FIÙR: A Benchmark Dataset for Eastern Lombard Machine Translation

Edoardo Signoroni
NLP Centre, Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czechia
e.signoroni@mail.muni.cz

Pavel Rychlý
NLP Centre, Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czechia
pary@fi.muni.cz

"En fiùr sul el fà mìa primaéra."
'A single flower alone does not make spring.'

— Lombard saying

Abstract

While recent initiatives have expanded machine translation benchmarks to include low-resource and minoritized languages, regional dialect continuums remain severely underrepresented. Current datasets for the Lombard language predominantly feature Western Lombard, with Eastern Lombard varieties almost absent from the digital landscape. In this paper, we introduce, a new benchmark dataset providing an Eastern Lombard translation of the FLORES+ `dev` and `devtest` splits. We detail the methodology of translating a largely oral, unstandardized language, including orthographic regularization and the handling of Italian loanwords. Finally, we establish zero-shot baselines by prompting current LLMs and MT models, demonstrating that existing systems struggle with Lombard varieties overall, and exhibit a strong Western Lombard bias, struggling even more to accurately generate Eastern Lombard text.

Introduction

The advancement of Machine Translation (MT) and Large Language Models (LLMs) relies on standardized benchmark datasets. Resources like FLORES+ have become foundational for evaluating cross-lingual generalization. Recent initiatives such as the WMT Open Data Shared Task actively aim to democratize natural language processing by extending coverage to under-resourced and minoritized languages. Formally integrating an endangered language into a globally recognized benchmark is a crucial milestone for its digital survival; it not only increases visibility but provides the infrastructure necessary to develop and evaluate downstream community-driven NLP tools.

However, bringing language continuums into the digital sphere introduces complex challenges regarding representativeness. While linguistic variation is pervasive (Lutgen et al., 2026), language continuums are frequently flattened into single, monolithic ISO codes.

This also holds true for the language varieties of Italy (Ramponi, 2024) and for Lombard (ISO 639-3: `lmo`), an endangered Gallo-Romance language continuum spoken in Northern Italy and Southern Switzerland. As illustrated in Figure 1, the continuum is characterized by a high degree of internal variation and is broadly divided into Western and Eastern branches. Current iterations of FLORES+ and major web-scraped corpora predominantly feature Western Lombard (Signoroni and Rychlý, 2026), leaving Eastern Lombard vastly unrepresented. To address this gap, we present our Eastern Lombard translation of the FLORES+ `dev` and `devtest` splits.

1 Linguistic Background

1.1 Lombard

Lombard is a Gallo-Romance language (Brasca, 2026), with a speaker base of roughly 2.5 million.[1] Its geographic footprint covers the Italian regions of Lombardy and Piedmont (specifically the provinces of Novara and Verbano-Cusio-Ossola), alongside the Swiss cantons of Ticino and Graubünden. Functioning as a dialect continuum rather than a strictly standardized system, Lombard exhibits notable lexical, morphological, and phonetic diversity, yet speakers largely retain mutual comprehensibility (Coluzzi et al., 2018; Bonfadini, 2010;

[1] Based on recent demographic assessments (ISTAT, 2026), this figure reflects a drop of around 1 million speakers compared to the older, frequently cited estimate of 3.5 million (Moseley and Nicholas, 2010).

Proceedings of the Eleventh Conference on Machine Translation, pages 2547–2560
October 28-29, 2026 ©2026 Association for Computational Linguistics

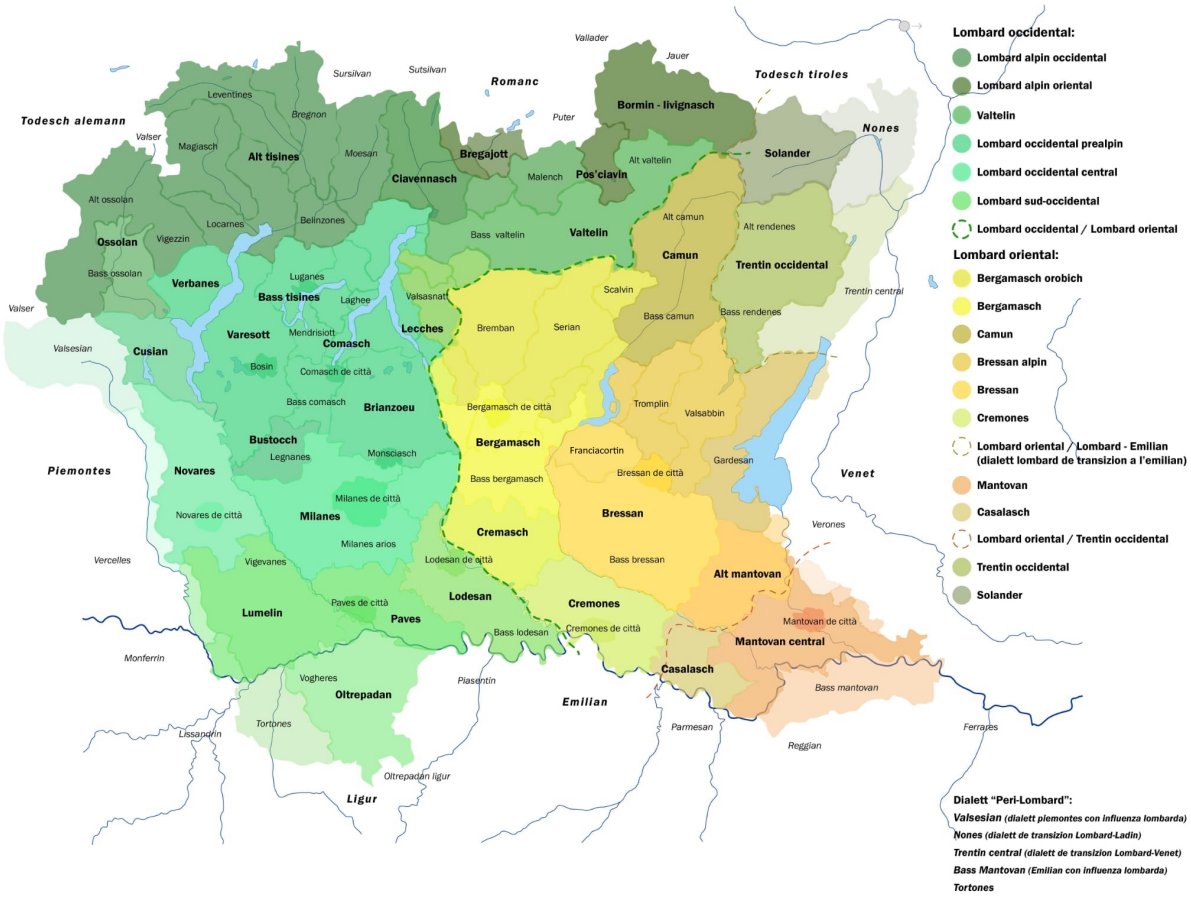

Figure 1: Map of the Lombard variants (names are in Western Lombard). While the internal subdivisions of the Lombard continuum are not always clear even to scholars, this illustrates the high degree of the internal diversity and overlap of varieties, and shows the distinction of the two main agreed-upon branches: Western Lombard (green shades) and Eastern Lombard (orange shades), divided by the dashed line.

