standardized design enables systematic extension toward read-speech data collection.

This dual purpose aligns with broader efforts in low-resource language documentation, where unified multimodal datasets have proven crucial for accelerating the development of both natural language processing (NLP) and speech technologies (Adda et al., 2022; Besacier et al., 2014). Recent multilingual initiatives such as *GlobalPhone* (Schultz, 2002), *CMU Wilderness* (Black, 2019), and *BABEL* (Gales et al., 2014) have shown that building balanced, linguistically informed resources across modalities can significantly reduce data scarcity bottlenecks. However, for Amazigh languages—and especially for Tashlhiyt—no comparable resource currently exists that provides clean, normalized text suitable for phoneme-level modeling or cross-modal alignment.

The paper is organized as follows. Section 2 reviews prior work and key challenges in Tashlhiyt NLP. Section 3 introduces the **DATASHI** corpus and its linguistic design. Section 4 details the orthographic normalization framework. Section 5 outlines the evaluation setup, and Section 6 presents results and phonological analyses before concluding with future directions in Section 7.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1947–1956
11–16 May 2026.

2. Background

2.1. Amazigh and Tashlhiyt NLP

Moroccan Amazigh (Berber) languages remain among the most under-resourced languages in natural language processing. Existing digital resources are scarce, unevenly distributed, and lack large annotated corpora or standardized processing tools (Ataa Allah and Boulaknadel, 2012; Akallouch et al., 2025). Available datasets mainly target part-of-speech tagging, named-entity recognition, and small-scale parallel corpora, but remain insufficient for large-model training and evaluation (Maarouf, 2025; Amri and Zenkouar, 2017).

Computational work has produced a limited number of morphological analyzers, OCR systems for Tifinagh[1], and experimental machine-translation models, although existing developments remain small-scale, domain-specific, and rarely integrated into end-to-end NLP pipelines (Akallouch et al., 2025). Comprehensive surveys consistently identify morphological richness, dialectal heterogeneity, and orthographic variation as primary barriers to progress. Moreover, resource distribution remains asymmetric: Standard Moroccan Tamazight benefits from comparatively larger datasets, while Tashlhiyt and Tarifit varieties remain critically underrepresented (Amri and Zenkouar, 2017).

For Tashlhiyt, linguistic investigations document its particularly complex morphology, including nonconcatenative root-and-pattern processes, extensive use of consonant gemination, and intricate agreement systems for gender and number (Ridouane, 2014; Oussou, 2021; Riad, 2022).

Substantial script and orthographic variation within Tashlhiyt—from Latin, Neo-Tifinagh, and Arabic scripts to non-standard user-generated orthographies—renders tasks such as tokenization, sentence alignment, and normalization significantly more difficult than in more standardized languages (Ataa Allah and Boulaknadel, 2012; Chaker and Sellès, 2016).

Moreover, available corpora that cover Tashlhiyt (or include it) remain very limited, and high-quality, large-scale parallel corpora aligning Tashlhiyt with a major language are notably absent (Akallouch et al., 2025). The development of a dedicated parallel and standardized corpus therefore constitutes an essential step toward enabling reproducible evaluation, advancing normalization research, and supporting the broader creation of NLP resources for Tashlhiyt and related Amazigh varieties.

[1]Tifinagh is an alphabet of ancient Amazigh origin used to write modern Amazigh languages, officially adopted in Morocco for standardized orthography.

2.2. Normalization

The orthographic diversity observed in Tashlhiyt reflects the broader multiscriptal situation of Amazigh writing in Morocco. Since the official adoption of the Neo-Tifinagh alphabet by the Royal Institute of Amazigh Culture (IRCAM) in 2003, it has served as the institutional standard for Amazigh literacy. However, Latin-based transcription remains predominant in research, education, and digital communication (Ataa Allah and Boulaknadel, 2012; Chaker and Sellès, 2016).

In informal contexts—particularly on social media—users employ both the standard Arabic script and, frequently, Latin orthographies adapted to online conventions. This latter system involves employing numerals and non-standard characters (e.g., “3”, “7”, “9”) to represent phonemes absent from the Latin script. (Boulaknadel, 2018). This hybrid digital orthography, combined with the coexistence of these two scripts with Tifinagh, amplifies inconsistency in written data and hinders the design of unified NLP pipelines. As noted by (Akallouch et al., 2025), such script heterogeneity poses structural challenges for text normalization, tokenization, and corpus alignment in low-resource Amazigh settings.

