

How Do Italian Varieties Behave in Embedding Spaces?

Tommaso Errico Federico Ferrari Edoardo Paccagnella
3377169@studbocconi.it 3226571@studbocconi.it 3352759@studbocconi.it

Raffaele Simbari Samuele Nicolò Straccialini
3185670@studbocconi.it 3199302@studbocconi.it

Bocconi University

Abstract

Distributed representations are increasingly used in NLP to study how language varieties relate. Italian dialects offer a particularly rich testbed: a deep, well-documented continuum shaped by centuries of cross-linguistic contact. Yet no systematic cross-method comparison exists for these varieties. We compare six representation families (TF-IDF, Word2Vec, FastText, XLM-R, CANINE, LaBSE) on 17 varieties (six Italo-Romance dialects, eleven comparison languages), training on Wikipedia and OLDI text, evaluating on the parallel FLORES+ centroids, and comparing the resulting distance matrices against a documented historical-influence reference and a lexicostatistical gold (LDND, Swadesh-207). All three pretrained models collapse the dialects zero-shot; after dialect adaptation, only character-level CANINE remains usable ($\rho = .725$). Surface methods recover linguistic structure better than pretrained encoders overall: character TF-IDF reaches $\rho = .905$ against LDND. The strongest findings come from cross-method agreement, which separates robust linguistic patterns from single-method artefacts.

1 Introduction

The Italian peninsula hosts a deep dialect continuum shaped by Latin descent, Romance neighbours, and centuries of external contact. We focus on six varieties from this continuum (Friulian, Ligurian, Lombard, Sardinian, Sicilian, Venetian), all low-resource, lacking standard orthographies and barely covered by mainstream multilingual models.

We ask one question: *do different embedding families agree on which Italo-Romance dialects pattern with which other languages, and do those patterns reproduce established linguistic references?* We hold variety set, training data, and evaluation fixed and vary only the representation. Six families (TF-IDF, Word2Vec, FastText, XLM-R, CANINE, LaBSE) train on Wikipedia and OLDI;

pretrained encoders are tested both zero-shot and after dialect-targeted continued training. Every model is evaluated on the same 1,827 parallel FLORES+ centroids of 17 varieties (the six dialects above plus eleven comparison languages from neighbouring Romance, Germanic, Slavic, and Uralic families) against a lexicostatistical gold (LDND) and a historical-influence reference.

Our contributions are three. **(i)** A reproducible benchmark on 17 varieties, built on carefully cleaned Wikipedia data and manually reviewed parallel corpora.¹ **(ii)** Surface methods outperform pretrained encoders ($\rho = .905$ for character TF-IDF against LDND), and only character-level CANINE remains usable among the pretrained after dialect adaptation ($\rho = .725$). **(iii)** Cross-method comparisons (silhouette, nearest-neighbour tables, Mantel correlations) that identify which dialect patterns are robust across architectures.

2 Data

Dataset choice is crucial for low-resource varieties, so we screen the available resources against three requirements. First, **volume**: each variety needs enough text to estimate stable representations. Second, **comparability**: distances between varieties only make sense on the same content, otherwise they reflect topical differences rather than linguistic ones. Third, **signal density**: modelling similarity requires sentences with real linguistic content, not boilerplate. Only three corpora jointly satisfy these requirements. **Wikipedia** dumps supply the only realistic source of large-volume monolingual text for Italian dialects, with per-variety sizes ranging from tens of thousands to over a hundred thousand sentences. **FLORES+** (NLLB Team, 2024) contributes 2,009 sentences professionally translated across hundreds of varieties, giving a fully parallel evaluation set. **OLDI Seed** (NLLB Team, 2024)

¹Code and outputs at github.com/sstraccialini/dialects.

adds 6,193 parallel sentences across the dialects, which we use as training material and as an additional parallel anchor.

The intersection of the three sources fixes our dialect inventory at **six Italo-Romance varieties**: Friulian, Ligurian, Lombard, Sardinian, Sicilian, and Venetian. We pair them with **eleven comparison languages** spanning the historical, geographic, and genealogical neighborhood of Italy: sibling Romance varieties (Italian, Spanish, French, Catalan, Portuguese, Occitan); Germanic languages (German, English); and the Slavic contact zone along the eastern border (Slovenian, Croatian). Hungarian enters as a control: a Uralic isolate should appear at the far end of any genuinely linguistic similarity measure.

2.1 Cleaning and normalization

Our cleaning operationalises the signal-density requirement at the sentence level: keep sentences that carry linguistic signal, drop those that do not.