Loporcaro, 2009). Scholars traditionally partition this continuum into two to four branches. The two most prominent are Western Lombard (spanning eastern Piedmont, Canton Ticino, and the provinces of Milan, Monza, Varese, Como, Lecco, Sondrio, Lodi, and Pavia) and Eastern Lombard (covering the Bergamo, Brescia, and Cremona provinces).[2] Figure 1 shows a map of the main varieties of Lombard.[3]

The continued survival of the language is severely threatened. UNESCO's "Atlas of the World's Languages in Danger" (Moseley and Nicholas, 2010) categorizes Lombard as "Definitely endangered".[4] Furthermore, under the Extended Graded Intergenerational Disruption Scale (EGIDS) (Lewis and Simons, 2010), its status hovers between "threatened" and "moribund".

The sociolinguistic reality of Lombard differs depending on the national context. In Switzerland, the language is supported by institutional frameworks and enjoys a relatively high degree of social prestige, spearheaded by the Bellinzona-based *Centro di dialettologia e di etnografia* (CDE). In contrast, within Italy, Lombard and comparable regional varieties face significant social stigma. They are often dismissed as mere *dialetti* ('dialects'), a term adopted even by the speakers themselves (Tamburelli et al., 2025). Because "dialect" frequently carries a negative connotation, incorrectly framing these languages as debased offshoots of Standard Italian rather than historically independent Romance languages, contemporary NLP literature prefers the bias-free term "language varieties" (Ramponi, 2024), bypassing the prejudice

[2]Due to historical emigration patterns in the 19th and 20th centuries, a variant of Eastern Lombard (Bergamasque) survives in parts of southern Brazil (Paganessi, 2017).

[3]Adapted from https://www.reddit.com/r/LearnLombardLanguage/comments/1s553nh/mappa_di_dialett_lombard/.

[4]This designation indicates that children are no longer learning the language natively at home, restricting the youngest fluent speakers to the parental generation or older.

still present in some traditional scholarship (Brasca, 2026).

In Italy, institutional recognition is virtually non-existent outside of symbolic regional legislation. For example, Lombardy's Regional Law 25/2016 formally commits to the promotion and protection of local Lombard varieties as expressions of intangible cultural heritage. However, these declarations have not translated into actionable preservation policies (Brasca et al., 2024). As a result, academics increasingly characterize Lombard and similar Italian linguistic systems as "contested" or "minoritized" languages (Tamburelli et al., 2025; Brasca et al., 2024).

Currently, the interaction between Lombard and Standard Italian takes the form of *dilalìa* (Berruto, 1987; Ramponi, 2024). Standard Italian exclusively occupies high-prestige, official domains, pushing Lombard into purely informal environments where it is constantly blended with the national language—which now serves as the primary language of socialization for children. This "vertical bilingualism" (Tamburelli, 2012; Brasca et al., 2024) actively drives language loss, as the community struggles to find practical reasons to maintain the local variety. Recent statistics highlight this rapid decline: ISTAT data from 2024 (ISTAT, 2026) shows that just 4.2% of Lombardy's population relies primarily on the local variety in domestic settings, with 18.1% alternating it with Italian. This is a noticeable drop from the 2015 data (ISTAT, 2017), which reported 5.6% and 26.1% respectively. Usage rates in public spaces or among friends are even lower, worsening the fracture of the continuum.

Although Lombard has a rich historical literary tradition, it operates almost exclusively as an oral language in the modern era. Various standardized spelling systems have been designed,[5] but none have achieved widespread community acceptance or official codification. As a result, individuals who write in Lombard usually rely on ad-hoc, subjective phonetic spelling rules to represent the sounds of their specific local variety.

1.2 Eastern Lombard

Assuming the same numbers we used before, we can roughly estimate the number of Eastern Lombard speakers at 0.6–1.0M (of whom only about 100,000 [4%] use it outside their immediate household or family circle).[6]

Eastern areas began to develop some innovations with respect to the then more uniform Milanese model starting from the Late Middle Ages (Bonfadini, 2015). Historically, the Eastern area never had a "reference" dialect comparable to *Milanès* for the West.[7] *Bergamàsch* and *Bresà* thus exhibit some differences, but also many shared distinguishable features and lexical choices that somewhat distance them from Western varieties (Bonfadini, 2010).

Specifically, vowel length is lost as a distinctive feature,[8] and Latin /s/, /ke, ki/, and /tj/ all merge into [s]. The close vowels /y/ and /i/ lower to [ø] and [e], respectively. Intervocalic /v/ is deleted, and syllable-coda /n/ after a stressed vowel is dropped without nasalization. In nominal and verbal morphology, nouns ending in /-t/ and /-n/ undergo plural palatalization to [-ʧ] and [-ɲ], the masculine singular past participle retains the final dental stop, and plural endings of feminine words in /-a/ shift to [-e].

Bresà has also its own peculiarities, sometimes due to proximity to *Vèneto*. For example BG, MI [nɔʧ] ('night') and [laʧ] ('milk') become BS [nɔ] and [lat] (ven. [ˈn ɔte]), BG [i ˈsøke] ('the pumpkins') becomes BS [le ˈsøke] (ven. [le ˈsuke], and BG [bat]/[baˈti] ('to pound') becomes BS [ˈbater] (ven. [ˈbatar]). Moreover, BG [ɔˈrɛʤa] ('ear') corresponds to BS [ɔˈrɛʧa], and the first person plural of BG [ˈnɔtɛr an ˈkanta] ('we sing') is BS [nɔˈaltɛr kanˈtɔm].

On the western borders, the distinction between *Bresà* and *Bergamàsch* fades, with transitional zones where features from both varieties overlap. Relevant to the present work, both translators are speakers of one of these contact varieties. Thus the translation features some phenomena due to the proximity to *Bergamàsch*, such as the alternative use of BG *i* and BS *le* for the definite feminine plural and of the verb forms (e.g. BG *scrif*/*scrìi* and BS *scrìer* 'to write'). Moreover, the indefinite masculine article *un* becomes *en*.

Some local phonetic features include: [o] and

[5] For example, the orthography used on the Lombard Wikipedia: `https://lmo.wikipedia.org/wiki/Wikipedia:GrafCat` (in Lombard).

[6] This is a very rough estimate done by using the percentages found in the most recent ISTAT report (ISTAT, 2026) and applying them to the combined populations of the provinces of Bergamo, Brescia, Cremona, and Mantova.

[7] Sanga (1997) seems to suggest that at least for a time *Bergamàsch* had a much more vibrant and prestigious status.

[8] According to Bonfadini (2015), distinctive vowel length existed and still exists (in some varieties) both in *Bergamàsch* and *Bresà*.

[ø] become [u] and [y] before the stress; [s] can (and almost always does) become [h]; unstressed word-final [a] is rendered as [ø].