Text normalization—the process of mapping nonstandard spellings and orthographic variants to a consistent form—is a critical component of NLP pipelines for low-resource languages. In this regard, normalization has been widely explored as a strategy to mitigate orthographic noise and lexical sparsity. Studies on Arabic dialects, African languages, and South Asian languages demonstrate that explicit normalization significantly improves downstream performance in machine translation, ASR post-processing, and morphological tagging (Zhang et al., 2023; Al-Sallab et al., 2018; Haque and Dossou, 2021).

In the case of Tashlhiyt, the literature still reports very limited work addressing orthographic normalization, either at the word or sentence level (Akallouch et al., 2025), leaving a clear methodological and empirical gap that this study seeks to address.

Recent research has investigated large language models (LLMs) for spelling correction and orthographic normalization across diverse languages. Models such as GPT-3, mT5, and BLOOM demonstrate promising zero- and few-shot generalization capabilities for normalization, even in languages absent from their training data (Madaan et al., 2023; Anastasopoulos et al., 2023). Prompt-based or instruction-tuned variants have been applied to dialect standardization, grapheme-to-phoneme normalization, and text cleanup in under-represented languages, often outperforming smaller supervised systems in low-data regimes (Madaan et al., 2023).

Language	Tokens	Avg. sentence len.	Min len.	Max len.	Vocabulary
English	43,654	8.68	2	16	3,160
Non-std. Tashlhiyt	35,714	7.11	2	18	7,261

Table 1: Statistics of the DATASHI Corpus

Language	Tokens	Avg. sentence len.	Min len.	Max len.	Vocabulary
Standard Tashlhiyt	12,211	8.14	2	21	3,683
Non-std. Tashlhiyt	10,698	7.13	2	18	4,583

Table 2: Statistics for the 1,500-sentence Tashlhiyt subset of the DATASHI corpus, comparing non-standard and expert-standardized versions used for orthography normalization.

These findings suggest that LLMs encode cross-lingual orthographic regularities that could be leveraged for Tashlhiyt normalization, providing both a methodological baseline and a diagnostic probe for language coverage in multilingual foundation models.

3. Corpus Creation

3.1. English and non-standard Tashlhiyt

DATASHI corpus [2] was designed to capture everyday linguistic diversity in Tashlhiyt while maintaining thematic balance and parallel alignment with English. To achieve this, we defined 20 thematic domains—including *home and daily life, food, health and body, greetings, money, technology,* and *religion*—and generated 250 sentences per domain, yielding a total of 5,000 English sentences. The English sentences were created using a semi-controlled elicitation procedure: initial templates were drafted by the authors and expanded to cover common communicative situations within each domain, then manually reviewed to ensure grammatical correctness, lexical diversity, and naturalness. Each domain was reviewed to ensure representativeness of everyday communicative contexts and coverage of a wide range of lexical and morphosyntactic constructions.

For the parallel alignment, 34 native Tashlhiyt speakers (17 male, 17 female), aged 20–50 years, were recruited. These participants were non-experts, unfamiliar with Latin-based or Neo-Tifinagh orthographies, and were instructed to transcribe the 5,000 English sentences using their own spontaneous non-standard writing conventions. The resulting material therefore reflects authentic variation in community writing practices.

[2]All resources used in this study—including the full dataset (5,000 sentence pairs and the 1,500 standardized subset), preprocessing and evaluation code, and the full prompt text—are publicly available at https://github.com/Nasseredd/datashi to support transparency and reproducibility.

Subsequently, a group of seven Amazigh language experts (four male, three female), all with advanced English proficiency (approximately CEFR level C1 or higher), manually standardized a subset of 1,500 sentences without any prior automatic or semi-automatic normalization step. This subset constitutes the gold-standard reference for the normalization experiments presented later in this study.

Table 1 summarizes corpus-level statistics for the full DATASHI dataset. The English portion comprises 43,654 tokens with an average sentence length of 8.68 tokens (ranging from 2 to 16), while the non-standard Tashlhiyt side contains 35,714 tokens, averaging 7.11 tokens per sentence (range 2–18). The vocabulary size reaches 3,160 unique types in English and 7,261 in non-standard Tashlhiyt, illustrating the high lexical variability induced by unregulated spelling and spacing conventions typical of informal writing.