For Wikipedia, we extract the raw dumps and assemble a sentence-level dataset through a pipeline that combines the cleaning steps from [Camposampiero et al. \(2022\)](#) and [Jauhiainen et al. \(2022\)](#) with our own additions. Dialect Wikipedias include considerable bot-generated content and templated articles. So we first apply length and structure filters, then manually inspect the surviving files to identify recurring patterns and add per-variety blocklists. Comparison-language dumps go through the same pipeline without per-variety blocklists: they are larger and less dominated by templated content. Sentence- and prefix-level deduplication close the pipeline. After cleaning, the six Wikipedia corpora range from roughly 22k (Friulian) to 128k (Lombard) sentences; comparison-language Wikipedias are all far larger.

The smaller size of FLORES+ and OLDI Seed makes manual review at the sentence level feasible. The parallel structure makes it cheap: a single decision on the Italian source line, the authors’ native language, propagates to all aligned variety columns. Indeed, a meaningless, foreign-language, or overly technical Italian sentence cannot yield meaningful dialect signal. All five authors independently review the parallel corpora. We drop any sentence flagged by at least three reviewers and discuss those flagged by two before resolving. We apply three jointly agreed rules: (i) drop outright junk; (ii) drop sentences too short to carry structural information; (iii) drop sentences with little usable lexical con-

tent, such as heavily domain-specific material or text dominated by proper nouns, foreign names, or item lists. Although Italian anchors each judgment, we keep the aligned translations in view: this reveals a failure mode where film titles, brand names, and quoted English phrases are localised in the Italian column but left unchanged in the dialect columns, producing spurious orthographic alignment. We remove these as well. Manual review reduced FLORES+ from **2,009 to 1,827** sentences per variety and OLDI Seed from **6,193 to 5,167**. We also create a normalized version of each corpus by stripping diacritics, digits, and non-Latin characters and lowercasing the result to plain Latin letters and single spaces. Normalization is a deliberate trade-off: it reduces noise and stabilizes estimates when data is scarce, but it also erases the surface variation that surface-based models are designed to capture.

3 Methods

Our training data comes from Wikipedia and OLDI, combining the largest data volume at our disposal with a clean cross-lingual signal. For each variety, the training set is capped at 100k sentences: for dialects, we use all OLDI Seed data and fill the rest of the budget with Wikipedia; for comparison languages, we use Wikipedia only. Full variety inventory and training sizes are reported in Table 1.

Surface methods use both *native* and *normalized* data, treated as complementary. Pretrained encoders use *native* text only, since normalization removes information they rely on.

3.1 Models

We pick three surface methods that cover complementary aspects of lexical similarity. We fit two **TF-IDF** variants on the pooled training data: word-level (unigrams + bigrams) and character-level (char_wb 3–5 grams), both with sublinear term frequency scaling. We also train a skip-gram **Word2Vec** ([Mikolov et al., 2013](#)) model on all pooled varieties. Sentence embeddings are computed by averaging in-vocabulary token vectors. Word2Vec tests whether word-level co-occurrence alone suffices to recover dialect relations. **Fast-Text** ([Bojanowski et al., 2017](#)) follows the same pipeline as Word2Vec, but additionally decomposes words into character n -grams ($n = 3-6$). Explicit subword modeling is particularly relevant for low-resource dialects with unstable orthographies and

rich morpho-phonological variation.

We choose three pretrained architectures that cover complementary properties of the embedding space. We use **XLM-R** (Conneau et al., 2020) as the canonical multilingual Transformer baseline. Its fixed subword vocabulary makes it a strong reference point for cross-lingual transfer and tests of how well a standard subword-based encoder can represent low-resource dialects. **CANINE** (Clark et al., 2022) is a tokenizer-free character-level encoder over Unicode. Its design should be more robust to orthographic variation and unseen forms, which are common in dialect data. **LaBSE** (Feng et al., 2022) is a sentence encoder trained with a translation-alignment objective. We include it as an informative counter-example: if semantic alignment suppresses language-specific signal too strongly, it should perform poorly as a model of dialect relatedness. We evaluate each pretrained encoder both zero-shot and after continued training on dialects (MLM on the training data for XLM-R and CANINE, MNRL on Italian-dialect parallel pairs for LaBSE). This separates pretrained multilingual structure from adaptation effects.

3.2 Evaluation

We evaluate on FLORES+: each variety is represented by the mean of its 1,827 sentence embeddings, yielding a 17×17 cosine-distance matrix between centroids. Macro-family cohesion is measured with the silhouette score over four typological groups (Romance, Germanic, Slavic, Uralic), complemented by 2D UMAP and t-SNE projections of the centroids for visual verification. This first check establishes whether a model has captured any meaningful typological structure at all; only models that pass it are trusted for the finer-grained analyses below. For finer-grained pairwise analysis, we examine nearest-neighbour relations in both directions: for each dialect, we rank the closest non-Italo-Romance standards (excluding Italian, whose position we verify separately), and for each non-Italo-Romance standard, we rank the closest dialects. We also rank dialects by their average cosine distance to the comparison languages, distinguishing dialects most integrated into the wider inventory from those at the periphery. We base this analysis on the historical-contact reference described in Appendix C. We further compare each distance matrix against a lexicostatistical gold reference (LDND, Wichmann et al., 2010, on a Swadesh-207 list; Appendix D.1). We compute

Spearman’s ρ both on the full upper triangle of the matrix and on each single-dialect vs. comparison-languages block. We assess reliability with 95% bootstrap CIs and Mantel permutation tests; details in Appendix D.3.

