Improving Parallel Sentence Mining for Low-Resource and Endangered Languages

Shu Okabe[1,2] **and Katharina Hämmerl**[1,2] **and Alexander Fraser**[1,2,3]
[1]School of Computation, Information and Technology
Technische Universität München (TUM)
[2]Munich Center for Machine Learning
[3]Munich Data Science Institute
{shu.okabe,k.haemmerl,alexander.fraser}@tum.de

Abstract

While parallel sentence mining has been extensively covered for fairly well-resourced languages, pairs involving low-resource languages have received comparatively little attention. To address this gap, we present BELOPSEM, a benchmark of new datasets for parallel sentence mining on three language pairs where the source side is low-resource and endangered: Occitan-Spanish, Upper Sorbian-German, and Chuvash-Russian. These combinations also reflect varying linguistic similarity within each pair. We compare three language models in an established parallel sentence mining pipeline and apply two types of improvements to one of them, Glot500. We observe better mining quality overall by both applying alignment post-processing with an unsupervised aligner and using a cluster-based isotropy enhancement technique. These findings are crucial for optimising parallel data extraction for low-resource languages in a realistic way.

1 Introduction

Parallel sentence mining aims to find matching sentence pairs from a source and a target language. This task is usually the first step in creating a corpus to train other NLP models, most famously Machine Translation (MT) systems. So far, significant progress has been reached for well-resourced languages and language pairs, such as in the BUCC Shared Task 2017 (Zweigenbaum et al., 2017), where English was paired with four other high-resource languages.[1] However, when one of the languages is low-resource and hence not yet well supported by the language models, similar performance cannot be achieved. Besides, sentence encoding models heavily rely on parallel data during pre-training. Without enough parallel sentences, decent sentence representation cannot be reached. For instance, Tan et al. (2023) rely on at least tens of thousands of *parallel* sentences in their sentence representation. Therefore, before reaching these parallel data size milestones, sentence encoders are not reliable enough for mining, which mainly affect very low-resource languages and, more critically, endangered languages.

The alternative is then to rely on language models pre-trained with only monolingual data by averaging word representations. Mining will be more difficult because sentence encoders perform better than mean-pooled embeddings for sentence mining. Thus, we focus on trying to improve such alternative sentence embeddings effectively through two methods: Alignment post-processing and isotropy enhancement in the embedding space. Hangya and Fraser (2019) used alignments with static embeddings to filter sentence pairs, but we updated it with an unsupervised multilingual aligner. The second method has already been shown to be effective on several language pairs on the related but different task of sentence matching (Hämmerl et al., 2023).

Our main contributions are: 1) We create BELOPSEM (**Be**nchmark of **lo**w-resource languages for **p**arallel **se**ntence **m**ining), which consist of three new realistic benchmarks for parallel sentence mining for three language pairs where the source language is low-resource and endangered. 2) We improve on baseline approaches to parallel sentence mining by applying isotropy enhancement and word-level alignment, neither of them requiring additional resources such as parallel sentences, unlike state-of-the-art sentence encoder models such as LaBSE (Feng et al., 2022). We publicly release the benchmark datasets[2] and the updated mining pipeline.[3]

[1]LaBSE (Feng et al., 2022) reaches more than 88 points of F-score on all four BUCC corpora.

[2]At: https://github.com/shuokabe/Belopsem/.

[3]At: https://github.com/shuokabe/PaSeMiLL/.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 196–205

2 Languages and corpora

2.1 Three language pairs

We focus on the following three language pairs, where the source language is an endangered language according to Ethnologue (Eberhard et al., 2024), and the target language is well-resourced: Occitan-Spanish, Upper Sorbian-German, and Chuvash-Russian. We choose these pairs to represent different mining difficulties with a range in language distance in the pair. They also happen to have received recent attention from the NLP community through WMT Shared Tasks (Libovický and Fraser, 2021; Sánchez-Martínez et al., 2024). We note that all three source languages are classified as 'Scraping-Bys' (1) in the classification of Joshi et al. (2020), i.e. low-resource.

Occitan-Spanish (OCI-ES) Occitan (ISO code: `oci`; Glottocode: `occi1239`) is spoken in the south of France, Italy, and Spain. Both Occitan and Spanish belong to the Western Romance language family and are written in the Latin script. This pair represents a close language pair which can heavily rely on related languages (French or Spanish directly).