3.2. Standard & non-standard Tashlhiyt

In addition to the full corpus, Table 2 reports statistics for the 1,500-sentence subset used for expert-based standardization. In this subset, the non-standard Tashlhiyt portion contains 10,698 tokens with an average sentence length of 7.13 tokens, while the standardized version includes 12,211 tokens and a slightly higher average length of 8.14 tokens. The maximum sentence length also increases from 18 to 21 tokens after normalization.

The difference in vocabulary size within the 1,500-sentence subset—between the standard and non-standard versions—is particularly informative. In this subset, the higher lexical diversity observed in the non-standard Tashlhiyt data arises primarily from inconsistent orthographic segmentation. In community writing, speakers often merge or separate morphemes and clitics unpredictably. For example, the expression ɣ akud ann (“At that moment”) can appear as ghakudan, ghakudane, gha kudan, or even rakudan. Such variation inflates the apparent vocabulary size by generating multiple orthographic forms for the same lexical

Phoneme type	Count
Vowels	16,101
Consonants	32,688

Table 3: Number of occurrences of vowels and consonants in the standardized Tashlhiyt corpus.

Consonants	Count
ɣ	1,193
ɛ	602
ḥ	470
x	327

Table 4: Number of occurrences of pharyngeal and uvular consonants in the standardized Tashlhiyt corpus.

Consonants	Count
g^w	82
k^w	42
$ɣ^w$	13
x^w	6
q^w	4
s^w	2
r^w	1

Table 5: Number of occurrences of labialized consonants (^{w}-marked) in the standardized Tashlhiyt corpus.

unit. The expert-standardized version mitigates this redundancy, reducing the vocabulary size from 4,583 to 3,683 unique types.

A similar phenomenon is observed in the complete 5,000-sentence corpus, where the non-standard Tashlhiyt side reaches 7,261 unique types. The broader thematic coverage and greater number of contributors amplify orthographic diversity.

The phonemic composition of the standardized Tashlhiyt corpus is summarized in Tables 3–6. As shown in Table 3, consonants dominate the dataset, with 32,688 occurrences compared to 16,101 vowels. This consonant-heavy profile is characteristic of Tashlhiyt phonotactics, which allows complex consonant clusters and even entirely vowelless syllables (Ridouane, 2014). The resulting skew towards consonantal segments highlights one of the main typological challenges for speech and text processing: syllabification, vowel insertion, and phoneme alignment are considerably less straightforward than in more vowel-rich languages.

Table 4 details the distribution of pharyngeal and uvular consonants, which are salient features of the Tashlhiyt phonemic inventory. The voiced uvular fricative /ɣ/ is the most frequent member of this class (1,193 tokens), reflecting its high lexical productivity and presence in both native and borrowed forms. The voiced pharyngeal /ɛ/ and the voiceless pharyngeal /ḥ/ also occur frequently, illustrating the robust presence of pharyngeal contrasts. These sounds are typologically significant and contribute to spectral and articulatory variability that can complicate both ASR and text normalization tasks (Ridouane, 2014; Boulaknadel, 2018).

Consonants	Count	Consonants	Count
tt	810	ll	465
ss	372	nn	397
dd	295	mm	150
qq	187	kk	182
yy	137	gg	139
ṣṣ	121	ṭṭ	111
zz	110	cc	101
ḍḍ	50	bb	48
jj	48	ww	46
rr	58	ẓẓ	41
ff	20	pp	13
xx	11	ḥḥ	5
hh	3	ɣɣ	1

Table 6: Number of occurrences of geminated consonant patterns in the standardized Tashlhiyt corpus, ranked from most to least frequent.

Labialization patterns, shown in Table 5, reveal that labialized consonants—especially g^w and k^w—are relatively rare but systematically attested. Their occurrence demonstrates that the corpus captures fine-grained phonological detail present in natural Tashlhiyt, where secondary articulations such as labialization and palatalization play a role in lexical contrasts. The inclusion of these forms is relevant for developing orthographic normalization schemes capable of handling diacritically marked consonants and superscript notations (e.g., ^{w}).

