Evaluating Latin and Ancient Greek Sentence Alignment through Parallel Sentence Mining

Sebastian Reichbauer[1] **and Shu Okabe**[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
Corresponding authors: {sebastian.reichbauer,shu.okabe}@tum.de

Abstract

Cross-lingual detection of intertextuality and translation in Latin and Ancient Greek through computational approaches is of great interest for classical studies. While several systems exist for parallel sentence detection, based on general multilingual or specific models for Latin–Ancient Greek, they have not been compared against each other. Therefore, we present a synthetic benchmark to evaluate the performance of language models regarding cross-lingual Ancient Greek and Latin parallel sentence mining. We first compare six language models to encode sentences and then further improve the cross-lingual alignment through post-processing, fine-tuning, and knowledge distillation. We find that the whitening transformation in combination with knowledge distillation provides excellent results. Specifically, SPhilBERTa, a trilingual language model for Ancient Greek and Latin, benefits the most from the improvements and achieves a notable mining score of 97.6 on our benchmark.

1 Introduction

Identifying instances of semantic similarity across texts is a major task in Classics and Philology, often phrased as intertextuality detection. Manual spotting of such passages is laborious, however, and will almost certainly stay incomplete, since scholars cannot memorize the whole corpus of Latin and Greek[1] texts to detect such paraphrases. Therefore, tools emerged to facilitate this task by applying computational concepts to automatically detect paraphrase candidates (Coffee et al., 2013). The more modern approaches rely mostly on the use of embeddings generated by language models, which have already shown promising results (Riemenschneider and Frank, 2023b). Nevertheless, only a reliable representation of tokens or sentences in the embedding space allows meaningful comparison between text passages.

To obtain reliable cross-lingual representations in language models, substantial training data from parallel corpora is needed. However, such resources are scarce for Latin and Greek, compared to modern and better-resourced language pairs. For instance, as of April 2025, there are fewer than 50 parallel sentences listed for the language pair on OPUS (Tiedemann, 2012), and open-access corpora for Classical Latin and Greek remain limited. Moreover, if Latin can be considered as a mid-resource language in Natural Language Processing (cf. language taxonomy from Joshi et al. (2020)), *Ancient* Greek remains extremely low-resource, with little to no representation in common multilingual language models.

In this article, we hence focus on retrieving cross-lingual Greek–Latin *translations*, as a first step towards paraphrase detection. We explore mining approaches which require few to no direct Greek–Latin parallel sentences for training, as this language pair faces the same challenges as other low-resource languages. Following an established approach, we first generate a bitext mining dataset to evaluate the Greek–Latin cross-lingual alignment. We compare six sentence encoding approaches as a backend of a mining pipeline. Moreover, we apply two types of post-processing to the sentence representations, whitening (Huang et al., 2021) and cluster-based isotropy enhancement or CBIE (Rajaee and Pilehvar, 2021), to improve cross-lingual alignment without training. We also consider a standard fine-tuning with limited parallel sentences in Greek–Latin from the Post-Medieval period and two knowledge distillation approaches (Reimers and Gurevych, 2020), where English or Latin are considered as anchor languages.

Our study extends Riemenschneider and Frank (2023b), which considered the task of parallel sentence *matching*. There, 1,000 parallel sentences

[1]'Greek' refers to Ancient Greek in this article, while Modern Greek is indicated by 'Modern Greek'.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 106–120

were shuffled and had to be paired back. Our mining benchmark is more challenging, as it features 2,000 translation pairs within a 23k corpus. Our contributions are the following:

(i) We create a synthetic benchmark to evaluate performance in cross-lingual Greek–Latin similarity detection.

(ii) We systematically compare six off-the-shelf language models for parallel sentence mining.

(iii) We study how to further improve mining quality with and without parallel sentences through post-processing, fine-tuning, and knowledge distillation.

(iv) We release the best-performing model and an open-source version of the benchmark.[2]

2 Related Works

Language Models for Greek and Latin A variety of monolingual language models for Greek and Latin exist, such as Latin BERT (Bamman and Burns, 2020) for Latin, while Krahn et al. (2023) provided BERT models for Greek, using knowledge distillation. Language models with explicit cross-lingual Greek–Latin training and understanding are much scarcer, however. Yousef et al. (2022) introduced Ugarit, a model based on XLM-R (Conneau et al., 2020), a multilingual language model, which was fine-tuned with both monolingual data and parallel sentences. Riemenschneider and Frank (2023a) introduced PhilBERTa, a trilingual language model for Greek, Latin and English. Based on this latter model, SPhilBERTa (Riemenschneider and Frank, 2023b) takes advantage of the knowledge distillation approach with English as an anchor language to create a language model for Greek–Latin intertextuality detection.

