Can contrastive learning help parallel sentence mining of low-resource languages?

Anonymous ACL submission

Abstract

Recent work on cross-lingual sentence representation focused on contrastive learning as an alternative to pre-training based on parallel sentences due to their scarcity, especially for lower-resourced languages. In this study, we assess the robustness of two contrastive learning strategies which either use transliteration or natural language inference datasets to create positive and negative pairs. Instead of sentence matching, we evaluate the quality of the more complex parallel sentence mining task on five language pairs with low-resource (and endangered) languages: Lower Sorbian-German, Chuvash-Russian, Corsican-French, Mingrelian-Georgian, and Mingrelian-English. We find that while contrastive learning based on NLI is better overall and improves the representation quality, it remains effective mostly for our experiments on language pairs in the same script or language family.

1 Introduction

There are two main ways to obtain multilingual sentence representations: either by simply averaging the word embeddings of a language model or using parallel sentences to further train the model, such as LaBSE (Feng et al., 2022) or LASER and its variants (Artetxe and Schwenk, 2019b; Heffernan et al., 2022). While the second approach achieves better performance overall, its effectiveness for a given language depends on the existence of parallel corpora or its proximity to any of the pre-training languages. This is why contrastive learning has been explored as a viable strategy to enhance multilingual sentence-level representation without requiring parallel sentences. In this article, we consider two such systems: one which relies on transliteration to improve the multilingual representation across writing systems (Liu et al., 2024) and another that only requires Natural Language Inference (NLI) datasets (Gao et al., 2021; Wang et al., 2022).

The cross-lingual sentence representation quality can notably be assessed with two related tasks: parallel sentence matching (or sentence retrieval), where sentence pairs are shuffled and the pairing should be found again, or parallel sentence mining (or bitext mining), where truly parallel pairs must be found among larger monolingual corpora. Previous works mostly focused on either the easier sentence matching (e.g., Tatoeba benchmark, Artetxe and Schwenk, 2019b) or parallel sentence mining but on high-resource languages (e.g., the BUCC benchmark, Zweigenbaum et al., 2017). In this work, we focus on parallel sentence mining for low-resource languages.

The main research question to answer is hence: to what extent does contrastive learning *without parallel sentences* scale to mine parallel sentences for low-resource languages? We extend an existing methodology for synthetic corpus creation to five pairs for four low-resource languages, covering three writing systems and three language families. We then apply the two mentioned contrastive learning approaches to multilingual language models to evaluate their cross-lingual capabilities. We release the trained models and benchmark corpora[1].

2 Languages and corpora

We study the following (source-target) language pairs: Lower Sorbian-German, Chuvash-Russian, Corsican-French, and Mingrelian paired with Georgian and English. All four *source* languages are classified as 'scraping-by' (1 on a scale from 0 to 5) in the taxonomy of Joshi et al. (2020) in terms of available resources. Besides, Ethnologue (Eberhard et al., 2025) considers all but Mingrelian as endangered, while the UNESCO (2010) lists Chuvash as vulnerable and the three other languages as definitely endangered. We designate the five well (or better)-resourced languages (classified as 3 or

[1] Anonymous link.

above on the taxonomy) as *target*.

Lower Sorbian-German (DSB-DE) Lower Sorbian (ISO code: dsb) is a Slavic language spoken in Germany (Brandenburg). It is related to Upper Sorbian, the other language from the Sorbian branch, or Polish. It is hence a pair of two Indo-European languages, but from a different branch, 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 very close, as both are from the Romance branch of the broader Indo-European language family and written in the Latin script.

Mingrelian-Georgian/English (XMF-KA/EN) Mingrelian (xmf) is a Kartvelian language (where Georgian also belongs), spoken in Western Georgia. For the XMF-KA pair, the source-target language distance is hence smaller than for XMF-EN (same language family and same Georgian script).

