Parallel Sentence Mining Without Direct Parallel Sentences: Scaling to Low-resource Languages

Anonymous ACL submission

Abstract

Parallel sentence mining aims to find translation pairs from comparable monolingual corpora and creates a valuable dataset for downstream tasks. Yet, it has been mostly considered for high-resource language pairs, and endeavours for low-resource languages still need a significant number of parallel sentences for success, although it remains challenging to gather them. In this study, we remove this requirement by only relying on *monolingual* data in the low-resource language or parallel sentences in related but different languages (i.e., *not direct*) to obtain a sentence representation. We show when three techniques (isotropy enhancement of the embeddings, contrastive learning, and knowledge distillation) can improve multilingual sentence representation under these settings. We extend a benchmark to cover sixteen language pairs, where eight low-resource languages from three families are paired with two languages to control within-pair language distance. While all three methods (notably distillation) improve representation quality by tackling the cross-lingual misalignment, monolingual pre-training and language proximity are essential factors to explain better performance. Mining can still be improved without direct parallel corpora even for harder pairs that suffer from script and language cluster phenomena.

1 Introduction

Parallel corpora remain a crucial resource for multilingual Natural Language Processing (NLP), as they still help cross-lingual performance of language models (Xu et al., 2024; Lin et al., 2025). Since they are costly to obtain, parallel sentence mining aims to automatically find translation pairs from comparable monolingual corpora. Its performance notably relies on the quality of the sentence representations in the language. Low-resource languages (LRLs) suffer here twice: not only are parallel corpora scarcely available or small, their representations in widely used models are poorer, leading to worse mining quality. While high-resource languages (and pairs) were extensively studied (e.g., the BUCC benchmark, Zweigenbaum et al., 2017) in general, unseen LRLs either required a large number of *parallel* sentences (>40k) to support an unseen language (Kvapilíková and Bojar, 2023; Tan et al., 2023) or did not achieve usable mining quality without them (Okabe et al., 2025).

In this work, we therefore focus on parallel sentence mining for low-resource languages by removing the need for **direct** parallel sentences for the LRL (i.e., only allowing high(er)-resource pairs). To do so, we extend an existing benchmark of parallel sentence mining (Okabe et al., 2025) to cover eight LRLs (Occitan, Upper and Lower Sorbian, Chuvash, Corsican, Mingrelian, Gilaki, and Mazandarani; four scripts), of which five are endangered. Each LRL is paired with a higher-resourced target language and its English translation (i.e., 16 language pairs). Our benchmark improves the original by introducing more difficult and diverse pairs with controlled language distances. We then consider three types of techniques to obtain better multilingual sentence representation *without* ***direct*** parallel sentences: isotropy-enhancing transformations, contrastive learning, and knowledge distillation. We highlight how they impact mining quality when combined with LRL monolingual pre-training and under two language pair distances per LRL, as the target sides are translated.

Our contributions are as follows: (i) we expand and release a harder parallel sentence mining benchmark to cover 8 typologically diverse source LRLs with 2 targets, (ii) we empirically show when language representation can be improved *without* any *direct* parallel sentences, i.e. featuring a LRL, and (iii) we further analyse each method per language distance and perform four realistic case studies.[1]

[1] Anonymous link for the benchmark and creation code.

2 Related works

Performance of parallel sentence mining is intertwined with the quality of multilingual sentence representation, which explains why mined corpora for LRLs tend to be noisier. Language models have better cross-lingual representation when trained on parallel corpora. Yet, the need for such resources is a major constraint, especially for LRLs.

Cross-lingual alignment can be improved without requiring any specific pre-training by using post-processing methods. Hämmerl et al. (2023) consider ZCA whitening (Huang et al., 2021) and cluster-based isotropy enhancement (Rajaee and Pilehvar, 2021), where both tackle the inherent anisotropy (i.e., clustered representation) in multilingual language models.

Relatedly, contrastive learning (Chopra et al., 2005; Hadsell et al., 2006) has been applied to improve English sentence representation, such as in SimCSE (Gao et al., 2021), without requiring parallel sentences. This framework can be supervised using the Natural Language Inference (NLI) sentence relationship. A positive and a negative example can be compared for a given sentence to bring similar sentences closer together. mSimCSE (Wang et al., 2022) extended it to multilingual sentence embeddings by using (English or multilingual) NLI datasets for better cross-lingual representations overall. This method proved to be more efficient than similar methods without contrastive learning, such as Goswami et al. (2021). For LRLs, such methods remain more accessible as they do not require parallel sentences for training at all. This contrasts with established sentence encoders, such as LASER (Artetxe and Schwenk, 2019b) or LaBSE (Feng et al., 2022), trained with large parallel corpora.

Knowledge distillation can also help extend a model to other languages (Reimers and Gurevych, 2020). However, it has mostly been used with *direct* parallel sentences in the studied language pair (Tan et al., 2023), limiting its application to better-resourced languages. Still, for low-resource languages, Heffernan et al. (2022) notably improve LASER using distillation with a multilingual teacher and monolingual student (LASER3). The training relies on both parallel corpora and monolingual sentences of the source LRL. Tan et al. (2023) extend this work by using contrastive learning with large parallel corpora (>40k sentences). For a given sentence pair, the English translation is considered as a positive example. We do not consider similar approaches because of the small size of the available parallel sentences for some of our language pairs (e.g., Mingrelian–English), and instead only use parallel sentences from *related* (i.e., not *direct*) language pairs (e.g., Georgian–English).

Finally, few works focused on parallel sentence mining for LRLs. Okabe et al. (2025) released a benchmark for this very task, BELOPSEM, but covering only three language pairs. Anisotropy enhancement notably led to consistent improvement but failed to outperform LaBSE for two of the pairs. We build upon this benchmark, introducing more linguistic diversity (language branches, families, and scripts) and show better mining quality, without using *direct* parallel sentences.

3 Languages and corpora

3.1 Language pairing

For notation purposes, we will refer to the *low-resource* language of a pair as 'source', and the *higher-resourced* language as 'target', reflecting a larger availability of sentences for the latter. We study the following eight *source* languages: Occitan, Upper Sorbian, Lower Sorbian, Chuvash, Corsican, Mingrelian, Gilaki, and Mazandarani. They are all classified as 'scraping-by' (1 on a scale from 0 to 5) at most in terms of available data in the taxonomy of Joshi et al. (2020) and can hence be considered as LRLs. Besides, Ethnologue (Eberhard et al., 2025) considers all but three languages as endangered, while the UNESCO (2010) lists five of the languages as 'definitely endangered'. Appendix A reports the exact status for all eight languages. We pair each source language with a well (or better)-resourced language (classified as 3 or above in the taxonomy from Joshi et al. (2020)). We choose them from a geographical and cultural standpoint and refer to them as *main target* languages (§3.2). We also pair all eight languages with English as a second and farther target (§3.3).

3.2 Main language pairs

Occitan–Spanish (OCI–ES) Occitan (ISO code: `oci`) is a Romance language spoken in southern France, Spain, and Italy. Both languages in the pair are part of the same branch in the language family (Ibero-Romance) and are written in the Latin script. This is the closest language pair in our benchmark.

Upper Sorbian and Lower Sorbian–German

Anonymous link for the best performing model.