4 Results

4.1 What embeddings recover

Pretrained encoders (*XLM-R*, *CANINE*, *LaBSE*) saw little or no dialect text during pretraining and behave unreliably on it. Their zero-shot silhouette can be deceptively high, yet 2D UMAP projections show the dialects collapsing into a single dense region rather than separating into the expected typological clusters, with XLM-R the most extreme case (Figure 2). Continued training on dialect text affects the three encoders differently. *LaBSE-MNRL*, whose underlying objective clusters sentence embeddings from different languages at the same point in space, suffers **catastrophic forgetting**: Italian, used as the anchor in the MNRL pairs, is pulled toward the dialects and drifts away from the other standards (average cosine similarity between pairs of comparison languages drops $0.98 \rightarrow 0.84$) (Figure 3). *XLM-R MLM* loses much of its original macro structure as well, with its silhouette score turning negative; only *CANINE-MLM* achieves a good silhouette overall (+0.57). Anisotropy is a further issue with the pretrained encoders, and the standard mitigations we tried only degraded the dialect signal further. The remainder of this discussion focuses on the surface methods and the dialect-adapted CANINE, the only pretrained encoder we deem reliable for the analyses below.

Silhouette scores rise from *TF-IDF* through *Word2Vec* to *fastText*, with *CANINE-MLM* best overall (Table 2); all of them recover the four typological macro-families.

We interpret the embedding-space geometry against the historical-influence reference (Table 3), which lists three documented contact languages per dialect. We focus on patterns that recur across methods and note method-level divergences where they shed light on the analysis.

Sardinian’s nearest external standard is Spanish in every method, and Spanish sits closer to Sardinian than Italian across all nine methods, consistent with Sardo-Iberian history (Viridis, 2023). Lombard’s nearest is Catalan in seven methods, Occitan in the other two (FastText native, CANINE-

MLM); Ligurian’s nearest is Portuguese in seven methods, Spanish in the other two (the Word2Vec variants). In both cases only the minority top-1 (Occitan for Lombard, Spanish for Ligurian) is in the reference, so the dissenting methods align more directly with documented contact. The remaining Romance reference contacts appear only sporadically, with FastText and Word2Vec recovering Sardinian-Catalan.

English also shows uniformly elevated similarities to all dialects, likely reflecting the large presence of English citations, titles, and named entities in the data that leak into the centroid.

Non-Romance contacts can appear only from the language-to-dialect direction; from the dialect’s side, Romance neighbours dominate. Slovenian and Croatian rank Venetian as their nearest dialect in eight of nine methods (CANINE-MLM places it second), consistent with Adriatic contact (Sadovski-Kornprobst, 2021), while other non-Romance reference contacts (e.g., German) emerge less robustly across methods. With only six dialects in the candidate set, we read counts as suggestive.

The integration ranking broadly reflects geographic position. Northern continental dialects (Lombard, Venetian) sit closest to the centre of the wider inventory (TF-IDF char native excepted); peninsular and coastal dialects (Ligurian, Friulian) occupy the middle; and island dialects (Sardinian, Sicilian) sit at the periphery. Sicilian is uniquely the most isolated in every method, in line with its multiple external influences and UNESCO recognition as a distinct language (Moseley, 2010). CANINE-MLM is an exception: it ranks Friulian as the most integrated dialect and as Slovenian and Croatian’s nearest dialect. Apart from this signal, Friulian’s documented Germanic-Slavic contacts are not consistently recovered.

4.2 Lexicostatistical alignment

The comparison with lexicostatistical gold has clear results. Surface lexical models perform best: char TF-IDF reaches $\rho = .905$ on the full matrix, with fastText and Word2Vec close behind. This is expected, since both LDND and character-level models are strongly driven by surface form similarity.

Among pretrained encoders, CANINE is the only one that substantially recovers the lexical structure. Dialect MLM continued training improves its full-matrix score from .664 to .725, and its dialect-external score from .700 to .803.

XLM-R and LaBSE behave differently. XLM-

R loses much of its global LDND alignment after continued training, dropping from .244 to .055 on the full matrix, while its dialect-external score rises from .261 to .382. LaBSE is the clearest negative case: both zero-shot and continued training versions remain near zero on the full LDND matrix, consistent with the language-signal suppression expected from its translation-alignment objective. Full scores and CIs are reported in Table 4.

5 Related Work

Italian dialect NLP has concentrated on resource description and identification. Ramponi (2024) surveys the language varieties of Italy and analyses their position within current NLP practice. Aeppli et al. (2022) introduce the VarDial ITDI shared task. Signoroni and Rychlý (2026) benchmark LLMs on Italian-dialect translation and report that current models struggle with these low-resource varieties.