Upper Sorbian-German (HSB-DE) Upper Sorbian (`hsb`; `uppe1395`) is a West Slavic language spoken in Saxony in Germany, while German belongs to the Germanic family. Hence, both are Indo-European languages and use the same Latin script. This pair is more challenging since the distance between the two languages is larger. Yet, Upper Sorbian can rely on related Slavic languages, namely Czech and Polish, which are better resourced and supported by existing language models.

Chuvash-Russian (CHV-RU) Chuvash (`chv`; `chuv1255`) is a Turkic language spoken in the Russian Federation, while Russian belongs to the Slavic language family. Both are written in the Cyrillic alphabet. The language distance for the pair is thus larger than the previous two pairs, which makes it the most challenging of the three.

2.2 Parallel sentence mining corpus creation

We follow the overall methodology used for the BUCC 2017 Shared Task (Zweigenbaum et al., 2017) to create a controlled experimental environment: we inject sentences from existing parallel sentences into monolingual corpora. We mainly use the Leipzig corpora (Goldhahn et al., 2012) for monolingual sentences and choose similar sources (e.g., same year of Wikipedia or news) for both languages in a pair. Parallel sentences for Upper Sorbian-German and Chuvash-Russian come from two editions of the WMT Shared Task in Unsupervised MT and Very Low Resource Supervised MT (Fraser, 2020; Libovický and Fraser, 2021), while we use parallel data from Wikimedia 2023, available on OPUS (Tiedemann, 2012), for Occitan-Spanish. We pre-process the sentences to ensure encoding consistency and filter out too-short sentences. We aim for 6% of true parallel sentences in the final corpora. Appendix A presents more details on the data curation.

Table 1 presents the number of sentences in the corpora of each language pair. We split the generated corpora to have 25% assigned to the training dataset and the remaining for testing. In the original BUCC Shared Task, the training and testing balance was 50:50. However, we chose a 25:75 split to simulate a scenario where we want to mine further parallel sentences from larger corpora.

	train	test
Occitan corpus	7,899	23,675
Spanish corpus	7,780	23,334
of which parallel	486	1,457
Upper Sorbian corpus	15,980	47,847
German corpus	15,999	47,999
of which parallel	1,000	3,000
Chuvash corpus	7,998	23,995
Russian corpus	7,994	23,985
of which parallel	499	1,498

Table 1: Number of sentences in the datasets for all three language pairs in BELOPSEM.

3 Parallel sentence mining methodology

3.1 Standard mining pipeline

We follow PASEMILL (Okabe and Fraser, 2025), which is similar to the established mining pipeline of Hangya and Fraser (2019), but updates it using contextual language models instead of static embeddings. First, we encode both source and target sentences in the same sentence representation space using a multilingual language model. When relying on mean-pooled sentence representations, we choose the 8th layer in line with previous findings (Hämmerl et al., 2023; Imani et al., 2023; Okabe and Fraser, 2025).

Then, we use the CSLS (Cross-Domain Similarity Local Scaling) score (Conneau et al., 2018) to compute the similarity between a source and target sentence. The CSLS score is related to established margin-based methods (Artetxe and Schwenk, 2019) and is defined as follows:

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

where $\text{NN}_k(x)$ indicates the k-nearest neighbours of vector x. In our experiments, we set $k = 20$. CSLS is known to suffer less from the hubness problem (Dinu et al., 2015) and to get better performance compared to the standard cosine similarity (Artetxe and Schwenk, 2019). In our preliminary work on the Upper Sorbian-German pair, we did not find major differences in mining, whether we used the margin-based score or the CSLS score to find the nearest neighbour.

Finally, we pair each source sentence with its closest target sentence. The actual computation of the nearest neighbours is carried out using the Faiss library (Johnson et al., 2019). Since not all source sentences have a matching counterpart, we filter the mined pairs according to a threshold θ. We define this similarity threshold dynamically, as in Hangya et al. (2018):

$$\theta = \text{mean}(S) + \lambda \times \sigma(S), \quad (2)$$

where we compute the mean and standard deviation (σ) values of the similarity scores (S), with λ as a hyperparameter. Details on computational efficiency can be found in Appendix B.