(HSB–DE and DSB–DE) Upper Sorbian (hsb) and Lower Sorbian (dsb) are two Slavic languages spoken in Germany. They constitute the Sorbian branch and are related to Czech or Polish, respectively. The pair features languages from two different branches but written in the Latin script.

Chuvash–Russian (CHV–RU) Chuvash (chv) is a Turkic language spoken in the Chuvash Republic in Russia. It is quite distant from other related languages, as it belongs to its own branch. Both languages in the pair use the Cyrillic script but belong to different language families.

Corsican–French (COS–FR) Corsican (cos) is a language spoken on the islands of Corsica in France and Sardinia in Italy. The language pair is hence close, as both are from the Romance branch and written in the Latin script.

Mingrelian–Georgian (XMF–KA) Mingrelian (xmf) is a Kartvelian language (where Georgian also belongs), spoken in Western Georgia. Both languages are written with the Georgian script.

Gilaki and Mazandarani–Persian (GLK–FA and MZN–FA) Gilaki (glk) and Mazandarani (mzn) are two Caspian languages from the Indo-Iranian branch, spoken in northern Iran. They hence belong to the larger Iranian branch as Persian, and all three are written in the (Perso-)Arabic script.

3.3 Pairing with English

We paired all eight source languages with English to introduce a different (and larger) language distance in the pairs. As it is a Germanic language written in the Latin script, close pairs, such as Mingrelian–Georgian, will then have a more challenging counterpart from the script and language family perspective. We also note that English is the language with the most resources, and language models have been extensively trained on it. Appendix B classifies the language pair difficulty (from 1 to 4).

3.4 Corpus creation

We create synthetic corpora for parallel sentence mining, following the BUCC Shared Task methodology (Zweigenbaum et al., 2017) used in BELOPSEM (Okabe et al., 2025). We mix gold parallel sentence pairs into monolingual corpora in each language, and the goal is to retrieve them back. We insist here on the importance of the quality of the parallel corpus in curating our benchmark and language pairs, limiting the set of studied languages.

English side For the first five languages, we have relied on machine translation using Google Translate.[2] As we translate from high-resource languages (Spanish, German, Russian, and French) and into English, we deem the output to be of decent quality.[3] On the other hand, the parallel sentences of the last three languages come from a three-way parallel corpus with English. We hence used actual parallel corpora between the main target language and English (e.g., Georgian–English) and treated them separately as a 'monolingual' corpus to create their benchmarks (XMF–KA and EN).

Data source For the three language pairs in BELOPSEM (OCI–ES, HSB–DE, and CHV–RU), we use the same *original* datasets. The HSB–DE, DSB–DE, and CHV-RU pairs were considered in WMT Shared Tasks, providing both parallel and monolingual source sentences. Other monolingual sentences were taken from the Leipzig corpora (Goldhahn et al., 2012). For COS–FR and OCI–ES, we use parallel corpora from OPUS (Tiedemann, 2012). For the last three source languages, we use three-way parallel corpora with English as the third language. For their target and English monolingual corpora, we thus use an actual parallel corpus from the OPUS collection instead of machine translation. While creating our BUCC-style corpus, we ensure that the shuffling and position of the parallel sentences are identical across the two variants of the dataset (main and English). Appendix D lists the exact corpora used.

Data preparation We pre-process both monolingual and parallel sentences using GlotLID (Kargaran et al., 2023) for language identification. For parallel sentences, we also filter out sentence pairs with a large difference in length and with a high proportion of overlapping words. After pre-processing, we inject the gold sentence pairs into the monolingual corpora. Then, we split each created corpus into training and test sets following a 25:75 ratio. The smaller training set is used to find the best model (and parameter setting) to mine sentences from a larger corpus (i.e., test set). Table 1 presents the datasets of the sixteen language pairs. We note that some datasets (e.g., XMF–KA or COS–FR) remain small due to the original parallel dataset size.

[2] With the googletrans library.
[3] Appendix C verifies the MT quality of Google Translate.

lang. pair	train			test		
	source	target	par.	source	target	par.
OCI–ES/EN*	5,981	6,828	331	17,923	20,489	996
HSB–DE/EN*	10,948	13,029	546	32,850	39,091	1,642
DSB–DE/EN*	9,996	11,899	499	29,992	35,699	1,499
CHV–RU/EN*	6,716	6,957	372	20,152	20,877	1,120
COS–FR/EN*	2,749	3,276	136	8,260	9,835	413
XMF–KA/EN	1,023	1,218	50	3,075	3,659	153
GLK–FA/EN	2,161	2,568	124	6,486	7,708	374
MZN–FA/EN	4,248	5,056	211	12,750	15,172	637

Table 1: Size of the training and test datasets, where source and target include the injected parallel sentences (par.). * denotes the use of MT for English.

4 Sentence representation

4.1 Baseline multilingual models

We use the standard approach of averaging the word embeddings[4] to get the sentence-level representation from a language model. We mainly compare two models: XLM-RoBERTa (base) or XLM-R (Conneau et al., 2020) and Glot500-m (Imani et al., 2023), which we will denote G500. The latter extends the former towards more than 500 languages, with a particular focus on low-resource languages, which will show the impact of dedicated pre-training on *monolingual* data only. We also compare our systems with LaBSE (Feng et al., 2022), a state-of-the-art sentence encoder which was trained using a large amount of *parallel* sentences. Appendix E summarises the language coverage of the three models. Importantly, G500 was pre-trained on six of our source LRLs (all except dsb and glk). The Appendix contains results for two more recent language models.[5] We only consider models with a vector dimension of 768.

4.2 Isotropy improving transformations

Multilingual sentence embeddings can be improved by tackling the anisotropy of the vectors in the multilingual space. Without relying on any additional data, operations such as whitening (Huang et al., 2021) can transform embeddings from language models before downstream usage. To improve mean-pooled representations for low-resourced language pairs on BELOPSEM, Okabe et al. (2025)[6] applied a cluster-based isotropy enhancement or CBIE (Rajaee and Pilehvar, 2021). This technique first clusters the vectors and then uses a Principal Component Analysis to remove the top 12 principal components, as in Hämmerl et al. (2023).[7] We also apply it to our sentence-level representation from G500. Furthermore, we consider whitening as an alternative post-processing method, which was not used on BELOPSEM. These settings will be called G500+{CBIE | whiten}, respectively.

4.3 Contrastive learning

For contrastive learning, we use the mSimCSE multilingual contrastive learning framework (Wang et al., 2022). We consider two types of datasets to train the models: NLI datasets and parallel sentences between English and the main target languages (i.e., no *source* language is involved).

We consider two NLI datasets: an English-only dataset (Conneau et al., 2017) and the multilingual XNLI dataset (Conneau et al., 2018b) to foster cross-lingual transfer. In this work, we apply both batch contrastive learning methods to XLM-R for comparison: we denote the former mSC-XLM-en and the latter mSC-XLM-multi. The first approach led to notable improvement on sentence mining in Wang et al. (2022), while it only used (monolingual) English sentences. The second approach led to further gains on retrieval tasks. Yet, these results concerned high(er)-resourced languages.