Corpus creation We create synthetic corpora for parallel sentence mining, following the BUCC Shared Task methodology (Zweigenbaum et al., 2017). We mix gold parallel sentence pairs in monolingual corpora in each language, and the goal is to retrieve them. Both the DSB-DE and CHV-RU pairs were considered in the WMT Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT (Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022), which gives us both parallel and monolingual sentences. For COS-FR, we use parallel corpora from OPUS (Tiedemann, 2012) and monolingual sentences from the Leipzig corpora (Goldhahn et al., 2012). For Mingrelian, we use the Megrelian Language Corpus (Gersamia and Lobzhanidze, 2022), which consists of three-way parallel sentences: Mingrelian, Georgian, and English. This enables us to create two corpora, XMF-KA and XMF-EN, with the same monolingual sentence pairs by using Georgian-English parallel sentences as target. Appendix A details the exact corpora that were used.

We split each created corpus into training and test sets following a 25:75 ratio. Table 1 presents the datasets of the five language pairs in descending order of size.

language	train			test		
	source	target	paral.	source	target	paral.
DSB-DE	42,365	49,715	1,497	127,015	148,992	4,496
CHV-RU	30,205	30,998	998	90,620	92,998	2,998
COS-FR	4,815	5,185	185	14,419	15,553	557
XMF-KA/EN	2,443	2,443	68	7,330	7,330	205

Table 1: Size of the training and test datasets, where source and target include the injected parallel sentences.

3 Language models

Multilingual language models We use the standard approach of averaging the word embeddings[2] to get the sentence-level representation from a language model. We mainly compare two models: XLM-RoBERTa or XLM-R (base) (Conneau et al., 2020) and Glot500-m (Imani et al., 2023). The latter extends the former towards more than 500 languages with a particular focus on low-resource languages through pre-training on monolingual data.

Contrastive learning with transliteration We consider Furina (Liu et al., 2024), an extension of Glot500-m, which uses a contrastive learning framework to improve cross-lingual transfer beyond script differences. They fine-tune with a Transliteration Contrastive Modelling objective, which pairs the original sentence with its transliteration in Latin script (positive) against other (negative) examples. Transliteration is also applied to Latin script languages (i.e., removing diacritics or converting special characters).

Contrastive learning with NLI datasets The other work we assess is mSimCSE from (Wang et al., 2022), which extends the English-focused SimCSE (Gao et al., 2021) in the multilingual space. Their main system is mSimCSE-en, which uses batch contrastive learning on XLM-R *large* using English NLI data only (Conneau et al., 2017; Reimers and Gurevych, 2019). Inside one batch, sentences with an 'entailment' relationship are considered as positive pairs, while the 'contradiction' relation is its hard negative. This approach (mSimCSE-en-L) led to significant improvement on both sentence matching and mining, while it only used (monolingual) English sentences.

[2]We use the 8th layer for all language models.

Wang et al. (2022) also apply the same approach but with a multilingual NLI dataset to further foster cross-lingual transfer. This approach led to further improvement on their retrieval tasks compared to mSimCSE-en-L. In this work, however, we replace the base model with XLM-R *base*, for a fairer comparison with our baseline, and train it on the same XLNI dataset (Conneau et al., 2018). We denote this setting mSimCSE-multi-B.

Additionally, we switch the base model from XLM-R to Glot500-m and carry out contrastive learning using English NLI, giving us the new mSimCSE-Glot500-m-en model.

Isotropy improvement Multilingual sentence embeddings can also be improved by tackling the anisotropy of the vectors in the multilingual space. A method that has been recently explored is to apply a cluster-based isotropy enhancement or CBIE (Rajaee and Pilehvar, 2021), as in (Hämmerl et al., 2023)[3]. This technique first clusters the vectors and then uses a Principal Component Analysis to remove the top 12 principal components. We apply it to the Glot500-m sentence-level representation This setting will be called Glot500-m+CBIE.

Sentence encoders All the previous systems are compared to LaBSE (Feng et al., 2022), a state-of-the-art sentence encoder which was trained using a large amount of *parallel* sentences.

Appendix B summarises the language coverage of the models.

4 Experimental setting

4.1 Mining pipeline

We use an established mining pipeline, where we 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 systems previously described. Then, sentences are compared according to a dedicated similarity metric, CSLS (Artetxe and Schwenk, 2019a). We select our similarity threshold based on the training set performance. The experiments are evaluated using the F-score. Appendix C lists computational details for reproducibility.