3.2 Multilingual language models

To represent the sentences, we stick to three language models in our analysis. First, we select a multilingual language model which is widely used: XLM-RoBERTa or XLM-R (base) (Conneau et al., 2020). The second model is LaBSE (Feng et al., 2022), a state-of-the-art sentence embedding model. Our preliminary experiments showed better performance of LaBSE over LASER (Artetxe and Schwenk, 2019; Costa-jussà et al., 2022) on our datasets. We do not fine-tune sentence encoders due to their need for *parallel* sentences because it is not accessible for languages which do not yet have such a sizeable corpus. None of the two models have seen sentences in Occitan, Upper Sorbian, or Chuvash; they can only rely on related languages for the mining process (e.g., other Romance languages for Occitan, Polish or Czech for Upper Sorbian, or Kazakh for Chuvash).

The main language model we work on is Glot500-m (Imani et al., 2023), which further pretrains XLM-R (base) with *monolingual* data for more than 500 low-resource languages with an extended vocabulary. We choose this model for two reasons: it has seen all three source languages, which means a better representation, and it can be directly compared to XLM-R. An additional point to consider for our three source languages is the data support from Glot500. Occitan is already well supported by the model (among the top 50 tail languages,[4] with more than 1M sentences). Fewer (100k) sentences have been seen for Upper Sorbian, which lies in the mid-to-low pre-training rank of Glot500. Finally, while representing Chuvash solely with related languages may be difficult (for XLM-R or LaBSE), better representation can be obtained from Glot500 (800k pre-training sentences).

3.3 Unsupervised alignment post-processing

Comparing the word-level alignment is a method to improve parallel sentence mining, as already documented in Hangya and Fraser (2019), for instance, but there they relied on *static* word embedding similarity. Instead, we use an unsupervised multilingual aligner, SimAlign (Jalili Sabet et al., 2020), which leverages pre-trained language models to compute alignment links. In our case, we use the 8th layer of Glot500 and count the number of alignment links. We choose the `argmax` method for higher-precision links. We divide the number of symmetrised links by the number of words in the source and target sentences to get the average alignment ratio. A full alignment between both sentences hence gets a score of 1.

We additionally filter alignment links according to their overall frequency, keeping only the most frequent (and hence, more reliable) word pairs. Finally, we also select the mined sentence pairs using a dynamic alignment threshold α (as in Equation 2), which depends on the mean and standard deviation of all scores and a hyperparameter μ.

Our preliminary experiments showed that using a lower filtering threshold θ and then applying alignment post-processing yielded a better F-score in the end. Such filtering increases precision; start-

[4] Languages *not* seen by XLM-R during pre-training.

ing from 'noisier' sentence pairs actually led to a quicker gain in precision for reduced loss in recall, and hence an overall improvement of the F-score. Due to this choice, alignment post-processing may lead to higher recall and lower precision than the standard approach (cf. complete results below).

3.4 Cluster-based isotropy enhancement

To enhance isotropy in sentence representations, Hämmerl et al. (2023) explored ZCA whitening (Huang et al., 2021) and cluster-based isotropy enhancement (CBIE; Rajaee and Pilehvar, 2021). Both yielded comparable results on cross-lingual sentence retrieval for 1K sentences. We choose CBIE for its better scalability when it comes to larger datasets, as in our case, and apply it to the mean-pooled sentence representation before a similarity search for each dataset separately. This method determines dominant directions not globally but for each cluster in the space using Principal Component Analysis and removes the top 12 components. Appendix C shows t-SNE visualisations (van der Maaten and Hinton, 2008) of anisotropy for all three language pairs.

We do not report the results of the CBIE transformation on LaBSE because sentence embeddings suffer less from anisotropic representation and benefit little from it, as noted in Hämmerl et al. (2023).

4 Experimental results

We tune the similarity and alignment score thresholds (θ and α), or more exactly, their associated hyperparameters (λ and μ), on each training dataset. We evaluate the experiments using the standard Precision (P), Recall (R), and F-score (F). Table 2 presents the mining results for all three language pairs on the test sets of BELOPSEM. The F-score in **bold** is the overall best score, while the underlined score is the best among variants of Glot500. Complete results with precision and recall scores are available in Appendix D.

From the F-scores of the two baseline models, we see that having closer source and target languages is beneficial, which is in line with our intuition on language distance. The number of sentences in related languages may be another factor: While Occitan can benefit from very well-resourced languages (namely, French or Spanish itself), and Upper Sorbian has close links with Czech and Polish, Chuvash is far from other Turkic languages (even those only supported by Glot500).