Parallel Sentence Mining To compare these methods, a suitable evaluation system must be designed to measure model performance in Semantic Textual Similarity (STS) tasks. The BUCC shared task (Zweigenbaum et al., 2017) introduced a synthetic mining corpus to evaluate the ability of models to detect parallel sentences for four high-resource language pairs. Following a similar protocol, Okabe et al. (2025) devised a mining benchmark for low-resource language pairs. A standard pipeline for mining relies on bilingual or multilingual sentence embeddings that are compared according to similarity metrics, such as the CSLS metric (Conneau et al., 2018). This score remains more robust and suited for mining than the cosine similarity, as it mitigates the hubness problem (Dinu et al., 2015).

Anisotropy One notable issue in multilingual language models remains the quality of cross-lingual alignment. Hämmerl et al. (2023) particularly explored the effects of anisotropic representations for STS tasks. Anisotropy thereby denotes the unbalanced use of the available vector space by embeddings, with some parts heavily occupied, while others remain largely empty. High anisotropy was found to induce significantly deteriorated performance for parallel sentence mining and matching tasks, explained by the poor cross-lingual alignment in the multilingual semantic space.

Thus, two approaches were considered to make the representations more isotropic without requiring parallel data. Whitening (Huang et al., 2021) lowers anisotropy by applying a linear transformation on the vector matrix to make the covariance matrix become the identity matrix. Another approach is CBIE (Rajaee and Pilehvar, 2021), which detects clusters in the embeddings, then identifies and removes the top x principal components in each cluster.

Knowledge Distillation Since direct parallel sentences between Greek and Latin are scarce, while English translations of Greek and Latin texts are more abundant, knowledge distillation (Reimers and Gurevych, 2020) can effectively extend the model's coverage to further languages. Distillation works by using a teacher model that creates high-quality embeddings in one language (e.g., English), and a student model that aims to represent parallel sentences (e.g., Greek–English or Latin–English) with embeddings closer to those of the teacher model. The student model can thus leverage the teacher model's robust language modelling without requiring extensive training data, and additionally keeps the additional capabilities from its initial multilingual embedding space. SPhilBERTa (Riemenschneider and Frank, 2023b) actually stems from distillation, where an English monolingual sentence encoder serves as teacher to extend a multilingual student model.

[2] https://github.com/TUM-NLP/latin-greek-mining

3 Benchmark Creation

We evaluate the quality of parallel sentence mining for Greek–Latin by creating a synthetic corpus. We follow the BUCC Shared Task approach (Zweigenbaum et al., 2017), where actual parallel sentences are injected into two monolingual corpora (one in each language). The data is split into a train partition that is used to find suitable mining parameters. Those are then applied to the test partition, which ultimately determines performance.

3.1 Datasets

Parallel Sentences We chose to focus on ancient Latin and Greek (i.e., not medieval or more recent) for parallel sentences to ensure measuring the performance of models in both language periods. This means that modern Latin translations (e.g., from the 19th century) of Greek texts are ruled out. Consequently, we only use 2,000 sentences from Boethius' translation in Latin of works of Aristotle on logic as parallel data, particularly *Analytica Priora* and *Topica*. Other works are not selected due to the limited online availability of texts or the high manual labour required to align them properly. The Bible has been specifically excluded because many models have already been trained on it, thereby increasing the risk of data contamination and reducing the meaningfulness of the results.

Monolingual Sentences Monolingual sentences should be as similar as possible semantically to increase the difficulty in detecting parallel sentences, while not creating accidental translation pairs. To achieve this, we chose to use authors largely from the same periods as Boethius and Aristotle, to match the language style, while keeping the topics different. Therefore, as shown in Table 1, we selected approximately 23,000 sentences in Greek and Latin from authors mostly from Late Antiquity (Latin) and the Classical Period (Greek). The topics include mostly historiography and theological texts, as well as philosophy, leaving out analytical works about logic. All text sources are detailed in Table 7 of Appendix A.