Finally, we perform contrastive learning with parallel data as positive pairs and a random sentence in another language as a hard negative (mSC-XLM-tr). We use the Parallel Sentences dataset collection (Reimers and Gurevych, 2019) for the six *target* languages paired with English (e.g., Spanish–English). In this way, we still do *not* use ***direct*** parallel sentences with the source LRL.

Changing the base model Additionally, we switch the base model from XLM-R to Glot500-m and carry out contrastive learning on the same three datasets. These newly trained models will be denoted as mSC-G500-{en | multi | tr} variants. These models enable us to identify the influence of monolingual training for the source languages.

4.4 Knowledge distillation

The third strategy we consider is a knowledge distillation approach. We choose LaBSE as the teacher model, based on its overall mining performance, and Glot500-m as the main student model, due to

[4] We use the 8th layer for all language models, following (Imani et al., 2023; Hämmerl et al., 2023).

[5] mmBERT (base) (Marone et al., 2025) and multilingual-e5-base (Wang et al., 2024) both underperformed their respective counterparts: G500 and LaBSE.

[6] We do not consider their alignment post-processing, as it only filters *after* mining and led to minimal improvement.

[7] https://github.com/KathyHaem/outliers.

its better language coverage. We use the Sentence Transformer framework for distillation (Reimers and Gurevych, 2020). We train the model on *target* language pairs: English paired with the six target languages (de, es, fa, fr, ka, ru). We use the same parallel dataset as for contrastive learning, with a maximum of 100k sentence pairs per language. We call the model G500-LaBSE. To understand the impact of monolingual pre-training, we also use XLM-R as a student model (XLM-LaBSE).

5 Experimental setting

5.1 Mining pipeline

We use an established mining pipeline (Okabe et al., 2025), which updated the system of Hangya and Fraser (2019) with contextual embeddings. It consists of two steps: first, it converts each sentence into embeddings in the same multilingual space using the models previously described. Then, sentences are compared according to the CSLS similarity metric (Artetxe and Schwenk, 2019a):[8]

$$\mathrm{CSLS}(x, y) = 2\cos(x, y) - \sum_{z \in \mathrm{NN}_k(x)} \frac{\cos(x, z)}{k} - \sum_{z \in \mathrm{NN}_k(y)} \frac{\cos(y, z)}{k}, \quad (1)$$

where $\mathrm{NN}_k(x)$ indicates all k-nearest neighbours of vector x to avoid the hubness problem (Dinu et al., 2015). We use $k = 20$ in our experiments. We select our similarity threshold based on the training set performance. The experiments are mainly evaluated using the F-score. Appendix G lists experimental details for reproducibility.

5.2 Main mining results

The left side of the language columns in Table 2 displays the sentence mining results for the eight *main* language pairs (i.e., from Section 3.2).

Baseline performance We first notice that monolingual pre-training on the language is crucial when using averaged word embeddings for sentence representation. Glot500-m consistently outperforms XLM-R on seen languages, while it similarly struggles with the unseen dsb and glk. The lower score for mzn could be due to the fewer pre-training sentences (74k) than for the other languages, which all comprise more than 100k sentences. Both approaches are also worse than the sentence encoder LaBSE, except for chv and xmf for G500. Explicit training on parallel sentences with a contrastive objective helps LaBSE represent unseen but related languages, whereas mean-pooled embeddings suffer from structural cross-lingual misalignment.

CBIE and whitening Tackling the inherent anisotropy of the language representation with post-processing leads to consistent improvement for all languages for G500. This closes the gap with LaBSE and widens it for chv and xmf. We note here that the least performing cases are still for DSB–DE and GLK–FA, which constitute the two language pairs where the source language has not been seen by G500. Besides, whitening leads to higher overall scores compared to CBIE.

Contrastive learning Methods based on contrastive learning noticeably improve their base model overall. While the English NLI dataset (-en) brings uneven effect, the cross-lingual training (-multi) leads to systematically better F-scores for all main language pairs and models. Our extension using Glot500-m clearly improves several language pairs from the baseline G500 but also from mSC-XLM-en. Contrastive learning is hence more effective, when the base model has been pre-trained on the language. Besides, we see that on average, using XNLI over English NLI is worse with XLM-R (17 vs. 21), while for G500, cross-lingual NLI benefits more (44 vs. 26), showing that the competitiveness found in Wang et al. (2022) seems to mostly hold, when both languages have been seen during pre-training. Finally, using translation datasets for contrastive learning (-tr) leads to rather comparable scores to the -multi models overall, with a few exceptions. A notable difference is that using XLM-R as a base model can potentially lead to a slightly worse score, whereas we see an improvement with G500, albeit sometimes small.

Knowledge distillation When we use knowledge distillation from LaBSE, we observe improvement for both base models. However, G500 as a student model better inherits LaBSE's multilingual representation, with some drastic differences, such as +43 points for XMF–KA. This stresses again the importance of monolingual pre-training in cross-lingual alignment in yet another strategy.

Across strategies Comparing G500+CBIE and mSC-G500-tr, we observe that the latter outperforms the former, except for the two closest (and easier) language pairs of OCI–ES and COS–FR. Whitening is an even better choice in this regard

[8] Appendix F describes the similarity metric.

source	OCI		HSB		DSB		CHV		COS		XMF		GLK		MZN	
target	ES	EN	DE	EN	DE	EN	RU	EN	FR	EN	KA	EN	FA	EN	FA	EN
XLM-R	47.44	35.72	0.40	1.10	0.42	0.20	0.64	0.35	16.06	12.75	4.55	0.00	1.64	0.00	3.84	0.00
G500	79.96	62.59	21.33	14.03	2.97	1.45	39.59	27.61	49.94	38.34	41.25	2.44	4.06	0.25	9.92	0.31
G500+CBIE	92.43	87.10	48.01	33.51	20.29	14.51	48.77	34.13	69.80	65.13	57.37	8.89	19.54	1.50	43.25	4.42
G500+whiten	93.40	88.43	66.39	49.95	35.34	25.64	57.25	47.47	72.14	69.91	54.49	3.72	20.51	2.62	39.43	3.72
mSC-XLM-en	62.64	50.24	13.45	10.46	11.91	8.55	7.11	2.11	52.44	43.31	13.01	2.92	20.46	4.39	30.37	3.31
mSC-XLM-multi	61.93	39.45	18.40	5.82	11.70	3.08	3.53	1.24	47.95	27.99	7.84	1.96	21.77	1.86	24.05	0.41
mSC-XLM-tr	67.95	63.69	21.28	20.12	15.79	13.61	2.50	1.47	44.18	43.85	5.46	2.16	18.64	7.62	24.77	10.01
mSC-G500-en	73.95	34.54	22.38	4.65	11.32	2.52	36.57	5.18	51.66	26.01	56.12	5.24	32.89	4.91	47.07	5.72
mSC-G500-multi	83.24	74.83	53.34	41.10	27.02	18.42	57.58	46.68	68.93	61.44	57.54	29.91	31.77	5.23	48.50	5.32
mSC-G500-tr	86.13	81.40	63.15	55.28	27.80	24.57	64.90	55.18	69.17	60.79	62.75	30.63	33.23	18.04	49.63	21.83
LaBSE	**98.20**	**97.04**	74.87	77.43	**69.60**	**74.69**	30.85	31.35	**90.32**	**90.39**	28.47	27.62	45.49	28.68	56.64	32.63
XLM-LaBSE	96.05	95.69	43.64	44.20	39.06	40.89	24.57	23.34	69.71	70.53	18.05	21.05	44.32	26.41	58.93	31.69
G500-LaBSE	97.11	96.92	**78.31**	**77.98**	54.57	54.56	**80.68**	**79.25**	83.70	85.58	**71.11**	**59.61**	**54.96**	**32.51**	**71.71**	**39.61**
$\Delta_{(G500-L)-L}$	-1.09	-0.12	3.44	0.55	-15.03	-20.13	49.83	47.90	-6.62	-4.81	42.64	31.99	9.47	3.83	15.07	6.98