LM	align.	CBIE	OCI-ES	HSB-DE	CHV-RU
XLM-R	NO	NO	48.08	1.43	3.06
LaBSE	NO	NO	**93.50**	**67.21**	28.24
Glot500	NO	NO	72.61	20.85	37.84
Glot500	YES	NO	77.19	32.70	37.39
Glot500	NO	YES	83.29	43.15	41.51
Glot500	YES	YES	84.46	50.82	**43.62**

Table 2: F-scores (%) on the test datasets of the three mining corpora in BELOPSEM.

XLM-R's performance also shows that relying only on related languages is not enough to carry out mining with mean-pooled sentence embeddings.

Regarding the standard Glot500 model, we notice better performance than XLM-R due to the additional pre-training (around or more than 20 points in F-score), but the mining quality is still worse than LaBSE. The notable exception is Chuvash, where the additional sentences seem to strongly help the model, gaining more than 9 points.

Using alignment post-processing only brings slight improvement on the final F-score, except for Upper Sorbian. For that corpus, this suggests that a significant number of actual parallel sentence pairs have a rather lower similarity score (and are thus discarded when the similarity threshold θ is too high). This seems to be mainly due to the language representation (i.e., number of pre-training sentences of Glot500), as it does not affect the other two languages as massively.

On the other hand, the CBIE transformation brings consistent improvement to Glot500: Related language pairs seem to benefit more from the transformation (+10 for OCI-ES and +22 for HSB-DE) than distant languages (+4 for CHV-RU). This confirms that the method also enables better representation of low-resource languages in the sentence embedding space. Finally, combining both improvement techniques leads to the best results. Still, the main increase seems to come from the CBIE transformation; the alignment post-processing rather complements it with more detailed filtering.

Qualitative analysis Table 3 presents an example of a sentence pair that was correctly mined after applying the CBIE transformation. We see that our base model pairs the Occitan sentence with a Spanish sentence unrelated to university work, resulting in a negative similarity score; hence, such a pair is discarded. On the other hand, isotropy

	sentence	sim.
OCI	A l'òra d'ara, qu'ensenha a l'Universitat de Nagoya.	
ES	A la fecha sólo sobrevive Pablo, quien vive en Parácuaro, Michoacán. *To date, only Pablo, who lives in Parácuaro, Michoacán, survives.*	-0.004
+ CBIE	Actualmente, trabaja en la Universidad de Nagoya. *He is currently working at Nagoya University.*	0.118

Table 3: Example of sentence mined for the Occitan-Spanish (OCI-ES) corpus before and after CBIE transformation with corresponding similarity scores.

enhancement enables a positive and higher similarity score with the correct Spanish translation of the sentence. The noticeable difference in similarity scores explains why this sentence pair has been mined by the second system. For this very sentence pair, we hypothesise that the named entity 'Universitat de Nagoya' (or Nagoya University) created confusion, coupled with the not obvious association of the four-word phrase 'A l'òra d'ara' with 'Actualmente' (currently).

5 Related works

Parallel sentence mining has mainly been studied to prepare a corpus to train MT models. Yet, few works focus on low-resource language pairs. Kvapilíková and Bojar (2023) already considered mining parallel sentences for the Upper Sorbian-German language pair using their pipeline (Kvapilíková et al., 2020). Their main goal was to crawl pseudo-parallel sentences for MT, for which they observed that 500k Upper Sorbian sentences were necessary to reach acceptable mining quality. Similarly, Heffernan et al. (2022) improve LASER with multilingual distillation (LASER3), requiring large amounts of monolingual data and a significant number of parallel sentences. Tan et al. (2023) rely on contrastive learning to further boost LASER3, but it still needs more than 40,000 *parallel* sentences coupled with back-translation even in the extremely low-resource scenario.

Two shared tasks provided comparable corpora: the BUCC Shared Tasks (Zweigenbaum et al., 2017, 2018), whose methodology we used to generate our corpora, and the Shared Task on corpus filtering for low-resource language pairs (Koehn et al., 2019, 2020). For the latter, the size of parallel corpora reaches several hundreds of thousands, which is difficult to obtain for endangered low-resource languages.