Finally, Table 6 summarizes the distribution of geminated consonant patterns. Gemination is pervasive in Tashlhiyt morphology and phonology, often signaling morphological boundaries, aspectual distinctions, or lexical contrasts. The most frequent geminates—*tt*, *ll*, and *nn*—reflect common morphological roots and derivational patterns. Interestingly, emphatic and pharyngeal geminates (*ṭṭ*, *ṣṣ*, *ḍḍ*, *ẓẓ*) are also well represented, underlining the phonological complexity that any normalization or grapheme-to-phoneme model must account for. The rarity of geminated /ɣ/ and /ḥ/ is consistent with phonotactic restrictions reported in descriptive grammars, where such patterns are typically disallowed or highly marked (Ridouane, 2014). Overall, the phoneme-level statistics confirm that the corpus preserves authentic phonological structures of Tashlhiyt, making it suitable for computational modeling that respects the language's typological and articulatory specificities.

Neo. shi	Ver. shi	English
tihirit	ṯṯumubil	car
tasnbḍayt	lmḥkama	court
tusnakt	lmaṭ	mathematics

Table 7: Examples comparing Neologism (Neo.) and Vernacular (Ver.) Tashlhiyt words.

Arc. shi	Con. shi	English
tawuri	lxdmt	work / function
taḍḍanga	lmuja	wave
arqqas	r̰r̰asul	messenger

Table 8: Examples of archaic (Arc.) Tashlhiyt terms and their contemporary (Con.) equivalents used in the corpus.

4. Methodology

4.1. Orthography

Normalizing user-generated Tashlhiyt text is challenging due to the absence of a single, universally applied writing standard. While Tifinagh is the official script promoted by IRCAM, literacy remains limited (Ait Laaguid and Khaloufi, 2023). Consequently, speakers often use non-standard Latin orthographies influenced by French and Arabic chat conventions, creating a gap between spontaneous writing and standardized forms required for NLP.

We adopted the Berber Latin script (Nait Zerad, 2011), following IRCAM's conventions (Boukhris et al., 2008), as it balances inclusivity and practical usability. To illustrate, consider:

Non-standard Tashlhiyt: `Irgha l7al os-sanad, our ssn7 manikhd kolo sigh ika l7mayad.`

Standardized Tashlhiyt: `Irɣa lḥal ussan ad, ur ssinɣ mani ɣ d kullu siɣ ikka lḥma ad.`

English: `It's been hot these days, I don't know where all this heat is coming from.`

Normalization operates on several levels. At the character level, numerals and digraphs representing absent phonemes (e.g., `7`→`ḥ`, `gh`→`ɣ`) are standardized (Kessai, 2018). Vowel representations influenced by French orthography (`o`, `ou`→`u`) and inconsistent word boundaries (e.g., `os-sanad`→`ussan ad`) are corrected systematically.

At the morpho-syntactic level, dialectal and grammatical variation requires linguistic normalization. For example, `our ssn7`→`ur ssinɣ` involves (a) replacing the regional suffix `-ḥ` (`-7`) with `-ɣ` (Chaker, 2019), and (b) enforcing the negative form of the verb `ssinɣ` under `ur` (Bensoukas, 2009).

Finally, phonological processes such as assimilation and fusion are resolved to preserve canonical forms. For instance, `ayt dari`→`ayddari` (voicing assimilation) and `tkkst tt`→`tkksst` illustrate phonetic spellings requiring morphophonemic normalization (Boukhris et al., 2008).

4.2. Vocabulary

A major challenge lies in reconciling standardized Amazigh lexicons—often prescriptive and neologism-rich—with contemporary spoken Tashlhiyt (Aït Ouguengay and Bouhjar, 2010; Bouhjar, 2008). Many official corpora employ terms unfamiliar to most speakers (Idhssaine, 2023). Our approach is descriptive, prioritizing lexical authenticity and modern usage.

We excluded artificial neologisms lacking community adoption, retaining common loanwords from Darija, French, and English, which form part of the living lexicon (Soulaimani, 2016). Table 7 illustrates typical contrasts between neologisms and their vernacular equivalents.

Archaic forms were included only when still widely understood, ensuring that the resulting corpus reflects real communicative usage rather than an idealized linguistic norm. This approach aims to create a dataset both linguistically sound and practically relevant for NLP applications.