Embedding similarity across languages has been studied independently. Choenni and Shutova (2022) probe multilingual sentence encoders against typological features. Dunn (2023) shows that variety-specific signal in embedding spaces is robust to baseline training instability. We extend this line to the low-resource set of Italian varieties: we combine Wikipedia, FLORES+, and OLDI Seed (used in isolation by prior work) with multi-family representation comparison and multi-gold-standard distance correlation.

6 Conclusion

We compared six families of embedding methods to investigate how low-resource Italo-Romance dialects behave in embedding space. Surface methods outperform large pretrained encoders; only CANINE’s tokenizer-free character-level architecture stands out as a promising direction for low-resource NLP. Beyond the four macro-families, the signal is delicate: the most stable findings are Sardinian’s pull toward Iberian standards and the broader geographic gradient (continental at the centre, islands at the periphery). Lexicostatistical alignment confirms the dominance of surface character methods; text normalization barely affects the linguistic signal. Our broader observation is that cross-method agreement gives the strongest signal: individual methods reflect one aspect of similarity more than others, while convergence across architectures points to robust linguistic structure.

Limitations

The main limitation of this work is also what makes the task interesting: **dialects are intrinsically hard to model**. For computational purposes, we treat labels such as “Venetian”, “Lombard”, or “Sicilian” as single varieties, but this is a strong simplification. Within the same labelled variety, speakers may use different lexical choices, structures, pronunciations, and writing conventions to express the same meaning. This issue is central to NLP for the language varieties of Italy, where a speaker-centric view is needed to avoid flattening internal variation.

We observed this problem directly during the manual review of Venetian data. Three of the five authors speak Venetian varieties (from Padua, Venice, and Verona), so they could directly assess part of the corpus. In many cases, they agreed on the formulation, structure, and meaning of a sentence. In other cases, however, their judgments diverged completely: each would have expressed the same content in a different local form. This does not simply reflect annotation disagreement. It reflects the fact that a dialectal label can group together several local systems that are not interchangeable.

This complexity also affects the datasets. Our selection criteria in Section 2 make the comparison possible, but they do not make the data ideal. Wikipedia is used because it is the only source that provides enough text for several Italian dialectal varieties; other resources are too sparse, incomplete, or difficult to use at the required scale. However, dialect Wikis remain problematic for this task. They contain bot-generated pages, templated text, repeated structures, and editorial conventions. More importantly, they often represent a single written and partly artificial version of a dialect, rather than the full range of local forms that speakers actually use. Our cleaning, filtering, and manual checks reduce noise, but they cannot turn Wikipedia into a fully representative dialect corpus.

Weak coverage is another consequence of the same constraint. We focus on six Italo-Romance varieties because Wikipedia, FLORES+, and OLDI Seed jointly provide enough comparable data for those only. Many relevant local, minority, and contact varieties spoken in Italy fall outside this inventory. Arabic and Greek, included in the initial proposal, are also absent: their non-Latin scripts require a separate preprocessing pipeline incom-

patible with the Latin-based normalization applied across the rest of the variety set. A broader catalogue would strengthen the analysis but requires new data collection and careful validation.

Finally, the complexity of the phenomenon requires stronger linguistic support. A linguist or dialectologist could help define variety boundaries more carefully, validate the data, and interpret model outputs with a more critical eye. The evaluation harness introduced here could also support a broader study of how LLMs behave in the presence of low-resource dialects; a preliminary zero-shot translation experiment is included in Appendix E for the interested reader. Future work should therefore combine **larger, purpose-built dialect corpora** with expert validation and a more speaker-centric approach, so that computational similarity can be evaluated against a more realistic account of dialect variation.

References

- Noëmi Aeppli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. [Findings of the VarDial evaluation campaign 2022](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Francesco Avolio. 2011. [Siciliani, calabresi e salentini, dialetti](#). In *Enciclopedia dell’Italiano*. Treccani. Accessed: 2026-04-29.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Giovanni Bonfadini. 2010. [Lombardi, dialetti](#). In *Enciclopedia dell’Italiano*. Treccani. Accessed: 2026-04-29.
- Giacomo Camposampiero, Quynh Anh Nguyen, and Francesco Di Stefano. 2022. [The curious case of logistic regression for Italian languages and dialects identification](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 86–98, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Rochelle Choenni and Ekaterina Shutova. 2022. [Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology](#). *Computational Linguistics*, 48(3):635–672.
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. [CANINE: Pre-training an efficient](#)

- tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Antonietta Dettori. 2011. [Sardi, dialetti](#). In *Enciclopedia dell’Italiano*. Treccani. Accessed: 2026-04-29.
- Jonathan Dunn. 2023. [Variation and instability in dialect-based embedding spaces](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 67–77, Dubrovnik, Croatia. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic bert sentence embedding](#). *Preprint*, arXiv:2007.01852.
- Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. [Italian language and dialect identification and regional French variety detection using adaptive naive Bayes](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 119–129, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *arXiv preprint arXiv:1301.3781*.
- Christopher Moseley, editor. 2010. *Atlas of the World’s Languages in Danger*, 3 edition. UNESCO Publishing, Paris. Third edition, entirely revised, enlarged and updated. Cartography by Alexandre Nicolas.
- NLLB Team. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630:841–846.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Alan Ramponi. 2024. [Language varieties of Italy: Technology challenges and opportunities](#). *Transactions of the Association for Computational Linguistics*, 12:19–38.
- Lena Sadovski-Kornprobst. 2021. [Multilingualism in venetian dalmatia: Studying languages and orality in written administrative documents from split \(fifteenth/sixteenth centuries\)](#). *Mediterranean Historical Review*, 36(2):217–236.
- 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.
- Marco Tamburelli and Lissander Brasca. 2018. [Revisiting the classification of gallo-italic: a dialectometric approach](#). *Digital Scholarship in the Humanities*.
- Fiorenzo Toso. 2010. [Liguri, dialetti](#). In *Enciclopedia dell’Italiano*. Treccani. Accessed: 2026-04-29.
- Flavia Ursini. 2011. [Veneti, dialetti](#). In *Enciclopedia dell’Italiano*. Treccani. Accessed: 2026-04-29.
- Laura Vanelli. 2010. [Friulani, dialetti](#). In *Enciclopedia dell’Italiano*. Treccani. Accessed: 2026-04-29.
- Federico Vicario. 2011. [Il friulano. cenni di storia linguistica, grammatica e lessico](#). In Raffaella Bombi and Vincenzo Orioles, editors, *Nuovi valori dell’italianità nel mondo. Tra identità e imprenditorialità*, pages 177–195. Forum, Udine.
- Maurizio Viridis. 2023. [Dinamiche linguistiche nella lunga età sardo-iberica](#). *RiMe. Rivista dell’Istituto di Storia dell’Europa Mediterranea*, 13/II n.s.
- Søren Wichmann, Cecil H. Brown, Eric W. Holman, Matthew S. Dryer, and Qibin Ran. 2025. [The ASJP database](#). Version 21.
- Søren Wichmann, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. [Evaluating linguistic distance measures](#). *Physica A: Statistical Mechanics and its Applications*, 389(17):3632–3639.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *ArXiv*, abs/1904.09675.

A Variety registry

Group	Variety (ISO)	Wiki	OLDI	FLORES+
Dialects	Friulian (fur)	22,256	5,167	1,827
	Ligurian (lij)	49,058	5,167	1,827
	Lombard (lmo)*	94,833	5,167	1,827
	Sardinian (sc)	60,948	5,167	1,827
	Sicilian (scn)	75,047	5,167	1,827
	Venetian (vec)*	94,833	5,167	1,827
Romance	Italian (ita), Spanish (spa)	100,000	—	1,827
	French (fra), Catalan (cat)	100,000	—	1,827
	Portuguese (por), Occitan (oci)	100,000	—	1,827
Germanic	German (deu), English (eng)	100,000	—	1,827
Slavic	Slovenian (slv), Croatian (hrv)	100,000	—	1,827
Uralic	Hungarian (hun)	100,000	—	1,827

Table 1: The 17 varieties used in our study, grouped by family, with per-variety sentence counts used for training (Wiki and OLDI) and for evaluation (FLORES+). For dialects, Wiki+OLDI is capped at 100k sentences: Lombard* and Venetian* are therefore sampled to 94,833 Wiki sentences, since $94,833 + 5,167 \text{ OLDI} = 100,000$, while the other dialects are kept in full. Comparison-language Wikipedias are sampled to 100k; FLORES+ is parallel across all varieties.

B Embedding-space visualisations and silhouette scores

Figure 1 shows the full 17×17 pairwise cosine similarity matrix for fastText, illustrating the results discussed in 4.1. Figure 2 depicts the UMAP results for XLM-R zero-shot. Figure 3 provides a visualization of the catastrophic forgetting in LaBSE model. Table 2 presents Silhouette scores for the different models.

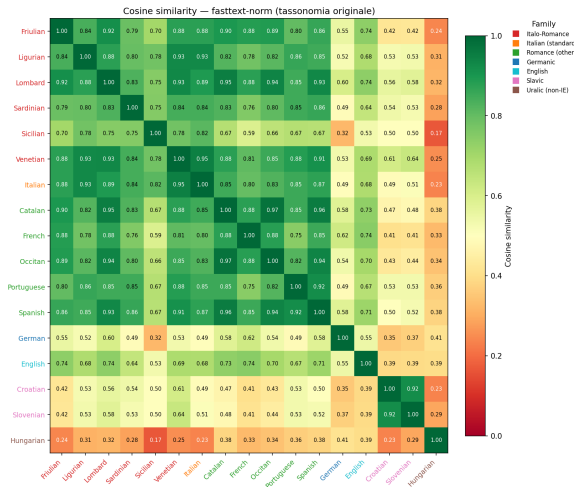


Figure 1: Cosine similarity matrix for **FastText** (normalised) ($\rho_{\text{full}} = .874$, $\text{sil}_{\text{macro4}} = +0.50$).