6 Conclusion

Low-resource languages, especially endangered ones, can benefit from bilingual sentence mining to gather a parallel corpus for downstream tasks, such as MT. However, current mining tools can show variable performance, with poorer results when one of the languages is neither supported nor close enough to other languages seen by state-of-the-art sentence encoders, such as LaBSE.

We thus focused on using a model which was trained using only monolingual data, Glot500. For controlled experimental conditions, we generated sentence mining corpora for three language pairs with varying degrees of linguistic similarity (BELOPSEM). We found that the CBIE transformation to enhance isotropy in the sentence representation consistently improves mining quality, while alignment post-processing helps when the language is not well-represented in the language model. These results hint at the possibility for low-resource languages to reach higher mining quality *without parallel data*.

The mining quality improvement we observed can help the early stages of parallel corpus creation where monolingual data is sufficiently present, but there are not enough parallel sentences to fine-tune sentence encoder models. We hope that the benchmark can also foster work on sentence mining for other language pairs. Future work will mainly extend this methodology to other language pairs (different scripts and lower or no support in Glot500).

Limitations

The choice of the language pairs can be discussed as they may not be the most challenging cases, with two of them from the same Indo-European language family and all three pairs using each the same script. Nonetheless, we stuck to plausible pairs with potential needs, having deployment in real-life settings in mind. For research purposes, considering pairs such as Chuvash-English, replacing the Russian target sentences with their translations might be an option to see further linguistic differences.

Moreover, compared to the state-of-the-art LaBSE model, mining performance can be notably lower with our mining pipeline using Glot500 (or any mean-pooled sentence embeddings). We

have, however, seen that for a language that is too far from the languages used to pre-train LaBSE, Glot500 is actually a better alternative for mining (cf. Chuvash-Russian). This means that by pre-training XLM-R with *monolingual* data on the language, we can reach higher mining quality. Eventually, with enough parallel sentences (possibly mined), we can fine-tune LaBSE or similar sentence encoders. The main goal of our work is to lower the data barrier for low-resource language pairs by removing the need for parallel data.

Finally, the size of the datasets remains small for all three settings. We restricted the number of sentences overall to stay in realistic dataset sizes (several tens of thousands of monolingual sentences) while maintaining a parallel sentence ratio of around 6%.

Acknowledgments

We thank Viktor Hangya for his help and the anonymous reviewers for their insightful comments. This work has received funding from the European Research Council (ERC) under grant agreement No. 101113091 - Data4ML, an ERC Proof of Concept Grant.

References

Mikel Artetxe and Holger Schwenk. 2019. 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.

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

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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. 2024. *Ethnologue: Languages of the World*, twenty-seventh 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.

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).

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.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. 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.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. 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.

Shu Okabe and Alexander Fraser. 2025. Bilingual Sentence Mining for Low-Resource Languages: a Case Study on Upper and Lower Sorbian. In *Proceedings of the Eighth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Honolulu. 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.

Felipe Sánchez-Martínez, Juan Antonio Perez-Ortiz, Aaron Galiano Jimenez, and Antoni Oliver. 2024. Findings of the WMT 2024 shared task translation into low-resource languages of Spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation*, pages 684–698, Miami, Florida, USA. 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).

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

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.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

A Data curation

We report the original data used to generate our three parallel sentence mining corpora.

Occitan-Spanish For the monolingual sentences, we use the 2021 Wikipedia data from the Leipzig corpora (30K) for both languages (Goldhahn et al., 2012). For the parallel sentences, we use the 2023 Wikimedia data from OPUS[5] (Tiedemann, 2012).

Upper Sorbian-German We use the WMT 2020 Shared Task data for both parallel sentences and monolingual Upper Sorbian data (Fraser, 2020). The monolingual data was provided by the Sorbian Institute. We use both development and test data (`devtest.tar.gz`) for the parallel sentences, which were controlled for the Shared Task evaluation. For German monolingual data, we use the Leipzig news corpora of 2020 (300K sentences).

Chuvash-Russian For the monolingual sentences, we also use the 2021 Wikipedia data from the Leipzig corpora (30K) for both languages. We chose this year to have a better correspondence with the parallel data provided by the WMT 2021 Shared Task (Libovický and Fraser, 2021). From the parallel data, we use the development set (`devel.chv-ru.{chv|ru}`), which was manually filtered for the Shared Task.