5. Experimental Setup

5.1. Dataset

For the evaluation of large language models on Tashlhiyt orthography normalization, we rely on the manually standardized subset of 1,500 parallel sentence pairs from the DATASHI corpus described in Section 3. In addition, a carefully curated set of 30 representative sentences—distinct from the 1,500-sentence evaluation subset—was selected from the broader corpus for few-shot prompting (see Subsection 5.3). These sentences were manually standardized and selected to maximize orthographic diversity, covering a broad spectrum of graphemic correspondences, character-level substitutions, and morphophonemic alternations characteristic of non-standard Tashlhiyt writing.

5.2. Large Language Models

To establish a strong baseline for Tashlhiyt, we evaluated the performance of several state-of-the-art Large Language Models (LLMs). The objective was to measure their intrinsic ability to handle the complex orthographic, morphological, and phono-

Model	Zero-Shot		Few-Shot	
	WER (%) ↓	LD ↓	WER (%) ↓	LD ↓
Claude-Sonnet-4.5	47.5	4.86	43.1	4.43
Gemini-2.5-Pro	**37.8**	**3.94**	**35.5**	**3.65**
GPT-5	51.9	5.60	48.7	5.25
Mistral-Large-2411	58.2	6.16	54.7	5.80
Qwen3-Max	63.4	7.46	56.8	5.99

Table 9: Average normalization performance across models in zero-shot and few-shot settings, measured using WER and LD.

Category	Claude-Sonnet-4.5	Gemini-2.5-Pro	GPT-5	Mistral-Large-2411	Qwen3-Max
Emphatics	271	188	307	360	317
Pharyngeals	10	10	11	17	15
Uvulars	8	11	9	12	13
Labialization	125	95	132	138	128
Gemination	2962	2144	3304	4056	3738

Table 10: Total **deletions** by phonological category across models. Lower counts indicate fewer deletion errors relative to the reference.

logical variations present in our dataset without any model fine-tuning.

Our selection of models was designed to cover the leading architectures from major AI research labs, representing the current frontier of language understanding and generation capabilities. The following five models were used in our experiments: GPT-5 (OpenAI, 2025), OpenAI's flagship multimodal model known for setting industry benchmarks in reasoning and multilingual tasks; Claude-Sonnet-4.5 (Anthropic, 2025), Anthropic's latest model recognized for its speed and strong performance in rule-intensive instruction following; Gemini-2.5-Pro (**?**), Google's state-of-the-art model distinguished by its massive context window and advanced multilingual reasoning; Mistral Large 2411 (AI, 2024), the top-tier proprietary model from Mistral AI, highly regarded for its powerful reasoning and multilingual fluency; and Qwen3 Max (Team, 2025), a top-performing open model from Qwen that has demonstrated capabilities competitive with leading proprietary systems.

5.3. Prompting

Our prompting strategy was designed to rigorously constrain the LLMs' outputs and ground their behavior in established linguistic rules. The prompt itself is multi-faceted: it opens with a direct instruction to normalize the input, immediately followed by an exhaustive character set (`a, b, c, ḍ...`) to prevent invalid orthography. The core of the prompt details the specific IRCAM-based rules for handling key phonological and morphological features of Tashelhiyt, including labialization (ʷ), emphatic/pharyngeal consonants (ṭ, ḍ, ḥ, ɛ), gemination, and vowel standardization. While this prompt provided the full context for the **zero-shot** evaluation, in the **few-shot** setting it framed the task, allowing the models to generalize from the provided examples based on these explicit rules.

5.4. Evaluation metrics

Normalization performance was evaluated using Word Error Rate (WER) and Levenshtein distance (LD). WER provides a token-level measure of insertion, deletion, and substitution errors between predicted and reference sentences, while the LD offers a complementary character-level metric sensitive to finer orthographic variations typical of Tashlhiyt spelling.

During preliminary experiments, some LLM outputs introduced characters not included in the orthographic inventory specified in the prompt (e.g., č, š). These cases were counted as insertion errors but are not reported separately in the analysis.

6. Results and Discussion

6.1. Overall Normalization Performance

Table 9 presents the average normalization performance of all evaluated models under both zero-shot and few-shot settings, using two complementary metrics: the WER and the LD.