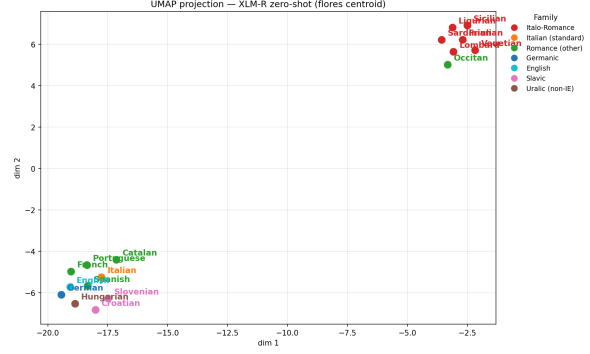


Figure 2: UMAP projection of XLM-R zero-shot variety centroids (FLORES+). The model separates dialects (upper right) from all standards (lower left) into two tight, internally undifferentiated blobs. Italian sits firmly in the standards cluster, far from the dialects it is lexically closest to. The apparent family silhouette (+0.14) reflects this coarse two-way split rather than genuine typological organisation.

Method	Variant	$\text{sil}_{\text{macro4}}$	Kept
CANINE-MLM	native	+0.57	✓
fastText	normalized	+0.50	✓
fastText	native	+0.41	✓
Word2Vec	normalized	+0.39	✓
CANINE zero-shot	native	+0.32	
Word2Vec	native	+0.32	✓
TF-IDF char	normalized	+0.24	✓
TF-IDF char	native	+0.21	✓
TF-IDF word	normalized	+0.15	✓
TF-IDF word	native	+0.14	✓
XLM-R zero-shot	native	+0.14	
XLM-R MLM	native	-0.01	
LaBSE MNRL	native	-0.21	
LaBSE zero-shot	native	-0.26	

Table 2: Silhouette score over four typological macro-families (Romance, Germanic, Slavic, Uralic) on FLORES+ variety centroids. Range $[-1, +1]$; higher is better. "Kept" marks the nine methods we retain for downstream analysis. Pretrained zero-shot encoders are excluded because they were not exposed to dialect text in pretraining; XLM-R MLM and LaBSE MNRL fail to recover the macro-family structure (negative silhouette).

C Historical influence

We construct a historical influence reference table to encode expected contact relations between each dialect and the comparison languages in our inventory. The table is built *a priori*: one author annotated the relations manually, independently of the experimental results. The annotation was based on dialectological and historical sources, especially encyclopedia entries (Vanelli, 2010; Toso, 2010; Bonfadini, 2010; Dettori, 2011; Avolio, 2011; Ursini, 2011) and studies of specific dialect areas and contact settings (Vicario, 2011; Viridis, 2023; Tamburelli and Brasca, 2018; Sadovski-Kornprobst, 2021). This step requires interpretation: the rel-