Licence Corpora downloaded from the Leipzig corpora are licenced under the Creative Commons Licence CC BY.[6] The Wikimedia data from OPUS has a CC BY-SA licence. For the resources from the WMT Shared Tasks, the Chuvash data has a CC0 licence, while the Upper Sorbian datasets have a CC BY-NC-SA licence. We hence release our corpora with the following licences: the Occitan-Spanish corpus with a CC BY-SA licence, the Upper Sorbian-German corpus with a CC BY-NC-SA licence, and the Chuvash-Russian corpus with a CC BY licence.

[5] `https://opus.nlpl.eu/wikimedia/es&oc/v20230407/wikimedia`.

[6] `https://wortschatz.uni-leipzig.de/en/usage`.

B Reproducibility

We ran our experiments on 1 GPU: we mainly used one NVIDIA H100 for faster computation, but also one NVIDIA Tesla V100. It is worth mentioning that Faiss (Johnson et al., 2019) or SimAlign (Jalili Sabet et al., 2020) can also be run on CPUs, albeit slowly, and the CBIE transformation does not require GPUs. In terms of actual computation time, the sentence representation step takes a few minutes on one V100 GPU, and so does the mining process (less than 10 minutes). The alignment post-processing on GPU is also very quick: Since we apply it on mined sentence pairs with a given threshold θ, we only consider a subset of the original monolingual corpora. Finally, the CBIE transformation takes only a few minutes on a CPU.

C Visualisation of anisotropy

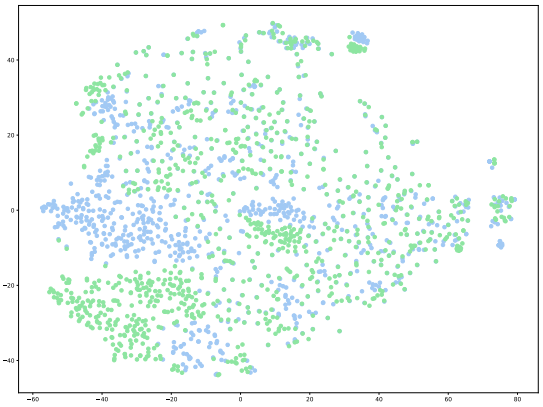
To visualise the impact of the CBIE transformation on the sentence representation space, we also follow the methodology of Hämmerl et al. (2023). We present t-SNE plots of 1,000 parallel sentences from each language pair before and after applying the CBIE transformation in Figures 1, 2, and 3.

As observed in Hämmerl et al. (2023), languages that have been less seen during pre-training suffer more from the issue of anisotropy.

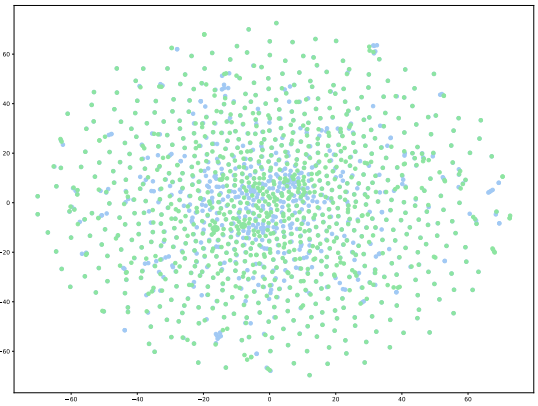
D Complete experimental results

Tables 4, 5, and 6 present the complete mining results for all three language pairs. As in Section 4, **bold** F-score stands for the overall best score, whereas the underlined score is the best among variants of Glot500.

We note here that the CBIE transformation mainly improves precision rather than recall. This implies that enhanced isotropy helps to better filter sentences.