2 Related Work

2.1 FLORES+

Since its first iteration (Guzmán et al., 2019), the FLORES+ dataset has evolved through the successive iterations of FLORES-101 (Goyal et al., 2022) and FLORES-202 (Team et al., 2022) to reach the current form of FLORES+, becoming an established benchmark dataset for MT.

The dataset is currently under the curation of the Open Language Data Initiative (OLDI), which hosts and makes available the data for an ever growing number of languages. Since 2024, OLDI has organized a shared task of the same name at WMT (Maillard et al., 2024; Dale et al., 2025) with the stated aim to add an increasing number of under-resourced languages to widely used open benchmark datasets.

The shared task has already yielded contributions for several languages: Uzbek (Mamasaidov et al., 2025), Hausa, Northern Sotho (Sepedi), Xitsonga, and isiZulu (Abdulmumin et al., 2024), Emakhuwa (Ali et al., 2024), Wu (Yu et al., 2024), Karakalpak (Mamasaidov and Shopulatov, 2024), Nko (Doumbouya et al., 2023), Aragonese, Aranese, Asturian, and Valencian (Perez-Ortiz et al., 2024), Tamazight (Oktem et al., 2025), Norwegian Bokmål (Mæhlum et al., 2025), American Sign Language (Tanzer, 2025), Erzya (Gordeev et al., 2024), Ladin (Frontull et al., 2025).

2.2 Lombard Datasets

Some recent research released data and tools for Lombard, such as parallel corpora for MT (Signoroni, 2022), multi-variety dataset for orthographical recognition (Signoroni and Rychly, 2026), and for code-mixed LangID (Signoroni et al., 2026).

Some multilingual web-scraped corpora such as NLLB (Team et al., 2022), WikiMatrix (Schwenk et al., 2021), XLEnt (El-Kishky et al., 2021), Glot500 (Imani et al., 2023), CulturaX (Nguyen et al., 2024), and HPLT 3.0 (Oepen et al., 2025), list Lombard among the languages they cover.

However, Kreutzer et al. (2022) have shown that these kind of resources are often unreliable or even unusable especially for under-resourced languages, with the text being only boilerplate, in the wrong language, or just non-linguistic noise altogether. These findings were also replicated for Lombard by Signoroni and Rychlý (2026), on a wide array of available text resources for Lombard.

On top of the low-quality alignments and examples found in the web-scraped corpora, they also show their lack of representativeness in terms of internal linguistic variation. Language varieties and variation (Lutgen et al., 2026) cannot faithfully be represented under a single ISO label (Ramponi, 2024).

3 Methodology

3.1 Translation

Translation Source We opted to translate from the Italian version of the FLORES+ `dev` and `devtest` datasets. Since most of the Lombard speakers are from the older generations, it was considered easier to find Italian-Lombard bilingual speakers, instead of English-Italian-Lombard trilingual individuals.

Searching for Volunteers and Language Perception The search for volunteers was conducted locally and in-person. We targeted cultural associations and similar groups, but there was a common perception that writing Lombard is "hard and a thing that people simply do not do". Even if anecdotal, this feedback on the perceived difficulty of the written language informed other parts of the methodology, such as the choice of orthography to adopt for the translation.

Orthography As mentioned in Section 1, Lombard does not have a standardized orthography, since nowadays it is mostly a spoken language.

During the preliminary phases, when presented with the alternative orthographies proposed, many (e.g. Pan-Lombard orthographies, devised for all Lombard variants) were considered by the translator as "another language", or even "incomprehensible" at times (due to unfamiliar characters or lexical choices).

Of the ones currently being proposed and used, we chose to adopt the *Ortografìa Orientàl Ünificàda* (LORUNIF). This was created for Eastern Lombard on the Lombard Wikipedia starting from the local orthographies for the most spoken varieties of Eastern Lombard, mainly *Bergamàsch* (*Ortografìa del Dücat*) and *Bresà* (*Ortografìa bresàna modèrna*)[9].

While agreeing in principle that the LORUNIF

[9] See Signoroni and Rychly (2026) for an overview of some of the proposed Lombard orthographies, mostly used on the Lombard Wikipedia

Split	Language / Variety	Sentences	Total Words	Words/Sent	Total Chars	Chars/Sent
dev	English	997	20,954	21.02	126,615	127.00
	Italian	997	22,984	23.05	148,692	149.14
	Western Lombard (`lmo_west`)	997	25,531	25.61	135,729	136.14
	Eastern Lombard (`lmo_east`)	997	27,164	27.25	141,014	141.44
devtest	English	1,012	21,901	21.64	133,428	131.85
	Italian	1,012	24,311	24.02	157,790	155.92
	Western Lombard (`lmo_west`)	1,012	26,954	26.63	142,878	141.18
	Eastern Lombard (`lmo_east`)	1,012	28,891	28.55	148,440	146.68

Table 1: Corpus statistics for the FLORES+ `dev` and `devtest` splits across English, Italian, Western Lombard, and Eastern Lombard.

	Mean CER (%) ↓		**Mean WER (%) ↓**	
Language Pair	**dev**	**devtest**	**dev**	**devtest**
`ita` ↔ `lmo_east`	**39.93**	**43.04**	**91.73**	**92.76**
`ita` ↔ `lmo_west`	62.16	62.47	97.79	98.39
`lmo_west` ↔ `lmo_east`	61.88	62.01	94.50	94.83
`eng` ↔ `ita`	74.72	75.01	103.99	104.93
`eng` ↔ `lmo_west`	71.01	71.16	106.26	107.30
`eng` ↔ `lmo_east`	77.74	77.82	112.14	113.72

Table 2: Bidirectional mean CER and WER between language varieties on FLORES+ `dev` and `devtest`. The

was closer to their representation of how the language should be written, in the beginning the volunteer proposed their own modifications to the system, out of concern that local phonological variations were not fully represented in the data, e.g. $ for the local pronunciation of [s] as [h] or @ for the grapheme <a> as [ø]. This is due to the perceived value and role of the language for personal sentimental and cultural heritage highlighted by the volunteer. After some discussion on the practicalities and on linguistic matters, it was decided to keep the current formulation of the spelling rules.

Appendix B gives the spelling conventions for the LORUNIF orthography. Accents are used to mark the word stress on the words with more than one syllable, and to distinguish between open (/ɛ/ /ɔ/) and closed (/e/ /o/) vowels.[10] Rounded vowels are always marked with umlauts, even if unstressed.[11]. The sequence /stʃ/ is marked as <s-c> (e.g. *s-cet* 'boy').

Regularization and Loanwords Together with the orthography, another recurrent issue is that of unregulated loanwords and idiolects: nowadays several words, even common ones, are increasingly being substituted by Italianized non-integrated loanwords. These cases are even recorded in modern-day dictionaries, such as the Bergamasque one (Francia and Gambarini, 2001). After analyzing the issue and discussing with the translator, we agreed to some modifications in post-editing, with the objective to avoid code-mixing (Muysken, 2000).