Table 2: F-scores (%) on the test set of the eight source languages paired with the main target language (left column for targets) and English (right). Results in **bold** indicate the best score. We highlight the difficulty of language pairs with varying shades of grey, following our classification based on language script and family (cf. Section B). $\Delta_{(G500-L)-L}$ indicates the difference between G500-LaBSE and LaBSE.

for closer pairs, while contrastive learning leads to more consistent improvement for harder pairs (cf. for chv, xmf, glk, and mzn). Overall, G500-LaBSE leads to the best score, with a similar resource cost to the -tr models. This suggests that specific training (contrastive learning or distillation) helps address deeper misalignment than isotropy enhancement, especially for harder pairs. CBIE or whitening, however, are easily scalable (i.e., no training) and language-independent. We also see that leveraging parallel sentences, even without the source language, remains more efficient in improving multilingual alignment than NLI datasets.

5.3 Results when paired with English

The right side of the columns in Table 2 shows the mining score when paired with the English translation of the target side. This allows for a direct comparison with the left side in a sometimes more challenging context: English can introduce a difference in script (e.g., CHV–EN for Cyrillic) and language family (XMF–KA to XMF–EN).

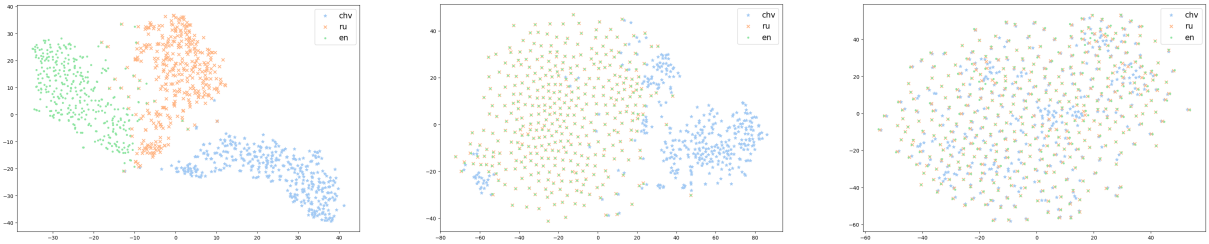
Most of the trends seen earlier still hold. Applying CBIE, whitening, contrastive learning, or knowledge distillation leads to consistent improvement over the averaged word representations from the base models. Monolingual pre-training still gives a noticeable advantage for cross-lingual alignment. Interestingly, NLI contrastive learning with only English, which is the target language here, can lead to lower scores than for the main language pair (e.g., 22 vs. 5 for mSC-G500-en in hsb), suggesting some limits to this otherwise robust approach (Wang et al., 2022). This underlines the importance of multilingual contrastive learning, whether with NLI or translation data, depending on availability.

Challenging pairs When comparing the mining scores for the same source language, they tend to be lower on average when paired with English. As the target side is identical in content, the decrease is very likely due to the larger distance between the languages. The most notable drop is for xmf, where the F-score decreases by 22 points on average for contrastive learning and 48 for CBIE. Both Caspian languages also see stark differences. This is mainly due to the source-target divergence, underlining the language or script 'cluster' phenomenon in multilingual language models (Wang et al., 2022; Liu et al., 2024). This negative effect, however, is reduced for LaBSE and distilled models, due to their better cross-lingual alignment, with relatively little difference. Overall, the script and language family barrier has not been fully overcome yet, as three of the four most difficult language pairs with English (XMF–, GLK–, and MZN–EN) significantly perform worse, showing that the three methods improved still largely *within* a language and script cluster.

6 Analyses

6.1 Language pairs for knowledge distillation

Beyond the setting of G500-LaBSE, we use two other sets of languages. First, a more comprehensive set of languages than G500-LaBSE with six ad-



(a) Glot500 (b) LaBSE (c) G500-LaBSE

Figure 1: t-SNE visualisation for three-way parallel sentences in Chuvash, Russian, and English.

ditional languages which are *related* to the *source* ones (`ca`, `cs`, `pl`, `tr`, `it`, `ku`; a total of twelve language pairs).[9] Second, we simulate a more generic and multilingual model by using five language pairs representing the United Nations languages: `ar`, `en`, `es`, `fr`, `ru`, `zh-cn`. The former, denoted G500-LaBSE-all, approximates the upper bound of distillation with cross-lingual transfer from better-resourced languages. The latter (G500-LaBSE-un) explores a more general model with widely spoken and institutional languages. For our benchmark, we note that some languages are covered by this -un set from the target side, from the script (e.g., CHV–RU or GLK–FA), but not all (e.g., XMF–KA).

	main pair	English pair
LaBSE	61.81	57.48
G500-LaBSE	74.02	65.75
G500-LaBSE-all	**78.98**	**69.72**
G500-LaBSE-un	73.40	63.66

Table 3: Averaged F-score for the three knowledge distillation settings on the 16 language pairs.

Table 3 compares the average scores of the three models (and LaBSE for reference); full results are in Appendix H. We note that providing related languages (-all) improves the mining quality noticeably. Besides, using some unrelated languages (-un setting) still gives better results than the baseline LaBSE, underlining the benefits of Glot500-m as a student model, but lags behind the other two settings. We notice lower but similar results for most language pairs, with the exceptions of `dsb`, `glk`, and `mzn`, showing low transfer for the Arabic script.

6.2 Anisotropy visualisation

Figure 1 displays the t-SNE representation (van der Maaten and Hinton, 2008) of the *three-way parallel* sentences in Chuvash, Russian, and English for three models. We observe that for Glot500, we easily distinguish three clusters that barely overlap, suggesting a high anisotropy. Monolingual pre-training, especially for languages from different families and scripts, struggles to properly align the sentences. LaBSE, on the other hand, manages to better align Russian–English sentences, as we cannot distinguish between both sentence representations any more in the figure. Chuvash, however, seems to suffer from poorer relative cross-lingual alignment with the two target languages.

Finally, knowledge distillation between the two models seems to alleviate cross-lingual misalignment, as the three sentence representations are now closer, with no visible language clusters. No parallel sentences for CHV–RU were needed to obtain this representation: it only relied on LRL *monolingual* sentences for pre-training and widely used Russian–English parallel sentences for distillation.

	Before	After CBIE	After whitening
XLM-R	7.82	20.79	**26.08**
G500	24.75	40.54	**45.65**
LaBSE	**59.64**	47.84	55.93
G500-LaBSE	**69.89**	52.02	66.09

Table 4: Averaged F-score over the 16 language pairs before and after applying CBIE or whitening.