In the zero-shot setting, Gemini-2.5-Pro clearly outperformed all other models, achieving the lowest WER (37.8%) and LD (3.94). Claude-Sonnet-4.5 followed with moderate results (WER 47.5%, Lev. 4.86), while GPT-5 and Mistral showed higher error rates (WER 51.9% and 58.2%, respectively). Qwen3-Max displayed the weakest performance, with both the highest WER (63.4%) and LD (7.46),

Model	D	S	I
Claude-Sonnet-4.5	3000	3215	1452
Mistral-Large-2411	4131	4405	**1293**
Gemini-2.5-Pro	**2182**	**2516**	1818
Qwen3-Max	3801	4673	4672
GPT-5	3357	3806	1732

Table 11: Total counts of deletions, substitutions, and insertions (D, S, I) across the Tashlhiyt corpus for each model output in few-shot setting.

suggesting difficulty in handling the morphological and orthographic variability of the corpus when no prior examples were provided.

In the few-shot setting, which incorporates a small number of demonstration examples, all models improved to varying degrees. Gemini-2.5-Pro again achieved the best overall performance (WER 35.5%, Lev. 3.65), confirming its robustness and adaptability to the normalization task. Claude-Sonnet-4.5 and GPT-5 also benefited from the few-shot configuration, reducing their WERs to 43.1% and 48.7%, respectively. Mistral and Qwen3-Max showed improvements, though their relative ranking remained unchanged compared to the zero-shot condition. Overall, these results highlight Gemini-2.5-Pro as the most reliable model for Tashlhiyt normalization, both in zero-shot and few-shot contexts. The consistent improvement across models in the few-shot setup also demonstrates the benefit of in-context learning for normalization tasks involving low-resource or morphologically rich languages.

6.2. Fine-Grained Phonological Error Analysis

In this section, we conduct a detailed analysis of edit operations, decomposed into deletions, substitutions, and insertions, as summarized in Table 11. To refine this analysis, Tables 10–13 present a phonological decomposition of these edit operations, showing how each model performs with respect to specific phonological categories, namely emphatics, pharyngeals, uvulars, labialized consonants, and geminates.

Deletions

Overall, the results indicate that Gemini-2.5-Pro consistently produces the lowest deletion counts across nearly all phonological classes, confirming the robustness already observed at the global level. Claude-Sonnet-4.5 and GPT-5 follow closely, showing moderate performance particularly on emphatic and uvular consonants. In contrast, Mistral and Qwen3-Max display substantially higher deletion frequencies, particularly in geminated segments, where deletions are an order of magnitude greater than in other categories.

The concentration of deletion errors in gemination and emphatic consonants reflects the models' sensitivity to non-concatenative morphological and morphophonological structures characteristic of Amazigh languages. These findings emphasize the importance of subword-level and orthography-aware modeling in improving normalization accuracy, particularly for morphologically rich and under-resourced languages such as Tashlhiyt.

Substitutions

Substitution errors are generally more frequent than deletions or insertions across all models, indicating that normalization discrepancies tend to involve incorrect replacements rather than omissions or additions. Overall, Gemini-2.5-Pro exhibits the lowest substitution occurrences across all phonological classes, followed by Claude-Sonnet-4.5 and GPT-5, while Mistral and Qwen3-Max display substantially higher rates.

A closer examination by phonological category (Table 12) reveals that gemination dominates substitution errors for all models, with Mistral and Qwen3-Max producing the largest counts (4,200 and 4,470, respectively). In contrast, Gemini-2.5-Pro maintains the lowest gemination substitutions (2,380), showing better preservation of consonant length contrasts. For emphatic consonants, substitution errors remain relatively stable across models, with the lowest values observed for Gemini-2.5-Pro (686) and the highest for Qwen3-Max (1,450). In pharyngeal and uvular categories, Claude-Sonnet-4.5 and Gemini-2.5-Pro show minimal confusion, whereas Mistral and GPT-5 exhibit more frequent replacements, suggesting weaker robustness in handling marked phonological features.

These results indicate that substitution errors are not uniformly distributed but rather concentrated in phonologically marked segments such as emphatics, pharyngeals, and geminates. While Gemini-2.5-Pro demonstrates the most balanced behavior across categories, Mistral and Qwen3-Max tend to over-normalize or distort these phonologically complex orthographic forms, highlighting their limitations in capturing segmental distinctions reflected in Tashlhiyt morphology and orthography.

Insertions

Insertion errors are generally less frequent than deletions and substitutions, indicating that normalization discrepancies arise predominantly from omissions or replacements rather than from added segments.