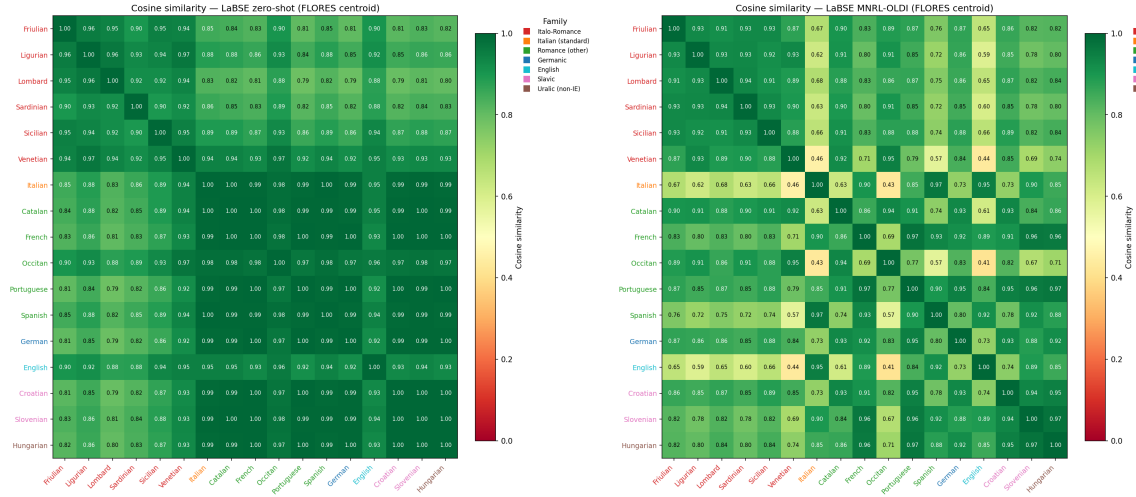


Figure 3: Cosine similarity matrices for **LaBSE zero-shot** (left) and **continued training LaBSE-MNRL** (right) models. In the zero-shot setting, translation-alignment pre-training compresses the 17-variety inventory into a near-uniform high-similarity band (0.79-1.00). After continued training on Italian-dialect pairs, dialect-dialect similarities increase, but cross-lingual alignment collapses: Italian drops to 0.46–0.68 against the dialects, and cross-family languages become similarly distant. This catastrophic forgetting removes meaningful dialect↔external signal, leaving the continued training LaBSE-MNRL worse than zero-shot on LDND Spearman correlation and Mantel significance.

evant relations are extracted from the sources and mapped onto the comparison languages (excluding Italian) available in our inventory. For each dialect, the table reports up to three historically relevant contact languages, as shown in Table 3. The selection criteria are documented contact, political history, settlement, and linguistic influence. Because the table is restricted to the languages included in our final dataset, it does not cover every historically relevant contact language discussed in the literature. The resulting table is therefore not a complete model of language history but rather a reference structure that encodes historically motivated expectations for our specific catalogue of varieties and comparison languages.

Dialect	Influence I	Influence II	Influence III
fur	slv	deu	hrv
lij	oci	spa	fra
lmo	oci	deu	spa
sc	spa	cat	
scn	spa	fra	cat
vec	hrv	slv	

Table 3: Historical influence reference table. The order does not indicate relative importance.

D Lexicostatistical gold

D.1 Construction

Our lexicostatistical gold (Figure 4) is a 17×17 distance matrix built from Wiktionary Swadesh-207

wordlists, i.e. 207 basic meanings translated across the varieties. For each pair of varieties, we average the Levenshtein distances between corresponding forms, after converting them to ASJPcode to reduce orthographic effects (Wichmann et al., 2025). We use LDND as our main score (Wichmann et al., 2010): it normalises the average Levenshtein distance by the distance expected between unrelated forms.

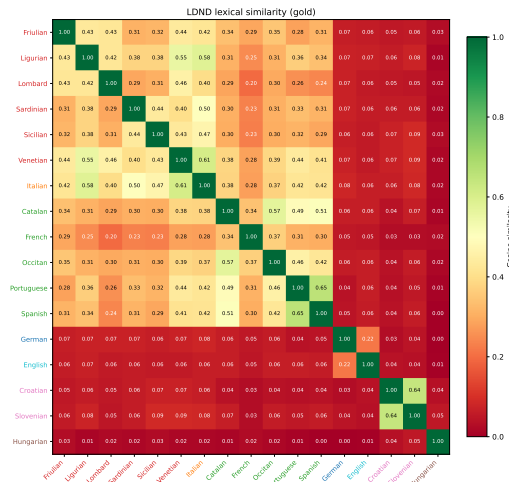


Figure 4: LDND distance heatmap via Swadesh list, 17×17 . Lighter shades indicate smaller distances.

D.2 Lexicostatistical comparison table

Table 4 shows the results of the comparisons.

Model	ρ_{full} [95% CI]	$\rho_{d \leftrightarrow e}$ [95% CI]
TF-IDF char (nat)	.905 [.903, .908]	.851 [.846, .857]
TF-IDF char (norm)	.881 [.878, .886]	.813 [.807, .821]
fastText (norm)	.874 [.872, .875]	.812 [.808, .813]
fastText (nat)	.867 [.865, .869]	.776 [.770, .781]
word2vec (norm)	.851 [.850, .854]	.778 [.773, .783]
word2vec (nat)	.841 [.839, .843]	.703 [.697, .710]
TF-IDF word (nat)	.795 [.791, .800]	.720 [.707, .727]
TF-IDF word (norm)	.788 [.785, .794]	.734 [.727, .744]
CANINE ft	.725 [.722, .729]	.803 [.797, .806]
CANINE zs	.664 [.658, .671]	.700 [.693, .713]
XLM-R zs	.244 [.239, .249]	.261 [.247, .275]
XLM-R ft	.055 [.048, .060] [†]	.382 [.371, .394] [†]
LaBSE ft	-.021 [-.044, .002] [†]	.327 [.272, .367]
LaBSE zs	-.070 [-.082, -.062] [†]	.266 [.245, .286] [†]

Table 4: Spearman ρ of each model’s FLORES-centroid distance matrix against the lexicostatistical LDND gold, with bootstrap 95% CIs. ρ_{full} : all 136 upper-triangle pairs; $\rho_{d \leftrightarrow e}$: 60-pair dialect $\times$ external-non-Italian block. Symbol [†] marks Mantel $p > 0.05$.