6.3 Post-processing and distillation synergy

We assess the synergy between isotropy-enhancing transformations and knowledge distillation (the better approach compared to contrastive learning) on mining quality. Table 4 shows the averaged F-score over the 16 language pairs before and after applying CBIE or whitening on the sentence representations. We note that the isotropy enhancement technique mainly improves the mean-pooled word representations, while it significantly degrades the performance of LaBSE or of the model obtained with distillation. Hämmerl et al. (2023) observed

[9] Kurdish is for both Caspian languages, and Georgian is already in the first set of training languages. See Appendix B.

that such a transformation provides limited benefits to models trained on parallel data; we confirm that already isotropic representations do not benefit from it. We thus find little to no synergy between CBIE (or whitening) and the other two approaches.

6.4 Case studies: for one language branch

One real-life scenario is to perform parallel sentence mining for a limited set of *related* language pairs rather than typologically diverse ones. We focus here on mining for two language branches, Sorbian and Caspian languages, as both feature a language *not* seen by Glot500 (dsb and glk).

We start from XLM-R and pre-train it in two ways: with the relevant languages seen by Glot500 (base setting)[10] and additionally with the unseen language (+ setting). We use 100k monolingual sentences maximum for each language, drawn from the Glot500 corpus. For the two unseen languages, we use different sentences than those in our benchmark. Based on the results, we compare the simple mean-pooled representation (PT), a CBIE transformed version (as in Okabe et al., 2025), and the best knowledge distillation with related languages (as in the -all setting in Section 6.1), denoted as PT-LaBSE. We note that models are developed separately per language branch.

language branch	Sorbian				Caspian			
source	hsb		dsb		glk		mzn	
target	de	en	de	en	fa	en	fa	en
G500	21.33	14.03	21.33	1.45	4.06	0.25	9.92	0.31
LaBSE	74.87	77.43	74.87	74.69	45.49	28.68	56.64	32.63
G500-LaBSE	78.31	77.98	54.57	54.56	54.96	32.51	71.71	39.61
G500-LaBSE-all	**89.67**	**88.26**	67.29	66.41	**59.30**	**32.06**	70.39	40.16
PT	41.89	4.00	2.48	1.82	2.90	0.00	37.11	0.96
CBIE	44.29	11.25	17.33	9.45	17.31	1.05	56.73	4.02
PT-LaBSE	87.39	83.18	66.42	63.47	47.62	26.23	77.44	43.42
PT+	49.24	17.04	59.64	18.08	2.86	0.00	34.62	0.46
CBIE+	47.40	26.32	51.93	28.10	20.48	0.34	55.72	3.13
PT-LaBSE+	89.63	86.10	**93.09**	**89.61**	48.64	24.95	**79.56**	**43.99**

Table 5: Mining scores for models specifically pre-trained for Sorbian or Caspian languages.

Table 5 reports the results for all eight language pairs. We still notice relatively poor F-scores for the two languages not supported in the base setting PT, dsb and glk, despite the branch-specific pre-training. This shows that G500 benefits also from cross-lingual transfer from less related languages seen during pre-training. Interestingly, while using monolingual data for dsb (+ setting) improves its representation, with a similar trend to hsb across settings (+26 points from distillation), glk benefits far less from the added pre-training, hinting at a potential abjad-script barrier. Still, the distillation approach remains reliable, as it consistently outperforms LaBSE. Besides, monolingual pre-training helped distillation, as PT-LaBSE+ leads to better mining than PT-LaBSE. Finally, G500-LaBSE-all appears as the best overall system.

We also perform another case study to integrate a bilingual lexicon, a realistically available resource for low-resource language pairs. We consider the two pairs involving Chuvash in Appendix I and find the positive effect of additional distillation with words and phrases. Appendix J summarises the general pattern per language pair difficulty.

7 Conclusion

Our study removed the *direct* parallel sentence requirement to mine LRLs. We created a benchmark to evaluate mining quality for eight low-resource languages, paired with a main target language (a locally relevant high-resource language) and its English translation. This setting allowed for direct comparison with two levels of language distance per LRL, both in terms of language family and script. We then show the compound effect of *monolingual* LRL pre-training (XLM-R vs. G500) with three types of approaches to improve multilingual sentence embeddings: isotropy enhancement (CBIE and whitening; needing no external resources), contrastive learning, and knowledge distillation, using NLI datasets or parallel corpora of *related* languages. None of the strategies relied on ***direct*** *parallel* data involving our eight source LRLs. Yet, we found substantial improvement for all methods across most language pairs, compared to a simply mean-pooled sentence representation. Distillation from LaBSE led to the best mining score for challenging pairs, suggesting mutual benefit with LRL pre-training (G500). More generally, our analyses stress the importance of monolingual pre-training to fully leverage methods for better cross-lingual alignment. We also show that while isotropy improvement is beneficial 'locally', for closer language pairs, its effect is limited on harder pairs with language and script divergence and does not synergise well with sentence-level models.

Future work will extend the study to more language pairs, notably on the source LRL side. Moreover, we will focus on improving the alignment between distant language pairs, such as GLK–EN.

[10] de, cs, pl, hsb for Sorbian languages and fa, ckb, mzn for Caspian languages. We exclude English for comparison.

Limitations

First, our sentence encoding baseline, LaBSE, is already very robust, despite having never seen most of the source languages we considered (none but Corsican). Relatedly, we observe that for both Romance languages (Occitan and Corsican), we only managed to get close to and not above the score reached by the baseline LaBSE. It has, nevertheless, been extensively pre-trained on related languages (cf. Section E; e.g., Catalan, French, Spanish for Occitan) or on the language itself (for Corsican), using parallel sentences, which can explain its 'zero-shot' performance for mining. Besides, while LaBSE achieves higher performance for Indo-European languages written in the Latin script, other languages are comparatively lagging behind. Our approach, on the contrary, scales to a variety of languages with a similar degree of improvement, even for the most challenging language pairs, such as CHV–EN or XMF–EN. The harder the source language and the further the language pair, the better our approach performs compared to LaBSE for mining. It effectively improves the sentence representation of a low-resource language using only monolingual pre-training in the language (as in Glot500) and parallel datasets from *related* parallel corpora.

Moreover, despite the high F-scores achieved by G500-LaBSE-all (more than 70 for 10 language pairs), we still see a relatively low mining quality for some pairs (e.g., Caspian languages paired with English, between 30–40), suggesting a poor representation. The lowest scores primarily occur when the two languages are distant in terms of script or language family, and with smaller or no monolingual pre-training.

Furthermore, the choice of language pairs (and corresponding datasets) in our benchmark is bound by the availability of resources (both monolingual and parallel). The main bottleneck is the number of *high-quality* parallel sentences, which restricts the overall dataset size and explains its variation. It also reduces the number of possible language pairs to study.

Finally, our benchmark and experimental settings rely on synthetically injected parallel sentences, which only approximate mining from real corpora. Our study, however, follows the methodology of the BUCC Shared Task (Zweigenbaum et al., 2017) and Belopsem (Okabe et al., 2025). For the most challenging pairs, we observe a range of mining scores, showing that the benchmark is neither too simple (cf. baseline performance) nor saturated (cf. best scores below 60).