Across models, Mistral exhibits the lowest overall insertion counts, followed closely by Claude-

Category	Claude-Sonnet-4.5	Gemini-2.5-Pro	GPT-5	Mistral-Large-2411	Qwen3-Max
Emphatics	1146	686	1183	1165	1450
Pharyngeals	41	34	250	576	391
Uvulars	103	37	283	570	796
Labialization	59	26	26	14	22
Gemination	3084	2380	3626	4200	4470

Table 12: Total **substitutions** by phonological category across models. Lower counts indicate fewer substitution errors relative to the reference.

Category	Claude-Sonnet-4.5	Gemini-2.5-Pro	GPT-5	Mistral-Large-2411	Qwen3-Max
Emphatics	10	14	12	0	11
Pharyngeals	9	10	30	15	109
Uvulars	3	3	16	1	46
Labialization	14	2	15	1	0
Gemination	1366	1714	1626	1185	4296

Table 13: Total **insertions** by phonological category across models. Lower counts indicate fewer insertion errors relative to the reference.

Sonnet-4.5 and Gemini-2.5-Pro, both of which show relatively controlled insertion behavior across categories. In contrast, Qwen3-Max displays markedly higher insertion rates, particularly for gemination (4,296) and pharyngeal segments (109). GPT-5 produces moderate insertion levels, with errors concentrated in geminated and labialized consonants, suggesting some instability in handling orthographic repetition and segmental marking.

The predominance of insertions in geminated consonants again underscores the models' limited capacity to preserve consonant length contrasts—a salient morphological and orthographic cue in Tashlhiyt. Occasional over-generation of pharyngeal and uvular consonants in lower-performing models may reflect inconsistent mappings of marked phonological segments to their normalized orthographic equivalents. Overall, insertion patterns corroborate the trends observed for deletions and substitutions: Mistral, Gemini 2.5, and Claude 4 maintain relatively stable normalization behavior, while Qwen3-Max exhibits the greatest redundancy and inconsistency across phonological categories.

7. Conclusion

In this paper, we presented **DATASHI**, the first parallel English–Tashlhiyt corpus designed for orthography normalization and broader Amazigh NLP development. By combining user-generated and expert-standardized texts, it captures real orthographic variation while providing a consistent benchmark for normalization and translation tasks.

Evaluation with state-of-the-art LLMs showed that Gemini-2.5-Pro achieves the best normalization accuracy and most stable phonological coverage, confirming that foundation models can generalize to low-resource settings but still face challenges with gemination and emphatic contrasts.

Beyond its textual contribution, DATASHI establishes a scalable foundation for multimodal extensions—particularly speech alignment and pronunciation modeling. Future work will expand the corpus with read-speech data, fine-tune multilingual models on normalized text, and integrate DATASHI into cross-Amazigh benchmarks. Together, these steps aim to advance resource equity and reproducible evaluation for Amazigh languages.

8. Ethical Statement

All participants gave informed consent prior to participation. Participation was voluntary, and contributors were informed about the research objectives before data collection. No personal data were collected. The corpus respects speaker privacy, cultural representation, and linguistic diversity. Data and scripts will be openly released for reproducible research, following the LREC Ethical Charter and ensuring responsible, community-centered development of Tashlhiyt language resources.

During preliminary experiments, we observed that several LLMs occasionally introduced characters not included in the orthographic inventory specified in the prompt, such as č or š. Because such outputs violate the explicitly defined character set and reflect orthographic conventions outside the scope of the normalization task, they were excluded from the error analysis reported in this paper.

9. Bibliographical References

Gilles Adda, Laurent Besacier, Sebastian Stüker, and Daan van Esch. 2022. Towards open multimodal corpora for low-resource languages. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Mistral AI. 2024. Mistral-large-instruct-2411 (123b parameters) [model card]. https://huggingface.co/mistralai/Mistral-Large-Instruct-2411. Accessed: 2025-10-25.

Brahim Ait Laaguid and Azeddine Khaloufi. 2023. Amazigh language use on social media: An exploratory study. *Jurnal Arbitrer*, 10(1):24–34. Accessed: 2025-10-25.

Youssef Aït Ouguengay and Aïcha Bouhjar. 2010. For standardised Amazigh linguistic resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Oussama Akallouch, Mohammed Akallouch, and Khalid Fardousse. 2025. Advances in amazigh language technologies: A comprehensive survey across processing domains. *Information*, 16(7):600.