D.3 Bootstrap and permutation methodology

For bootstrap reliability, we resample the 1,827 FLORES sentence embeddings within each variety, recompute the variety centroids and the 17×17 distance matrix, and repeat this $B=1000$ times to obtain 95% confidence intervals for each Spearman ρ . Moreover, we use a two-sided Mantel permutation test: we randomly relabel the varieties in the model distance matrix $B=10000$ times and ask whether the observed correlation is more extreme (in either direction) than correlations obtained by chance. The results show that lexicostatistical correlations are consistently reliable, while some encoder scores, especially for XLM-R and LaBSE in Table 4, are stable under bootstrap resampling but not significant under Mantel permutation.

E LLM translation experiment

If embedding similarity captures genuine linguistic proximity, a natural prediction follows: translating into a dialect from a linguistically closer source language should yield better output.

Setup. We use ChatGPT (GPT-5.5) and Gemini 3.0. Each model receives a zero-shot prompt instructing it to translate a given sentence from the source language into the target dialect, with no additional context. Source languages are English, French, Italian, and Spanish; target dialects are the six Italo-Romance varieties from Appendix A. We select 300 sentences from the FLORES+ evaluation set and score each output against the FLORES+ reference translation with chrF++ (Popović, 2015), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2019).

Significance testing. For each dialect and model, we run pairwise bootstrap tests ($B = 1000$ resamples) between all six source-language pairs, applying Bonferroni correction for the six comparisons per metric. We report p -values and 95% confidence intervals alongside point estimates.

Results. Table 5 reports the three metrics for all 48 conditions. BERTScore F1 is consistently high (0.87-0.92) across all conditions and broadly tracks the chrF++ ranking, confirming metric-level consistency. ChatGPT scores are uniformly higher than Gemini and show lower variance: the worst-to-best spread within a dialect rarely exceeds 4 chrF++ points, while Gemini’s spread reaches 12+ points for Ligurian and Venetian. For ChatGPT, Italian is the best or co-best source on every dialect. For Gemini, it remains competitive on Sicilian and Sardinian, while English leads for Friulian and Venetian, Spanish for Ligurian, French for Lombard.

Across both models, Italian and English are the strongest sources; French and Spanish yield no consistent advantage despite their Romance affinity. Italian matching English is linguistically expected but notable given LLMs’ general English-input advantage, and suggests that training-data coverage shapes quality at least as much as linguistic proximity.

Dialect	Model	chrF++				BLEU				BERTScore F1			
		EN	FR	IT	ES	EN	FR	IT	ES	EN	FR	IT	ES
Friulian	ChatGPT	61.42	58.18 [†]	61.10	58.39	34.53	30.42 [†]	34.47	30.81	91.7	91.0 [†]	91.5	91.0
	Gemini	59.68	50.76	44.83 [†]	56.14	31.17	24.18	20.51 [†]	27.17	91.3	90.0	89.5 [†]	90.6
Ligurian	ChatGPT	47.77	46.10	48.22	45.63 [†]	14.81	13.88 [†]	15.48	13.92	89.2	88.9	89.2	88.7 [†]
	Gemini	39.67	34.61 [†]	37.82	46.67	9.07	4.46 [†]	7.54	15.36	87.9	86.7 [†]	87.3	89.1
Lombard	ChatGPT	42.02	40.94	43.41	40.94 [†]	8.74	8.57 [†]	10.61	8.91	88.2	88.0	88.4	88.0 [†]
	Gemini	39.61	40.38	30.37	29.79 [†]	8.25	8.21	5.78	5.38 [†]	88.1	88.2	86.8	86.5 [†]
Sicilian	ChatGPT	48.72	46.78 [†]	49.67	47.46	14.26	13.07 [†]	15.10	14.33	89.0	88.6 [†]	89.1	88.8
	Gemini	46.67	37.98 [†]	47.70	44.20	12.70	9.02 [†]	12.39	10.47	88.7	87.6 [†]	88.5	88.1
Sardinian	ChatGPT	54.11	52.45	56.07	51.88 [†]	22.11	20.82	25.51	20.70 [†]	90.3	89.9	90.6	89.7 [†]
	Gemini	56.34	44.77 [†]	53.69	52.44	26.87	16.90 [†]	21.60	22.94	90.7	89.1 [†]	90.3	89.9
Venetian	ChatGPT	49.70	48.21	49.74	48.04 [†]	14.83	14.59	14.67	14.05	88.9	89.0	89.0	88.8 [†]
	Gemini	46.20	37.06	32.54 [†]	41.55	14.26	10.04	9.03 [†]	11.06	88.7	87.4	87.2 [†]	88.0

Table 5: Translation quality by source language, target dialect, and LLM. **Bold**: best score per row. [†]: significantly worst under pairwise bootstrap with Bonferroni correction; lowest value for BERTScore F1.