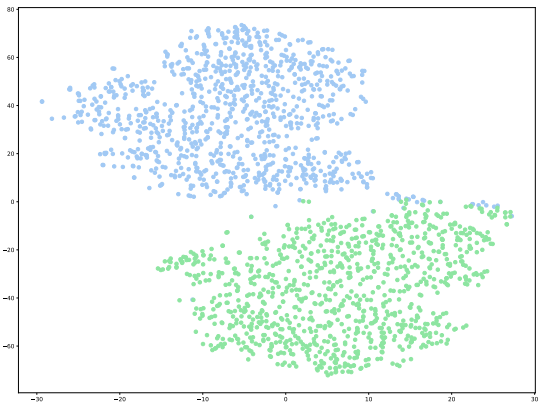


(a) Before transformation

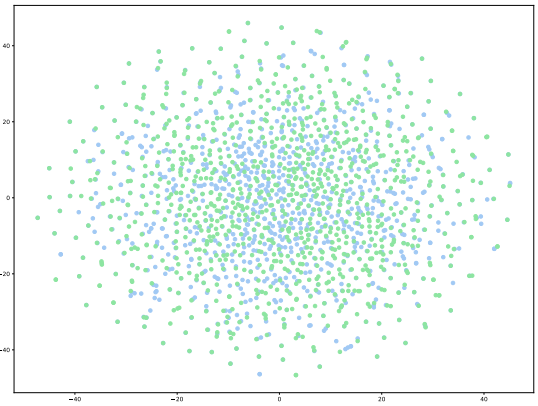


(b) After CBIE transformation

Figure 1: t-SNE plots for 1,000 parallel **Occitan-Spanish** sentences before and after CBIE transformation.

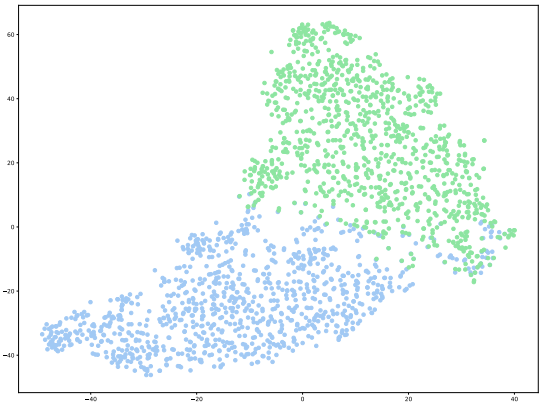


(a) Before transformation

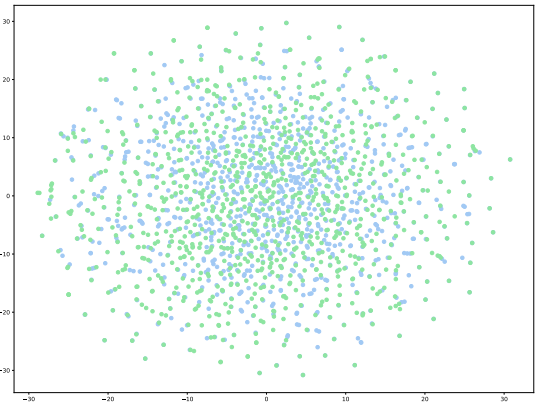


(b) After CBIE transformation

Figure 2: t-SNE plots for 1,000 parallel **Upper Sorbian-German** sentences before and after CBIE transformation.



(a) Before transformation



(b) After CBIE transformation

Figure 3: t-SNE plots for 1,000 parallel **Chuvash-Russian** sentences before and after CBIE transformation.

LM	align.	CBIE	P (%)	R (%)	F (%)
XLM-R	NO	NO	59.18	40.49	48.08
LaBSE	NO	NO	94.72	92.31	**93.50**
Glot500	NO	NO	72.81	72.41	72.61
Glot500	YES	NO	76.36	78.04	77.19
Glot500	NO	YES	90.50	77.14	83.29
Glot500	YES	YES	90.38	79.27	84.46

Table 4: Evaluation on the test dataset of the **Occitan-Spanish** corpus.

LM	align.	CBIE	P (%)	R (%)	F (%)
XLM-R	NO	NO	1.64	1.27	1.43
LaBSE	NO	NO	83.22	56.37	**67.21**
Glot500	NO	NO	22.72	19.27	20.85
Glot500	YES	NO	47.03	25.07	32.70
Glot500	NO	YES	60.74	33.47	43.15
Glot500	YES	YES	61.03	43.53	50.82

Table 5: Evaluation on the test dataset of the **Upper Sorbian-German** corpus.

LM	align.	CBIE	P (%)	R (%)	F (%)
XLM-R	NO	NO	3.84	2.54	3.06
LaBSE	NO	NO	42.18	21.23	28.24
Glot500	NO	NO	41.40	34.85	37.84
Glot500	YES	NO	39.57	35.45	37.39
Glot500	NO	YES	46.67	37.38	41.51
Glot500	YES	YES	54.62	36.32	**43.62**

Table 6: Evaluation on the test dataset of the **Chuvash-Russian** corpus.