Ahmad Al-Sallab, Wael Salloum, and Nizar Habash. 2018. Arabic dialect normalization for improved nlp performance. In *Proceedings of the LREC 2018 Workshop on Arabic NLP*, pages 47–55, Miyazaki, Japan.

Samir Amri and Lahbib Zenkouar. 2017. Coupling an annotated corpus and a lexicon for amazigh pos tagging. *Journal of Mobile Multimedia*, 13(3&4):222–232.

Antonios Anastasopoulos, Peng Qi, and Graham Neubig. 2023. Large multilingual models for low-resource languages: Capabilities and limitations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10245–10258, Toronto, Canada.

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2025-10-25.

Fadoua Ataa Allah and Siham Boulaknadel. 2012. Natural language processing for amazigh language: Challenges and future directions. In *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, Rabat, Morocco.

Karim Bensoukas. 2009. The loss of negative verb morphology in tashlhit: A variation approach. *Asinag*, (2):89–110. Accessed: 2025-10-25.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Alan W. Black. 2019. Cmu wilderness multilingual speech dataset. In *Proceedings of ICASSP 2019*.

Aicha Bouhjar. 2008. Amazigh language terminology in morocco or management of a "multidimensional" variation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Fatima Boukhris, Abdallah Boumalk, El Houssaïn El Moujahid, and Hamid Souifi. 2008. *La nouvelle grammaire de l'amazighe*. Institut Royal de la Culture Amazighe, Rabat, Morocco. Accessed: 2025-10-25.

Siham Boulaknadel. 2018. Sociolinguistic patterns and orthographic practices in online amazigh communication. *Revue des Études Amazighes*, 12:45–59.

Salem Chaker. 2019. Syntaxe. *Encyclopédie berbère*, 43:7644–7655. Accessed: 2025-10-25.

Salem Chaker and Anissa Sellès. 2016. Writing and rewriting amazigh/berber identity: Orthographies and language ideologies. *Writing Systems Research*, 8(1):1–18.

Mark Gales, Kate Knill, Anton Ragni, and Santosh Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project overview. In *Proceedings of SLTU 2014*.

Rezaul Haque and Bonaventure Dossou. 2021. Low-resource text normalization for african languages: A case study on wolof and hausa. In *Proceedings of the AfricaNLP Workshop at EACL*, pages 120–129.

Abdellah Idhssaine. 2023. A critical review of the sociolinguistics of the amazigh language in morocco: Documentation, teaching, and officialization. *Langues et Littératures*, 28:179–201.

Fodil Kessai. 2018. *Élaboration d'un dictionnaire électronique de berbère*. Ph.D. thesis, Institut National des Langues et Civilisations Orientales. Accessed: 2025-10-25.

Otman Maarouf. 2025. Amazigh linguistic dataset: Part-of-speech tagging, named entity recognition, and parallel corpus (tifinagh-english).

Aman Madaan, Abhilasha Suresh, and Graham Neubig. 2023. Prompting Large Language Models for text normalization in low-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13012–13025, Singapore.

Kamal Nait Zerad. 2011. *Mémento grammatical et orthographique de berbère : Kabyle – Chleuh – Rifain*, 1 edition. Éditions L'Harmattan, Paris. Accessed: 2025-10-25.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-10-25.

Said Oussou. 2021. Amazigh language morphology: Examples from tashlhiyt in ayt hdidou. *Lingua. Language and Culture*, 1:212–224.

Tomas Riad. 2022. The secret morphology of tashlhiyt berber. *Brill's Journal of Afroasiatic Languages and Linguistics*, 14(2):273–307.

Rachid Ridouane. 2014. Tashlhiyt berber. *Journal of the International Phonetic Association*, 44(2):207–221.

Tanja Schultz. 2002. Globalphone: A multilingual speech and text database developed at karlsruhe university. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Dris Soulaimani. 2016. Becoming amazigh: Standardisation, purity, and questions of identity. *The Journal of North African Studies*, 21(3):485–500.

Qwen Team. 2025. Qwen3-max: Just scale it.

Yi Zhang, Rob van der Goot, and Barbara Plank. 2023. Universal text normalization for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12456–12468, Singapore.