Loanwords and Lombard forms were used interchangeably without a clear pattern. In some other cases, the mixing occurred due to the perceived register of the sentence: Italian loanwords were consistently used for "higher" topics and meanings (e.g. Italian *inizio* 'beginning' was translated as *prensépe*, but *principio* 'principle' was used as-is in *princìpio*, whereas there should not have been any distinction, since it. *principio* is a synonym of *inizio*, of which lmo *prensépe* is a cognate).

Words such as *viàgio* ('journey'), *elètrone* ('electron'), *giudìce* ('judge') and others were found across the original draft text. Without a common, widespread reference to use as foundation, we opted to reinstate the Lombard forms where possible in case of unmodified lexical borrowings (e.g. *rìschio* → *rìsech* 'risk'), or integrate the loans coherently with Lombard sound changes (e.g. *càcia* → *càsa*, as the verb *casà*, 'to hunt'). And so, *viàgio* becomes *viàs*, *elètrone* becomes *eletrù*, and *giudìce* becomes *giödes* (as in the idiom *fà giödése* 'to be responsible'). In some cases, such as the latter, a further change could be possible: older Brescian dictionaries (Pellizzari, 1759), for example, spell *giustìsia* ('justice') as *zustìsia*, thus reflecting the *gi-* → *z-* change, still recognizable in more frequent words such as *zöch* ('game, toy'), *zugà* ('to play'), *zontà* ('to join'), *zenér* ('January'), *zèner* ('son-in-law') for which the forms in *gi-* are not used. Table 3 gives an overview of the rules we applied for these integrations.

Another widespread phenomenon in Eastern

[10]These accents can be omitted if there is no danger of confusion between similar words, e.g. sét /set/ 'thirst' sèt /sɛt/ 'seven'.

[11]Some edge cases still remain in the rare cases in which a rounded vowel also carries the word stress (e.g. *cümü* 'township')

Draft	Change	Example
gi	s, z	*viagiatùr* → *viazadùr*
one	ù	*protóne* → *protù*
smo	zem	*buddìsmo* → *buddizèm*
rno	ren	*intórno* → *entùren*
o	∅	*concèrto* → *consèrt*
e	∅	*variànte* → *variànt*
ci	s	*cèrto* → *sert*
sc	s	*sciensiàto* → *sensiàt*
io	e	*telescòpio* → *telescòpe*

Table 3: Summary of the changes we adopted to regularize Italian loanwords.

Lombard speech is elision and assimilation. The most represented case is the masculine singular article BS *èl*/BG *ól* ('the'), which almost always becomes *'l*. Thus, the form *l'* is always feminine and *'l* is always masculine. We kept this rule also for the prepositions: e.g. *de 'l* (masculine 'of'), *dela* (feminine 'of'). Other cases of elision due to assimilation are words starting with a vowel: e.g. *andàl'ndà* ('to go'); *importànt/'mportànt* ('important'), and so on.

Another point that we tried to address was the deletion of the *v*, e.g. *vìga/ìga* ('to have'). We kept the form in *v-* when it was preceded by a vowel (e.g. *gh'o vest* instead of the phonetically closer *gh'o est* 'I have seen'). A few instances of intervocalic *-v-* were removed (e.g. *università* → *uniersità* 'university').

Resources To aid in the translation process, we suggested several available resources in addition to the Lombard Wikipedia LORUNIF page: *Vocabolario bresciano e toscano* (so-called "*del seminario*" 'of the seminar') (Pellizzari, 1759), *Vocabolario Italiano-Bresciano* (Melchiori, 1817), *Vocabolario Bergamasco-Italiano per ogni classe di persone e specialmente per la gioventù* (Zappettini, 1859), *Vocabolario dei Dialetti Bergamaschi Antichi e Moderni* (Tiraboschi, 1873), *Vocabolario Italiano-Bergamasco* (Francia and Gambarini, 2001).

Translators The work was carried out by two people (one of them being the first author), both native speakers of Eastern Lombard. The translation was done by one volunteer. While holding a Master's Degree, they are not a professional translator. We made sure that they were familiar with the OLDI Translation guidelines, which we translated into Italian, and briefed them on some relevant linguistic concepts (e.g. dialect vs language, phonetic vs phonemic, loanwords, code-switching) to address some of the volunteer's concerns and give them some theoretical framework.

Translation Workflow The volunteer was presented with the Italian version of FLORES+ `dev` and `devtest` splits in `.txt` format. After the translation, the first author manually reviewed the translation with the help of a custom web app. The translation work spanned around 90 hours, while the post-editing took around 25 hours. Table 2 reports CER and WER between the Italian, English, and Lombard versions of FLORES+. Table 1 gives some figures regarding the size of the datasets.

4 Evaluation

To evaluate the dataset, we conduct some brief translation experiments with off-the-shelf models: NLLB-200-distilled-600M (NLLB Team, 2024), Minerva-7B-Instruct-2407 (Orlando et al., 2024), LLaMAntino-3-ANITA-8B-Inst-DPO-ITA (AI@Meta, 2024; Basile et al., 2023; Polignano et al., 2024), Mistral-Nemo-Instruct-2407 (Mistral AI Team, 2024), EuroLLM-22B-Instruct-2512 (Ramos et al., 2026), and Qwen2.5-32B-Instruct (Qwen Team, 2025). We use ChrF++ (Popović, 2017) since it is currently the least suboptimal automated metric for under-served languages (Mathur et al., 2020; Signoroni et al., 2025). We still compute and report (in Table 5) other metrics, such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), to allow for comparison.

To establish a baseline for lexical and orthographical overlap, we first compute the ChrF++ scores for a verbatim copy baseline. The highest overlap occurs between Italian (28.69 for Italian → lmo_east) and the Lombard varieties, reflecting the shared phylogenetic roots. The copy baseline between `lmo_east` and `lmo_west` sits at roughly 25 ChrF++, indicating that their surface-level orthographic and lexical realizations in the translation differ significantly. There is minimal overlap with English (16.36 ChrF++).

As for actual MT, we prompt the models cited above in a zero-shot scenario. Figure 2 presents the zero-shot MT performance of our selected LLMs. While models like EuroLLM 22B and Qwen 32B perform robustly on standard high-resourced, Italian-English pairs, their performance drops when generating Lombard. NLLB-200 achieves the highest scores when translating from English to both Lombard varieties, but a distinct