4.2 Mining results

Table 2 displays the results on the five synthetic corpora. We first notice that pre-training on the language is crucial when using averaged word embeddings for sentence representation, as Glot500-m significantly outperforms XLM-R on seen languages, while it struggles with the unseen dsb.

LM	DSB-DE	CHV-RU	COS-FR	XMF-EN	XMF-KA
XLM-R	0.66	3.06	21.67	1.50	5.84
G500	2.24	14.01	48.96	4.23	26.46
Furina	5.46	11.44	47.73	2.51	29.10
mSC-en-L	8.92	6.15	57.77	4.81	14.63
mSC-multi-B	8.87	5.39	41.88	2.29	10.72
mSC-G500-en	7.23	20.62	42.12	3.28	34.57
G500+CBIE	10.86	**29.69**	64.00	3.02	**39.79**
LaBSE	**55.12**	22.00	**84.22**	**25.35**	38.89

Table 2: F-scores (%) on the five test sets. mSC stands for mSimCSE, while G500 designates Glot500-m. Results in **bold** indicate the best score.

Furina has an ambivalent influence on the mining quality compared to Glot500-m. The additional transliteration contrastive learning seems to benefit DSB-DE and XMF-KA only, degrading performance otherwise. For the former pair, the transliteration of the specific characters in dsb might have improved cross-lingual transfer, while for the latter, the approach seems to help because the script is less represented in the pre-training dataset.

Methods based on mSimCSE lead to noticeable improvement compared to XLM-R (base), especially when the languages are close (COS-FR or XMF-KA). It lags behind Furina for both CHV-RU and XMF-KA, which happen to be written in non-Latin scripts (source and target). The multilingual extension (mSimCSE-multi-B) mostly helps closer language pairs than distant ones (e.g., XMF-EN), and, despite its longer training time, it is not necessarily a better approach than mSimCSE-en[4].

Our extension using Glot500-m brings significant improvement to both mSimCSE-en but also Glot500-m in CHV-RU and XMF-KA, thanks to both additional pre-training and better cross-lingual transfer. It, however, struggles to outperform Glot500-m in COS-FR and XMF-EN and mSimCSE-en on DSB-DE.

The isotropy enhancement technique improves the mining quality of Glot500-m by a large margin, except for the XMF-EN pair where it degrades it. This method even manages to outperform the otherwise best approach, LaBSE, for two language pairs. But, LaBSE can rely on related languages

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

[4] Additional experiments with XLM-R (*base*) and English NLI indicate better scores than mSimCSE-multi-B.

for the other three language pairs: Polish for DSB-DE, Italian for COS-IT and Georgian for XMF-EN. We suppose that this makes its language and script alignment more robust.

4.3 Case study of Mingrelian

The Mingrelian-Georgian and English corpora enable us to compare the quality of the bilingual sentence representation directly. We notice that for all models, the XMF-EN pair is more challenging than XMF-KA. This is mainly due to the divergence in script and language family, underlining the language or script 'cluster' phenomenon in multilingual language models (Wang et al., 2022; Liu et al., 2024).

Contrastive learning with transliteration or NLI datasets both improved over Glot500-m for the XMF-KA pair, while it degraded for XMF-EN. We thus note that the script and language family barrier has not been overcome yet, as contrastive learning improved within a language and script cluster to the detriment of the Latin-Georgian script or Indo-European-Kartvelian alignments.

Additionally, we also applied CBIE to the mSimCSE-Glot500-en model and achieved F-scores of 36.82 for Georgian and 4.17 for English. This means that the isotropy enhancement can still improve cross-lingual transfer even after contrastive learning; however, there is no clear synergy since it remains worse than using CBIE directly on Glot500-m for XMF-KA.

5 Related works

Contrastive learning (Chopra et al., 2005; Hadsell et al., 2006) is used to improve the sentence representation by comparing a positive and a negative example for a given sentence, bringing similar sentences closer together. It has been applied for English sentence representation, such as in SimCSE (Gao et al., 2021). This framework can be either unsupervised, using dropout to corrupt the sentence, or supervised using the NLI sentence relationship.