Ethical considerations

We create our benchmark from openly accessible corpora, respecting their licences. For that reason, each language pair has a different licence in our released version. Moreover, the machine translation we use to generate the English side for four *target* languages is deemed to be of decent quality, due to the nature of the translation direction and involved high-resource languages. We acknowledge, however, the risk it entails and try to measure it by verifying the translation quality of the system in Section C. We also note that the most challenging pairs of our benchmark (XMF–EN, GLK–EN, and MZN–EN) are curated with *non-translated* English sentences.

References

Sina Ahmadi, Razhan Hameed, and Rico Sennrich. 2025a. Literary translations and synthetic data for machine translation of low-resourced Middle Eastern languages. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 110–118, Vienna, Austria (in-person and online). Association for Computational Linguistics.

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

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

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, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the Workshop Track at ICLR*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas. Online version.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rusudan Gersamia and Irina Lobzhanidze. 2022. Megrelian Language Corpus. The Megrelian Language Corpus, 4 April. 2022(v1). Web.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Harald Hammarström and Robert Forkel. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web Journal*, 13(6):917–924.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised parallel sentence extraction from comparable corpora. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 7–13, Brussels. International Conference on Spoken Language Translation.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. 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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Ivana Kvapilíková and Ondřej Bojar. 2023. Boosting unsupervised machine translation with pseudo-parallel data. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 135–147, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Peiqin Lin, Andre Martins, and Hinrich Schuetze. 2025. A recipe of parallel corpora exploitation for multilingual large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4038–4050, Albuquerque, New Mexico. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Shu Okabe, Katharina Hämmerl, and Alexander Fraser. 2025. Improving parallel sentence mining for low-resource and endangered languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–205, Vienna, Austria. Association for Computational Linguistics.

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.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. Tep: Tehran english-persian parallel corpus. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, page 68–79, Berlin, Heidelberg. Springer-Verlag.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490, Dubrovnik, Croatia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

UNESCO. 2010. *Atlas of the world's languages in danger*, 3rd edition. Paris, France.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

A Language status

ISO	Glottocode	LF	NLP resource	Eth.	UNESCO
oci	occi1239	IE-R	1	E	SE–DE
hsb	uppe1395	IE-S	1	E	DE
dsb	lowe1385	IE-S	1	E	DE
chv	chuv1255	T	1	E	VU
cos	cors1241	IE-R	1	E	DE
xmf	ming1252	K	1	S	DE
glk	gila1241	IE-II	1	I	NE
mzn	maza1291	IE-II	0	S	NE

Table 6: Classification of languages for NLP resource availability and endangerment. 'LF' indicates the language family, and 'Eth.' refers to Ethnologue (Eberhard et al., 2025).

Table 6 reports the status of the eight source languages according to Joshi et al. (2020), Ethnologue (Eberhard et al., 2025), and UNESCO's *Atlas of the world's languages in danger* (UNESCO, 2010). For Ethnologue, E stands for endangered, S for stable, and I for institutional, in order of language vitality. UNESCO classifies as follows, ordered

from the most critical state to safe: SE: severely endangered, DE: definitely endangered, VU: vulnerable, and NE: safe.[11] For Occitan, the classification depends on the considered dialect. We also report the Glottocodes (Hammarström and Forkel, 2022) of each language and language families (IE: Indo-European (R: Romance, S: Slavic, II: Indo-Iranian), T: Turkic, and K: Kartvelian).

Besides, only two languages are present in the high-quality and multilingual FLORES+ benchmark for machine translation (NLLB Team et al., 2024): Chuvash and Occitan. The remaining six languages have thus not been considered by the popular multilingual sentence matching task.

B Language pair

source	main target	English	closest
Occitan	Spanish: 1	2	Catalan
Upper Sorbian	German: 2	2	Czech
Lower Sorbian	German: 2	2	Polish
Chuvash	Russian: 3	4	Turkish
Corsican	French: 1	2	Italian
Mingrelian	Georgian: 2	4	Georgian
Gilaki	Persian: 2	4	Kurdish
Mazandarani	Persian: 2	4	Kurdish

Table 7: Language pair difficulty when paired with the main target and English, from the easiest (1) to the hardest (4). Harder pairs are highlighted with varying shades of grey.

Table 7 summarises the language pair difficulty into four levels. Level 1 represents the easiest pairs with the same language branch, script, and relatedness to several high-resource languages (e.g., OCI–ES). Level 2 indicates a pair where the language family is the same but from a different branch (e.g., HSB–DE) or from the same branch but with a non-Latin script (e.g., XMF–KA). Level 3 is only for CHV–RU, as they are from different language families but with the same script. Finally, the hardest is level 4, where the script and broader language branch or family differ (e.g., GLK–EN).

We also report the *closest high-resource* language according to the taxonomy of Joshi et al. (2020). For Mingrelian, no language other than Georgian was deemed to be close enough. We use these languages paired with English during the additional distillation of LaBSE into Glot500 (G500-LaBSE-all) in Section 6.1.

[11] There are two other levels below SE.

C MT quality

Since part of our benchmark relies on Machine Translation into English, we evaluate the quality of Google Translate on the multilingual Flores+[12] benchmark (NLLB Team et al., 2024). We evaluate the translations according to three widely-used metrics: BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and xCOMET(-XL) (Guerreiro et al., 2024). We use sacreBLEU (Post, 2018) for the implementation of the first two. The third was ranked best in the WMT 2024 metrics Shared Task (Freitag et al., 2024) and supports the five considered languages.

lang. pair	BLEU	chrF++	xCOMET
DE→EN	48.77	69.46	98.32
ES→EN	33.73	60.37	97.14
FR→EN	51.97	71.18	97.21
RU→EN	41.29	64.73	97.09

Table 8: BLEU, chrF++, and xCOMET scores for Google Translate on Flores+ for translations from German, Spanish, French, and Russian into English.

Table 8 presents the translation quality for the four language pairs where we used MT into English. While BLEU and chrF++ are fairly average, the xCOMET score shows very high quality semantically for all four language pairs. As it is the metric with the highest correlation with human evaluation in Freitag et al. (2024), we deem the MT quality from Google Translate to be sufficient for our use.

D Corpora details

Below are the resources that we use to create our synthetic corpus for parallel sentence mining.

Occitan-Spanish Following Okabe et al. (2025), we use the Wikimedia corpus from OPUS (Tiedemann, 2012) for the parallel data. Similarly, the monolingual sentences for both languages come from the Leipzig corpus (Goldhahn et al., 2012), more specifically, the Wikipedia data from 2021 (30k sentences).

Upper Sorbian-German For Upper Sorbian sentences (monolingual and bilingual), we rely on the WMT 2020 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT (Fraser, 2020). The monolingual sentences in German come from

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

the 2020 Wikipedia data of the Leipzig corpus (Goldhahn et al., 2012) with 300k sentences.

Lower Sorbian-German We use the data from the WMT 2022 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT[13] (Weller-Di Marco and Fraser, 2022) for both monolingual and parallel sentences in Lower Sorbian. More precisely, we use the `66408_DSB_monolingual.txt.gz` file for the monolingual part and the 2022 training data files (`40194_train_dsb_de`) for the parallel corpus. For German, we use the news data from the Leipzig corpora (Goldhahn et al., 2012) in German (2022, 100k sentences).