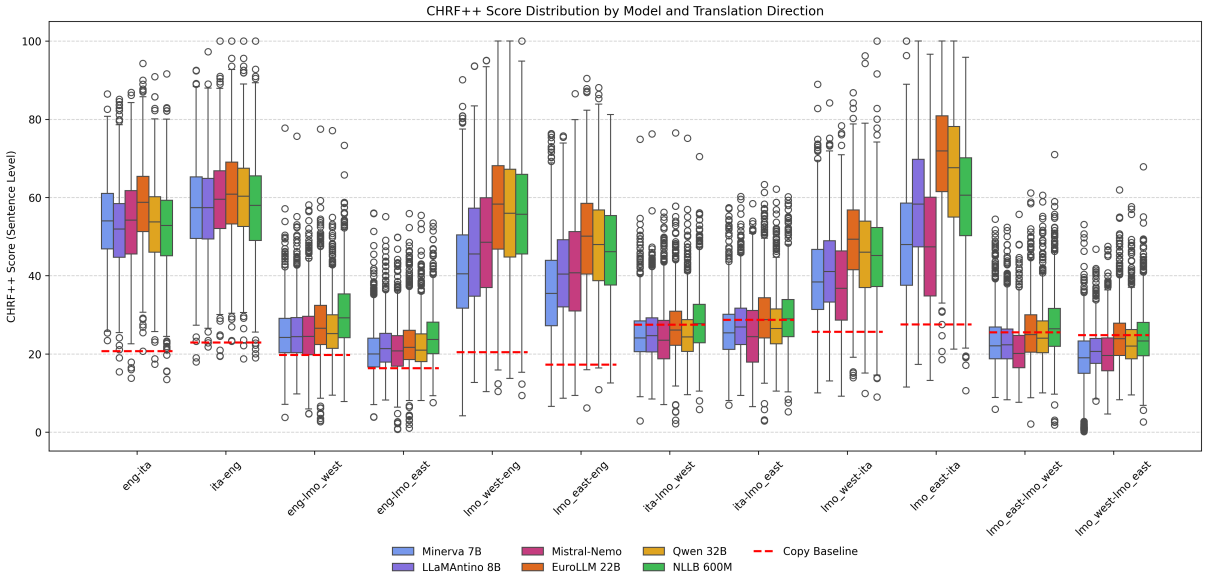

Figure 2: Zero-shot performance on MT (ChrF++) of the models we tested. We include English-Italian as a high-resource reference.

performance gap remains: all score close or below the copying baseline when translating into a Lombard variety, and universally lower when translating into `lmo_east` compared to `lmo_west`. Translating from Italian seems to be easier overall, with *lmo_west* being the harder one instead.

The discrepancy when translating from English, suggests a training data imbalance. Western Lombard is more heavily represented in existing web-scraped corpora (Signoroni and Rychlý, 2026), and models default to Western morphology and lexicon even when prompted for Eastern Lombard. By manually examining some examples, we notice that the models, even when the scores are high, tend to output Western Lombard. To verify this, we evaluate the output of the models when they are asked to translate into Eastern Lombard against the current Western reference. When translating from English, the models are more likely to generate Western Lombard (avg. +3.95 ChrF++), whereas when the source text is in Italian they are more reliant on the surface similarities with the Eastern Lombard text (avg. -2.09 ChrF++). When substituting *lmo_west* for *lmo_east* in inter-variety translation, ChrF++ scores spike (+24.62), signaling that the models are copying the source text. Similarities in the structure of the text can also explain the higher scores for *lmo_east* when translating from Italian.

These scores mirror those found by Signoroni and Rychlý (2026) when evaluating similar LLMs on a broader selection of language varieties from Italy.

5 Conclusions

In this paper we introduced a high-quality Eastern Lombard translation of the FLORES+ benchmark designed to address the lack of evaluation datasets for Eastern Lombard varieties. Translating an endangered, predominantly oral language continuum required navigating complex methodological challenges that should be addressed by involving local communities and volunteers. Our zero-shot evaluations confirmed that current LLMs and MT models are virtually blind to Eastern Lombard, consistently defaulting to Western forms due to their prevalence in training data. By providing the critical infrastructure to measure and mitigate this bias, our translation lays the groundwork for future efforts in variant-aware NLP for Lombard.

Limitations

While this translation represents a significant step forward for Lombard NLP, the dataset has some limitations. First, the translation relies on the effort of a single volunteer and the post-editing of the first author. Consequently, the dataset inevitably reflects their specific idiolects and generational linguistic habits. Second, both contributors speak a contact variety situated on the border between the Bergamasque and Brescian areas. While this could, arguably, be an advantage, the dataset does not capture the full geographic spectrum of East-

ern Lombard, and ideally each variety should be represented by its own data.

Finally, the adoption of a unified orthography, while necessary for computational consistency, forces an orthographic compromise. As noted during the translation process, the volunteer initially expressed a desire to use highly localized phonetic spelling (e.g., marking the local aspiration of /s/ to [h]). By standardizing the spelling to maintain NLP reliability, we actively flattened these localized phonetic features, which restricts the dataset's ability to represent the true phonetic granularity of the spoken language.

Ethical Considerations

Creating an NLP benchmark for an endangered, predominantly oral language continuum introduces some ethical tensions. By standardizing spelling and regularizing vocabulary in a widely used benchmark dataset, we risk implicitly crowning our specific translation as a "correct" or prestige version of Eastern Lombard. This artificial standardization can inadvertently alienate speakers whose local varieties or personal writing habits differ from our conventions, flattening the highly varied nature of the continuum. We encourage speakers to contribute translations in other varieties and orthographical standards to enable for a more comprehensive representation of Lombard.

Furthermore, we acknowledge a significant domain mismatch. Eastern Lombard varieties are socially perceived and utilized almost exclusively as a colloquial, informal heritage language. The FLORES+ benchmark, however, includes highly technical, scientific, and journalistic domains. Translating these sentences required the extensive adaptation of Italian loanwords and the generation of neologisms that may not organically exist in the everyday spoken language. Consequently, while the dataset is grammatically and lexically accurate, some of the resulting text may feel artificial to native speakers. We sought to strike a balance between respecting the natural linguistic reality of the community and the engineering necessity of creating a benchmark, prioritizing the creation of a functional NLP resource. To mitigate this issue, we provide extensive information in the text regarding which variety we are representing and the linguistic and orthographical choices we made. In this spirit, we store (on a secure server) the original draft translation as given by the translator, to be able to revert the regularization we applied.

Acknowledgments

We would like first and foremost to acknowledge Luigi Signoroni, the volunteer translator without whose patient collaboration and tireless effort this dataset would not exist.

We also wish to thank Alex Caporali, for the insightful feedback and discussion on Lombard spelling and orthography; David Dale and Samuel Frontull, for the useful suggestions regarding evaluation and validation of the data, and the reviewers for suggesting improvements to the text.

The work described herein has been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Projects No. LM2023062 LINDAT/CLARIAH-CZ.

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.

AI@Meta. 2024. Llama 3 model card.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Expanding FLORES+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA. Association for Computational Linguistics.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *Preprint*, arXiv:2312.09993.

Gaetano Berruto. 1987. Lingua, dialetto, diglossia, dialìa. In Johannes Holtus, Günter e Kramer, editor, *Romania et Slavia adriatica. Festschrift für Zarko Muljačić*, 1st edition, pages 57–81. Buske, Hamburg.

Giovanni Bonfadini. 2010. lombardi, dialetti. In Treccani Eds., editor, *Enciclopedia dell'italiano*. Treccani.

Giovanni Bonfadini. 2015. Il dialetto bresciano alla luce delle ricerche più recenti.