The next step was its extension to multilingual sentence embeddings, such as mSimCSE, where Wang et al. (2022) showed that using English NLI datasets could improve the overall cross-lingual generalisability of the representation. This method proved to be more efficient than similar methods without contrastive learning, such as Goswami et al. (2021). Another approach was to use a monolingual setting, where transliteration is applied to the original sentence (Liu et al., 2024).

These methods remain more accessible for low-resource languages as they do not require parallel sentences for training at all, as opposed to sentence encoders such as LASER (Artetxe and Schwenk, 2019b) or LaBSE (Feng et al., 2022). 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. Tan et al. (2023) extend this work using contrastive learning with large parallel corpora (>40K sentences); for a given sentence pair, the English translation is considered as a positive example, while fairly similar English sentences are considered as a negative example. This is due to the better representation on the target side. We do not consider similar approaches because of the size of the available parallel sentences for some of our language pairs (e.g., with Mingrelian).

6 Conclusion

We evaluated two contrastive learning approaches to improve multilingual sentence embeddings: Furina, which uses transliteration for robustness across scripts, and mSimCSE, which relies on mono- or multilingual NLI datasets. Parallel sentence mining experiments on five corpora representing various levels of language distance (both in terms of language family and script) show that if contrastive learning does improve the cross-lingual representation on average, it still struggles for challenging pairs (e.g., XMF-EN) with too different source and target languages. We also note that having a poor representation of the language in the model (e.g., dsb) cannot be patched with contrastive learning alone.

We saw that considering languages paired with English only occults the extent of (mis-)alignment in the multilingual space (e.g., with Mingrelian). Since contrastive learning significantly improved for the closer XMF-KA pair, this suggests that 'locally', Mingrelian is better represented for cross-lingual transfer with Georgian.

Future work includes the extension of the study to more language pairs: either on the source side with more low-resource languages or on the target side through automatic translation of the current parallel sentences. Moreover, we will focus on improving the alignment between distant language pairs such as Mingrelian-English.

Limitations

Although our extension of mSimCSE with Glot500-m did not bring any state-of-the-art performance, our aim was to assess its effectiveness in improving the sentence-level representation for low-resource languages. We found that it indeed helps overall, albeit not as much as other techniques such as CBIE or direct supervision with a large number of parallel sentences. More generally, we see that LaBSE is a robust baseline model, despite having never seen some of the source languages we considered. It has, nonetheless, been extensively pre-trained in related languages.

In terms of mining quality, we observe that for most pairs (all except Corsican-French), the representation quality is still low. This is likely due to the different challenges we looked for in each language pair: the absence from the pre-training data for DSB-DE, the language distance (but in the same script) for CHV-RU, and the diverse script and language family for XMF-EN.

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

References

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, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. 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.

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.

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.

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.

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 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.

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.

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.

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.

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.

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 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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

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.

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 Corpora details

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

Lower Sorbian-German We use the data from the WMT 2022 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT[5] (Weller-Di Marco and Fraser, 2022) for both monolingual and parallel sentences in Lower Sorbian. More precisely, we use `mono.dsb.gz` (actually released in 2021) for the monolingual part and combine the `devtest.dsb-de.tgz` (2021) and the 2022 training data files for the parallel corpus. For German, we use the news data from the Leipzig corpora (Goldhahn et al., 2012) in German (2021, 300K 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 combined the development and test datasets (`devtest.chv-ru.tgz`) to have the parallel sentences, while we use `monocorpus_chv.zip` for monolingual Chuvash data. The Russian monolingual sentences also come from the Leipzig corpora (Wikipedia 2021).

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[6] 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.

B Language coverage of the models

Table 3 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, of which Chuvash (859,863 sentences), Corsican (3,015,055), and Mingrelian (174,994). We note that LaBSE is trained on more than 109 languages, including Corsican (both monolingual and parallel sentences). All three models have seen the five target languages (English, French, Georgian, German, and Russian, in alphabetical order). We recall that Furina is based on Glot500-m and mSimCSE on XLM-R.

	XLM-R	Glot500-m	LaBSE
dsb	✗	✗	✗
chv	✗	✓	✗
cos	✗	✓	✓
xmf	✗	✓	✗

Table 3: Languages seen during pre-training for the three back-end multilingual language models.

C 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 DSB-DE). 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. None of them took more than two hours.

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.

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

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