Chuvash-Russian This language pair was studied in the WMT 2021 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT (Libovický and Fraser, 2021). We rely on the development dataset (in `devtest.chv-ru.tgz`) to have the parallel sentences. For both Chuvash and Russian monolingual data, we use the Leipzig corpora (Wikipedia 2021; 30k sentences)

Corsican–French For this language pair, we use parallel sentences from OPUS (Tiedemann, 2012). Given the dataset size, we combine the Wikimedia dataset and the eight sentences from Tatoeba and filter sentences with two or fewer words. We also manually corrected some sentences that were misaligned. On the monolingual side, we rely on the Leipzig corpora and use the Wikipedia corpus for both Corsican and French. We namely combine all three available corpora for Corsican (Wikipedia 2014, 2016, and 2021, each with 10k sentences). We use the 2021 30k Wikipedia corpus for French.

Mingrelian–Georgian/English The three-way parallel dataset from the Megrelian Language Corpus[14] (Gersamia and Lobzhanidze, 2022) is released under a CC BY-NC-SA 4.0 licence. We use the 2021 10k Wikipedia corpus for Mingrelian. For the Georgian or English monolingual side, we use a Georgian–English parallel corpus from OPUS: OpenSubtitles[15] (Lison and Tiedemann, 2016).

Gilaki and Mazandarani–Persian/English For both Gilaki and Mazandarani, we use the three-way parallel dataset collected in the PARME dataset (Ahmadi et al., 2025b). It is licensed with a MIT licence. For the source monolingual sentences, we use the 2016 10k Wikipedia corpus for Gilaki and for Mazandarani, the 2021 30k Wikipedia corpus from the Leipzig corpus. For the target side, we use an actual parallel corpus for Persian–English: the TEP corpus (Pilevar et al., 2011) from OPUS, which can be used for research and non-commercial purposes. We note here that both source languages were namely studied for mining in Ahmadi et al. (2025a), paired with English only.

[13] https://www.statmt.org/wmt22/unsup_and_very_low_res.html.

[14] https://xmf.iliauni.edu.ge/.

[15] https://www.opensubtitles.org.

E Language coverage of the models

Table 9 summarises the languages present in the pre-training dataset of the three language models that we compare, XLM-R, Glot500-m, and LaBSE. Glot500-m extends XLM-R (base) to more than 500 languages, including six of our source languages. Corsican has the most monolingual sentences among them (3,015,055), while Mazandarani has the fewest comparatively (73,719). We note that LaBSE is trained on more than 109 languages, including Corsican (both monolingual and parallel sentences). All three models have seen English and the six target languages (French, Georgian, German, Persian, Russian, and Spanish, in alphabetical order).

	XLM-R	Glot500-m	LaBSE
`oci`	✗	✓ (1.4M)	✗
`hsb`	✗	✓ (104k)	✗
`dsb`	✗	✗	✗
`chv`	✗	✓ (860k)	✗
`cos`	✗	✓ (3.0M)	✓
`xmf`	✗	✓ (175k)	✗
`glk`	✗	✗	✗
`mzn`	✗	✓ (74k)	✗

Table 9: Languages seen during pre-training for the three back-end multilingual language models.

F Methodological details

Contrastive learning with NLI datasets When using a NLI dataset for contrastive learning, sentences with an 'entailment' relationship are considered as positive pairs, while the 'contradiction' relation is its hard negative.

Similarity metric for sentence mining The CSLS metric (Conneau et al., 2018a; Artetxe and

Schwenk, 2019a), presented in Equation (1), is more effective than the standard cosine similarity for the mining task. It notably better finds source (resp. target) sentences close to multiple target (resp. source) sentences (i.e., 'hubs'), as they are more likely to attract noise than actual translations. Including the nearest neighbours in the similarity (the two subtracted terms) helps to penalise such situations, while actual parallel pairs would stand out thanks to higher cosine similarity, relatively to the remaining terms.

Mining threshold setting We follow the existing mining approach and consider a dynamic threshold (Hangya et al., 2018) based on the mean and standard deviation of the similarity scores to filter sentence pairs.

G Experimental details

G.1 Computational details

The parallel sentence mining pipeline relies on the creation of the sentence-level representation and the mining itself, both scaling with the dataset size (i.e., the longest experiments being for HSB–DE, the shortest for XMF–KA). The mining part carries out similarity search using Faiss (Johnson et al., 2019), which can also run with GPUs for faster results. We used 1 GPU (NVIDIA A100 or H100) for all our experiments. The whole pipeline runs in less than one hour for one source language paired with the two target languages (a few minutes for the smallest pair).

We have also trained or retrained models using the mSimCSE framework (Wang et al., 2022). To ensure comparability, we use the same pre-training parameters as the default implementation. Using the same GPU resources as above, the training took a few hours; the longest computation time was for multilingual contrastive learning with XLNI or translation datasets, which reached around one million instances.

Finally, distillation with LaBSE took around eight hours for the largest dataset (-all setting) with 3 epochs. For our distillation experiments, convergence can happen earlier, depending on the language pairs used.

G.2 External datasets for model training

We list the datasets used to train or pre-train our models. For contrastive learning, we use the English NLI dataset (276k sentences) for the -en variants and the XNLI dataset (2.2M sentences) for the -multi variant. The -tr variant uses the same dataset as the base dataset used for knowledge distillation: each language pair (English paired with French, Georgian, German, Persian, Russian, and Spanish) is represented through 100k sentences (maximum).

For the additional knowledge distillation settings (Section 6.1), the same 100k threshold is applied for each language pair, from the Parallel Sentences dataset. For the -all setting, this amounts to 1.0M sentences with the language pairs added (English paired with Catalan, Czech, Italian, Kurdish, Polish, and Turkish). For the -un setting, this leads to a dataset of 500k instances for the English–Arabic, simplified Chinese, French, Russian, and Spanish combined language pairs.

We use the same 100k threshold for our specific pre-training experiments (Section 6.4). We consider the Glot500 corpus as a source for the languages covered by the model. For the Sorbian branch, there are 400k monolingual sentences in Czech, German, Polish, and Upper Sorbian. The monolingual sentences for Lower Sorbian come from the *2021* edition of the WMT Shared Task (the benchmark relies on the *2022* edition). For the Caspian branch, we have 271,710 sentences for Central Kurdish, Mazandarani (less than 100k), and Persian. For Gilaki, we use the other dataset available from the Leipzig corpus: the 2017 community dataset. We make sure to pre-process the datasets before applying the 100k threshold to avoid internal duplicates, but also overlaps with our created benchmark sentences.

H Complete mining results

Tables 11 and 12 display the F-scores for all models tested on the eight main language pairs and English versions, respectively. We also include results from two more recent baseline models: mmBERT (base) (Marone et al., 2025) and multilingual-e5-base (multi-e5) (Wang et al., 2024). On average, both models underperform compared to G500 and LaBSE, respectively, on our new mining benchmark.

Compared to Table 2, we additionally report the results for XLM-R with CBIE and whitening post-processing (XLM-R+{CBIE|whiten}) and detailed results from Table 3.