Lissander Brasca. 2026. Comparing intelligibility between occitan, gallo-"italic", and tuscan: Classification of gallo-"italic" empirically tested. *Lingua*, 335:1–44.

Lissander Brasca, Marco Tamburelli, Ianto Gruffydd, and Florian Breit. 2024. A matter of strength: Language policy, attitudes, and linguistic dominance in three bilingual communities. *Journal of Multilingual and Multicultural Development*.

Paolo Coluzzi, Lissander Brasca, Marco Trizzino, and Simona Scuri. 2018. Language planning for italian regional languages: the case of lombard and sicilian. In Dieter Stern, Motoki Nomachi, and Bojan Belić, editors, *Linguistic Regionalism in Eastern Europe and Beyond: Minority, Regional and Literary Microlanguages*, pages 274–298. Peter Lang, Frankfurt am Main.

David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. Findings of the WMT 2025 shared task of the open language data initiative. In *Proceedings of the Tenth Conference on Machine Translation*, pages 495–502, Suzhou, China. Association for Computational Linguistics.

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

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carmelo Francia and Emmanuele Gambarini. 2001. *Dizionario italiano-bergamasco*. Grafital, Torre Boldone.

Samuel Frontull, Thomas Ströhle, Carlo Zoli, Werner Pescosta, Ulrike Frenademez, Matteo Ruggeri, Daria Valentin, Karin Comploj, Gabriel Perathoner, Silvia Liotto, and Paolo Anvidalfarei. 2025. Bringing Ladin to FLORES+. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1061–1071, Suzhou, China. Association for Computational Linguistics.

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. FLORES+ translation and machine translation evaluation for the Erzya language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 614–623, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

ISTAT. 2017. L'uso della lingua italiana, dei dialetti e di altre lingue in italia. `https://www.istat.it/it/archivio/207961`. Accessed: 2025-08-07.

ISTAT. 2026. L'uso della lingua italiana, dei dialetti e delle lingue straniere - anno 2024. https://www.istat.it/comunicato-stampa/luso-della-lingua-italiana-dei-dialetti-e-delle-lingue-straniere-anno-2024/. Accessed: 2026-02-17.

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.

M. Paul Lewis and Gary F. Simons. 2010. Assessing endangerment: Expanding fishmans's gids.

Michele Loporcaro. 2009. *Profilo Linguistico dei Dialetti Italiani*, 1st edition. Manuali Laterza. Editori Laterza, Bari.

Anne-Marie Lutgen, Alistair Plum, Verena Blaschke, Barbara Plank, and Christoph Purschke. 2026. Variation is the norm: Embracing sociolinguistics in

nlp. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11332–11344, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Petter Mæhlum, Anders Næss Evensen, and Yves Scherrer. 2025. Improved Norwegian Bokmål translations for FLORES. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1124–1132, Suzhou, China. Association for Computational Linguistics.

Jean Maillard, Laurie Burchell, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. Findings of the WMT 2024 shared task of the open language data initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.

Mukhammadsaid Mamasaidov, Azizullah Aral, Abror Shopulatov, and Mironshoh Inomjonov. 2025. Filling the gap for uzbek: Creating translation resources for southern uzbek. *Preprint*, arXiv:2508.14586.

Mukhammadsaid Mamasaidov and Abror Shopulatov. 2024. Open language data initiative: Advancing low-resource machine translation for Karakalpak. In *Proceedings of the Ninth Conference on Machine Translation*, pages 606–613, Miami, Florida, USA. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Giovan-Battista Melchiori. 1817. *Vocabolario Bresciano-Italiano*. Franzoni e socio, Brescia.

Mistral AI Team. 2024. Mistral nemo: Our new best small model.

Christopher Moseley and Alexandre Nicholas. 2010. *Atlas of the World's Languages in Danger*, 3rd edition, volume 19 of *Memory of Peoples*. UNESCO, Paris.

Pieter Muysken. 2000. *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and 1 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:383–392.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. Hplt 3.0: Very large-scale multilingual resources for llms and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *Preprint*, arXiv:2511.01066.

Alp Oktem, Mohamed Aymane Farhi, Brahim Essaidi, Naceur Jabouja, and Farida Boudichat. 2025. Correcting the tamazight portions of FLORES+ and OLDI seed datasets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1072–1080, Suzhou, China. Association for Computational Linguistics.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Giulia Paganessi. 2017. *Brazilian Bergamasch: an Italian language spoken in Botuverá (Santa Catarina, Brazil)*. Leiden University.

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.

Bartolomeo Pellizzari. 1759. *Vocabolario bresciano e toscano compilato per facilitare a' bresciani col mezzo della materna loro lingua il ritrovamento de' vocaboli, modi di dire e proverbj toscani a quella corrispondenti*. Pietro Pianta, Brescia.

Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Confer-*

ence on Machine Translation, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Glauco Sanga. 1997. Lombardy. In Martin Maiden and Mair Parry, editors, *The Dialects of Italy*. Routledge.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Edoardo Signoroni. 2022. Piötòst ché nièn, mèi piötòst-a manually revised lombard-italian parallel corpus. *RASLAN 2022 Recent Advances in Slavonic Natural Language Processing*, page 105.

Edoardo Signoroni, Emma Bednaříková, and Pavel Rychlý. 2026. Rebelòt: Datasets and token-level language identification for lombard-italian-english code-mixing. In *Proceedings of the SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL (LREC-COLING 2026)*, pages 253–262.

Edoardo Signoroni and Pavel Rychlý. 2026. Can LLMs translate Italy's language varieties? In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 69–77, Rabat, Morocco. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychlý. 2026. Chi nas dal soch el sent de legn - auditing text corpora for lombard. In *Text, Speech, and Dialogue. TSD 2026*, volume 16940 of *Lecture Notes in Computer Science*. Springer, Cham.

Edoardo Signoroni and Pavel Rychly. 2026. Lombardographia: Automatic classification of lombard orthography variants. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 3270–3280, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Edoardo Signoroni, Pavel Rychly, and Ruggero Signoroni. 2025. Efficient architectures for low-resource machine translation. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 39–64, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Matthew Snover, Bonnie Dorr, Rich 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: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Marco Tamburelli. 2012. Dalla diglossia al bilinguismo: note sul mantenimento del patrimonio linguistico d'italia. In *L'enseignement des langues locales Institutions, méthodes, ideologies. Rome: Aracne*.

Marco Tamburelli, Ianto Gruffydd, Florian Breit, and Lissander Brasca. 2025. Modelling language attitudes: Attitudinal measurements and linguistic behaviour in two bilingual communities. *Journal of Language and Social Psychology*, 44(3-4):257–296.

Garrett Tanzer. 2025. FLEURS-ASL: Including American Sign Language in massively multilingual multitask evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6167–6191, Albuquerque, New Mexico. 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. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Antonio Tiraboschi. 1873. *Vocabolario dei Dialetti Bergamaschi Antichi e Moderni*. Editrice Fratelli Bolis, Bergamo.

Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. Machine translation evaluation benchmark for Wu Chinese: Workflow and analysis. In *Proceedings of the Ninth Conference on Machine Translation*, pages 600–605, Miami, Florida, USA. Association for Computational Linguistics.

Stefano Zappettini. 1859. *Vocabolario Bergamasco-Italiano per ogni classe di persone e specialmente per la gioventù*. Pagnoncelli, Bergamo.

A Automated Classification

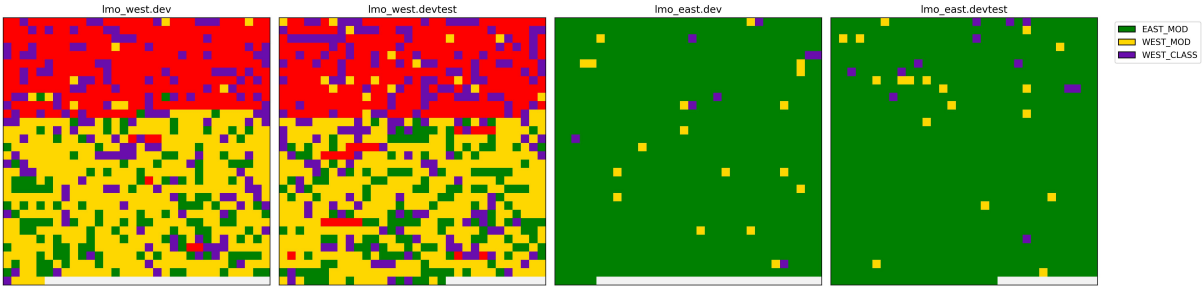

Figure 3: Plots of the orthographical composition of the FLORES+ `dev` and `devtest` according to automated classification (1 square = 1 example). Western Classical in purple; Western Modern in yellow; Eastern Modern (our) in green; *Scriver Lombard* (SL) in red.

As we mentioned in the Introduction, the motivation for this work is the lack of a benchmark dataset for Eastern Lombard. This claim is supported by audits of the data (Signoroni and Rychlý, 2026), and by our own observations, which show that the current `lmo` data in FLORES+ is almost exclusively in Western Lombard. Hence, we will refer to it provisionally as *lmo_west*.

To validate our manual observations on the whole dataset, we run the `LombardoGraphia` classifiers (Signoroni and Rychly, 2026) on both the current `lmo_west` and our `lmo_east` translation.

Following Signoroni and Rychlý (2026), we aggregate the orthographical tags in four macro-groups to smooth out variant under-specification and differing performance of the classifiers on minority classes. The grouping is guided by both linguistic and empirical evidence. The first group is "Western Classical" (WESTCLAS), which comprises the Milanese Classical Orthography (MILCLASS) and the *Noeva Ortografia Lombarda* (NOL), a Pan-Lombard orthography which is heavily influenced by MILCLASS. The second group is Western Modern (WESTMOD), formed by the *Urtugrafia insübrica ünificada* (LOCC) and by the LSI orthography, that are both "modern" (i.e. phonological, using umlauts) systems for Western (and Alpine) Lombard. "Eastern Modern" (EASTMOD) covers all the major Easter Lombard variants (LORUNIF *Ortografèa Orientàl Ünificàda*, BERGDUC *Ortografea del Dücat*, BREMOD *Ortografèa Bressàna Modèrna*, CRES *Urtugrafìa Cremàsca*), which are similar to one another. Lastly, *Scriver Lombard* (SL), another Pan-Lombard orthography, constitutes a separate group due to its use of proprietary characters (e.g. *ç*, *x*) and orthographical choices (e.g. *qe* vs *che*) which render it somewhat removed from the other proposed orthographies.

While orthographies cannot be the sole indicator of linguistic variety, they cannot be firmly disentangled, not even in the case of Pan-Lombard systems. However they can still be usable proxies for the provenance of the lexical content of the text (Signoroni and Rychly, 2026). Simply put, we maintain that one can safely assume that a text written in e.g. MILCLASS will be lexically *Milanés*, and not in e.g. *Bergamàsch*; and vice-versa a text orthographically in BERGDUC, will be most assuredly not from a Milanese speaker, but from a Bergamasque one.

Figure 3 visualizes the results for the automatic classification. Even at a first glance, the difference in composition for the two versions is striking: the `lmo_west` version presents an almost clean split between SL and WESTMOD. As is expected, some examples are classified as WESTCLAS and EASTMOD, but most of these are misclassifications due to the models not being perfect or the sentences being ambiguous or under-specified orthographically.

The presence of multiple orthographic systems in a benchmark dataset poses problems for the representativeness and the reliability of the dataset itself, confounding both models and humans about the nature of the Lombard language continuum and the current MT performance for its varieties.

Our `lmo_east` translation is, instead, mostly uniform. EASTMOD dominates the classification plot, with other tags being misclassified examples.

B Orthographical Conventions

Phoneme	Grapheme	Examples
Vowels		
/a/	a	*áqua* (‘water’)
/e/	é	*sét* (‘thirst’)
/ɛ/	è	*sèt* (‘seven’)
/i/	i	*ìga* (‘to have’)
/o/	ó	*póm* (‘apple’)
/ɔ/	ò	*òm* (‘man’)
/u/	u	*ùlem* (‘elm’)
/y/	ü	*giü* (‘one’)
/ø/	ö	*öa* (‘grapes’)
Consonants		
/p/	p	*scarpù* (‘boot’), *pà* (‘bread’), *gróp* (‘knot’)
/b/	b	*bò* (‘ox’), *sbòss* (‘sketch’)
/t/	t	*tòch* (‘piece’)
/d/	d	*dit* (‘finger’), *stràda* (‘road’)
/ʧ/	c (before *e, i*) ci (before *a, o, u*) cc (word-final)	*cél* (‘sky’), *cìca* (‘marble, pellet’) *ciar* (‘clear’), *ciós* (‘field’), *paciüch* (‘mud’) *öcc* (‘eyes’), *gacc* (‘cats’)
/ʤ/	g (before *e, i*) gi (before *a, o, u*)	*getà* (‘to throw’) *giàca* (‘jacket’), *giöst* (‘right, correct’), *giupì* (‘clown’)
/k/	c (before *a, o, u*) ch (before *e, i* or word-final) q (before *u* /w/)	*còl* (‘head’), *àca* (‘cow’), *cuacià* (‘to caress’) *chèst* (‘this’), *sorèch* (‘mouse’) *quat* (‘how much’)
/g/	g	*gombèt* (‘elbow’), *sórga* (‘rat’)
/f/	f	*röfa* (‘dandruff’), *féra* (‘fair, market’), *catìf* (‘evil’)
/v/	v	*vagù* (‘wagon’)
/s/	s	*servèl* (‘brain’), *spössà* (‘to smell bad’), *fés* (‘more’)
/z/	z	*zùen* (‘young’)
/ʃ/	sc (before *e, i*) sci (before *a, o, u*)	*scèr* (‘Turkey oak’), *scì* (‘ski’) *sciàl* (‘shawl’), *sciùra* (‘lady’), *sciòr* (‘gentleman, rich man’)
/m/	m	*mestér* (‘job’), *s-ciöma* (‘foam’), *salàm* (‘salami’)
/n/	n	*nas* (‘nose’), *snazà* (‘to smell’), *pan* (‘cloth’)
/ɲ/	gn	*gnèch* (‘sturdy’), *scàgna* (‘chair’), *règn* (‘kingdom’)
/l/	l	*löch* (‘place’), *balòss* (‘mischievous’), *pól* (‘chicken’)
/ʎ/	gli	*sbaglià* (‘to make a mistake’)
/r/	r	*röda* (‘wheel’), *geröl* (‘pebble’), *tòr* (‘tower’)
/j/	i	*pìat* (‘flat’)
/w/	u	*quach* (‘some’)

Table 4: Spelling conventions of the LORUNIF orthography.