Additional experiment with fine-tuning Contrastive learning and knowledge distillation appear as efficient approaches to keep existing language structures (for LRLs) while improving cross-

	CHV–RU	CHV–EN
G500-LaBSE-all	85.17	85.07
dict-word	**90.61**	**90.59**
dict-sent	86.43	83.94

Table 10: Effect of a bilingual dictionary for Chuvash paired with Russian or English.

lingual alignment. We additionally considered a standard fine-tuning setting for comparison, using a parallel corpus of *related* languages (i.e., the twelve languages used in the -all setting for distillation). We fine-tuned both Glot500-m and LaBSE using the Sentence Transformer framework (with the Multiple Negatives Ranking Loss), but did not find improvement compared to the standard LaBSE or our distilled models based on G500 (G500-LaBSE). The main assumption behind this lack of improvement lies in the nature of the parallel data: even with hard negatives sampled from the batch, the languages remain only *related*, which seems to worsen alignment with the LRL and, hence, the mining scores.

I Case study with a bilingual lexicon

We also study strategies to integrate a realistically and commonly available bilingual source for low-resource language pairs: a dictionary between the source and target languages. We consider G500-LaBSE as a student model and LaBSE as a teacher model. We use the Chuvash-Russian dictionary (63k entries) provided during the WMT Shared Task (Libovický and Fraser, 2021) and evaluate its impact on two language pairs, CHV–RU and CHV–EN. We try two approaches: either with words (and phrases) given directly as a pair (dict-word) for distillation or with 'code-switched' Russian sentences, where we replace words covered by the dictionary in Chuvash (dict-sent). This allows us to find which method leads to a better representation from a lexicon.

While using the word lexicon directly improves the mining score for both language pairs in Table 10, the 'code-switched' sentences only brought a slight boost in CHV–RU and degraded the score when paired with the unseen English.

J Summary of mining methods for language pairs

We observe three patterns on our benchmark according to the language pair difficulty:

1. For source languages (LRL) which are close to high-resource, well-represented languages, such as Occitan, or which are already seen during extensive bilingual training (Corsican for LaBSE): LaBSE is already strong and state-of-the-art.

2. For language pairs with only one better-resourced language in LaBSE's pre-training (e.g., hsb): pairs with moderate difficulty (level 2 in our taxonomy; both with Latin script but different language branch, e.g., HSB–DE, or non-Latin script but same language family, e.g., XMF–KA) can be improved using knowledge distillation from LaBSE, if the student model has been pre-trained using monolingual sentences (here, Glot500-m). Monolingual pre-training clearly helps to better benefit from distillation, leading to combined strengths from both models and decent mining quality (F-score above 70).

3. For the most challenging cases (levels 3 and 4, e.g., CHV–EN or MZN–EN), with diverging language family or script: knowledge distillation coupled with monolingual transfer is still the best setting. The achieved mining score can, however, be lower, specifically when the script is less present in pre-training (Georgian or Perso-Arabic scripts, in our case).

LM	OCI–ES	HSB–DE	DSB–DE	CHV–RU	COS–FR	XMF–KA	GLK–FA	MZN–FA
XLM-R	47.44	0.40	0.42	0.64	16.06	4.55	1.64	3.84
G500	79.96	21.33	2.97	39.59	49.94	41.25	4.06	9.92
mmBERT	61.13	4.50	1.98	1.07	25.21	12.80	4.72	6.30
XLM-R+CBIE	81.66	15.77	10.08	2.46	39.62	14.29	12.17	25.75
G500+CBIE	92.43	48.01	20.29	48.77	69.80	57.37	19.54	43.25
XLM-R+whiten	85.45	26.88	20.12	7.95	47.35	21.25	14.36	27.54
G500+whiten	93.40	66.39	35.34	57.25	72.14	54.49	20.51	39.43
mSC-XLM-en	62.64	13.45	11.91	7.11	52.44	13.01	20.46	30.37
mSC-XLM-multi	61.93	18.40	11.70	3.53	47.95	7.84	21.77	24.05
mSC-XLM-tr	67.95	21.28	15.79	2.50	44.18	5.46	18.64	24.77
mSC-G500-en	73.95	22.38	11.32	36.57	51.66	56.12	32.89	47.07
mSC-G500-multi	83.24	53.34	27.02	57.58	68.93	57.54	31.77	48.50
mSC-G500-tr	86.13	63.15	27.80	64.90	69.17	62.75	33.23	49.63
LaBSE	98.20	74.87	69.60	30.85	90.32	28.47	45.49	56.64
multi-e5	95.95	54.33	47.32	37.88	86.03	30.24	50.28	61.83
XLM-LaBSE	96.05	43.64	39.06	24.57	69.71	18.05	44.32	58.93
G500-LaBSE	97.11	78.31	54.57	80.68	83.70	71.11	54.96	71.71
G500-LaBSE-all	97.07	89.67	67.29	85.17	85.99	76.92	59.30	70.39
G500-LaBSE-un	97.63	87.83	56.98	82.56	84.76	73.84	39.20	64.43

Table 11: All F-scores (%) on the test set of the eight main language pairs for all models. mSC stands for mSimCSE, while G500 designates Glot500-m.

LM	OCI–EN	HSB–EN	DSB–EN	CHV–EN	COS–EN	XMF–EN	GLK–EN	MZN–EN
XLM-R	35.72	1.10	0.20	0.35	12.75	0.00	0.00	0.00
G500	62.59	14.03	1.45	27.61	38.34	2.44	0.25	0.31
mmBERT	38.82	3.26	0.73	0.24	18.36	0.00	0.00	0.00
XLM-R+CBIE	71.77	9.55	7.13	0.35	34.30	1.62	1.31	1.39
G500+CBIE	87.10	33.51	14.51	34.13	65.13	8.89	1.50	4.42
XLM-R+whiten	77.99	21.68	15.17	3.34	42.23	3.70	1.37	0.97
G500+whiten	88.43	49.95	25.64	47.47	69.91	3.72	2.62	3.72
mSC-XLM-en	50.24	10.46	8.55	2.11	43.31	2.92	4.39	3.31
mSC-XLM-multi	39.45	5.82	3.08	1.24	27.99	1.96	1.86	0.41
mSC-XLM-tr	63.69	20.12	13.61	1.47	43.85	2.16	7.62	10.01
mSC-G500-en	34.54	4.65	2.52	5.18	26.01	5.24	4.91	5.72
mSC-G500-multi	74.83	41.10	18.42	46.68	61.44	29.91	5.23	5.32
mSC-G500-tr	81.40	55.28	24.57	55.18	60.79	30.63	18.04	21.83
LaBSE	97.04	77.43	74.69	31.35	90.39	27.62	28.68	32.63
multi-e5	94.58	64.07	56.26	23.92	83.69	18.78	29.05	27.01
XLM-LaBSE	95.69	44.20	40.89	23.34	70.53	21.05	26.41	31.69
G500-LaBSE	96.92	77.98	54.56	79.25	85.58	59.61	32.51	39.61
G500-LaBSE-all	97.12	88.26	66.41	85.07	85.75	62.92	32.06	40.16
G500-LaBSE-un	97.24	83.72	52.97	79.80	83.81	60.70	20.41	30.60

Table 12: All F-scores (%) on the test set of the eight source languages paired with **English** for all models. mSC stands for mSimCSE, while G500 designates Glot500-m.