C Prompting Results

Direction	Model	BLEU ↑	chrF++ ↑	TER ↓	CER (%) ↓	WER (%) ↓
eng-ita	EuroLLM 22B	**32.99**	**58.46**	**54.44**	**41.77**	**59.04**
	LLaMAntino 8B	24.66	52.02	65.80	49.23	70.33
	Minerva 7B	26.36	53.96	60.50	46.17	65.03
	Mistral-Nemo	22.53	53.65	77.42	62.20	82.06
	NLLB 600M	25.99	52.59	61.56	45.25	65.32
	Qwen 32B	24.20	53.35	71.80	56.38	76.28
ita-eng	EuroLLM 22B	**33.70**	**61.35**	**55.63**	**43.73**	**60.97**
	LLaMAntino 8B	26.30	57.41	67.34	53.44	72.54
	Minerva 7B	29.88	57.97	59.51	46.88	64.63
	Mistral-Nemo	31.33	59.73	58.34	46.44	63.42
	NLLB 600M	31.87	57.65	58.13	44.18	62.41
	Qwen 32B	31.70	60.36	58.52	47.00	63.91
eng-lmo_west	EuroLLM 22B	4.99	28.00	91.23	59.14	92.18
	LLaMAntino 8B	3.15	25.60	126.49	96.31	127.82
	Minerva 7B	3.08	25.63	90.44	60.96	91.09
	Mistral-Nemo	2.76	25.17	129.03	107.44	129.81
	NLLB 600M	**6.12**	**30.66**	**82.36**	**51.77**	**83.37**
	Qwen 32B	3.79	26.70	97.46	67.45	98.14
lmo_west-eng	EuroLLM 22B	31.36	**57.92**	59.40	41.68	61.86
	LLaMAntino 8B	12.44	45.90	132.73	107.78	135.68
	Minerva 7B	14.98	42.27	85.50	62.16	88.21
	Mistral-Nemo	14.41	47.39	117.18	96.66	119.71
	NLLB 600M	**32.38**	56.06	**55.98**	**39.64**	**58.08**
	Qwen 32B	25.75	55.81	76.71	59.46	79.34
eng-lmo_east	EuroLLM 22B	3.50	23.01	93.07	66.67	94.14
	LLaMAntino 8B	2.24	22.29	124.38	98.69	125.71
	Minerva 7B	2.20	21.10	93.41	67.91	94.13
	Mistral-Nemo	1.86	21.08	142.00	127.71	142.92
	NLLB 600M	**3.85**	**24.94**	**87.65**	**60.88**	**88.86**
	Qwen 32B	3.00	22.42	105.10	81.40	105.97
lmo_east-eng	EuroLLM 22B	20.45	**50.26**	77.74	57.23	82.07
	LLaMAntino 8B	8.23	40.88	153.18	126.22	156.90
	Minerva 7B	9.21	36.59	102.08	75.07	105.32
	Mistral-Nemo	8.80	40.56	144.75	119.49	148.43
	NLLB 600M	**20.70**	47.08	**74.18**	**54.89**	**77.98**
	Qwen 32B	16.39	48.33	97.30	78.36	101.74
ita-lmo_west	EuroLLM 22B	4.35	27.52	91.83	64.54	93.17
	LLaMAntino 8B	2.58	25.65	131.85	103.74	133.49
	Minerva 7B	2.89	25.66	93.06	67.54	93.88
	Mistral-Nemo	2.28	24.20	137.52	120.61	138.63
	NLLB 600M	**5.08**	**29.06**	**88.37**	**61.16**	**89.90**
	Qwen 32B	3.31	25.77	101.15	74.51	102.15
lmo_west-ita	EuroLLM 22B	**23.44**	**49.62**	**66.69**	**48.94**	**71.05**
	LLaMAntino 8B	9.46	41.23	140.13	108.03	144.07
	Minerva 7B	14.09	39.96	82.80	60.78	86.40
	Mistral-Nemo	5.81	37.03	213.95	179.02	217.32
	NLLB 600M	18.94	45.33	71.53	51.02	74.96
	Qwen 32B	14.22	45.60	107.30	84.90	111.59
ita-lmo_east	EuroLLM 22B	**6.80**	**30.15**	**80.45**	**44.67**	**80.98**
	LLaMAntino 8B	4.19	27.82	112.58	81.18	113.38
	Minerva 7B	4.54	26.67	86.11	52.53	86.43
	Mistral-Nemo	3.14	25.07	143.71	127.00	144.21
	NLLB 600M	6.68	30.11	80.69	46.34	81.34
	Qwen 32B	5.27	27.98	98.54	65.95	98.94
lmo_east-ita	EuroLLM 22B	**52.95**	**71.27**	**35.92**	**23.19**	**36.70**
	LLaMAntino 8B	21.20	57.65	123.82	100.29	124.93
	Minerva 7B	24.14	48.84	70.49	49.26	72.30
	Mistral-Nemo	11.75	47.20	201.69	171.85	202.81
	NLLB 600M	38.01	60.28	49.88	30.41	50.96
	Qwen 32B	34.02	65.57	77.13	62.24	77.94
lmo_west-lmo_east	EuroLLM 22B	**3.78**	**24.76**	**89.72**	**60.92**	**91.02**
	LLaMAntino 8B	1.38	21.46	186.05	165.72	187.31
	Minerva 7B	1.94	19.02	96.30	74.38	96.97
	Mistral-Nemo	1.11	19.82	241.10	248.89	242.04
	NLLB 600M	3.76	24.70	89.81	62.47	91.18
	Qwen 32B	2.74	23.37	113.63	92.11	114.70
lmo_east-lmo_west	EuroLLM 22B	3.76	26.08	95.67	64.43	97.08
	LLaMAntino 8B	1.51	23.35	195.46	169.33	196.93
	Minerva 7B	2.39	23.94	98.75	73.75	99.60
	Mistral-Nemo	1.07	20.77	249.16	241.69	250.01
	NLLB 600M	**4.74**	**27.96**	**92.44**	**62.61**	**94.20**
	Qwen 32B	2.90	25.36	109.22	81.50	110.50

Table 5: Comprehensive evaluation results across all 12 translation directions for all models (BLEU, chrF++, TER, CER, and WER). Bold values indicate the best performance per direction.