3.2 Preprocessing

Sentences should not be distinguishable with regard to formal features to prevent models from detecting parallel sentences due to their similar external appearance. Therefore, we preprocess our selected sentences as follows: The sentence length for monolingual as well as parallel sentences is bounded by a minimum of 10 and a maximum of 50 words. This provides a similar word length per sentence, as shown in Table 1. Additionally, due to the large amount of combinable characters in the Greek alphabet, we apply an NFC normalization on the data to ensure the same byte representations for the same characters. Latin sentences that cite Ancient Greek using the Greek alphabet are discarded to prevent recognition of the same alphabet as parallel. Excess whitespaces are removed, and Latin texts are normalized orthographically. This means specifically that spellings with 'j' are automatically replaced by 'i'. We also ensure manually that no texts using the orthographic variant of 'e' instead of 'ae' as casus endings (for genitive, for instance) are included.

	Latin		Greek	
	mono.	par.	mono.	par.
N_{sent}	22,727	2,000	21,641	2,000
$N_{words/sent}$	22.67	19.68	22.36	21.76

Table 1: Statistics of our synthetic benchmark datasets. N_{sent} indicates the number of sentences, while $N_{words/sent}$ denotes the average number of words per sentence. Parallel sentences (par.) are not counted in the monolingual sentence number (mono.).

4 Mining without Additional Data

We now detail our mining methodology, where no additional training data is needed by default.

4.1 Mining Pipeline

We use the parallel sentence mining pipeline suitable for BUCC-style corpora from Okabe and Fraser (2025) to evaluate the out-of-the-box performance of language models. Once sentences are converted into embeddings, we compute the CSLS metric (Conneau et al., 2018) for every combination of source and target sentence representation, as presented in Equation (1):

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

where $NN_k(x)$ denotes the k nearest neighbours of a vector x. We use this method, as it has been shown to achieve better results than the standard cosine similarity. We follow Okabe and Fraser

(2025) and set k to 20. Each source sentence is then paired with its closest target sentence based on their CSLS score. Since many sentences are not parallel and therefore do not have a corresponding sentence in the target corpus, we rely on a dynamic threshold, as defined by Hangya et al. (2018):

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

where S denotes the similarity scores and σ the standard deviation over S. If sentence pairs achieve a value greater than this threshold θ, they are accepted as parallel. λ serves as a hyperparameter here, which can be adjusted on the train split of the BUCC-style corpus, as described above, to maximise the sentence mining quality.

We evaluate the mining results according to the standard Precision, Recall, and F-score.

4.2 Base Benchmark

To assess the off-the-shelf performance of different models utilizing the described pipeline, we select specific models and conduct a baseline comparison on the created corpus. This provides comparability between the base versions of the considered models and underlines possible improvement paths.

4.2.1 Multilingual Language Models

We first consider multilingual language models to represent sentences, as they may have been trained on Greek or Latin. We include the following models in the benchmark:

Token-level models XLM-R (Conneau et al., 2020) is a token-level language model trained on 100 languages, including Latin (390 million tokens), but not Greek. Glot500 (Imani et al., 2023) extends XLM-R to support many low-resource languages and may therefore be well-suited for this task. It has been pre-trained on 1.2 million Latin and 380,000 Greek *monolingual* sentences.

Sentence encoder LaBSE (Feng et al., 2022), in contrast, is a state-of-the-art sentence encoder that was not pre-trained on Greek (only Modern Greek), but is trained to specifically set non-similar sentences apart in the embedding space. It has notably been pre-trained using English–Latin *parallel* sentences. It serves as a reliable reference point for the Greek–Latin mining performance of multilingual sentence encoders.

LLM-based embeddings Qwen3-embeddings (Team, 2025) is an LLM-based approach to obtain text embeddings. It consequently allows for examining the ability of LLMs for the mining task. To maintain comparability with the other models, we choose to use its rather smaller version, with 0.6 billion parameters and 1,024 dimensions, even though versions with up to 30 billion parameters exist. We denote this model Qwen below.

4.2.2 Ancient Language Specific Models

We choose Ugarit (Yousef et al., 2022) and SPhilBERTa (Riemenschneider and Frank, 2023b) as representatives of models that have been specifically trained for Greek–Latin semantic similarity detection. Ugarit is based on XLM-R and has been fine-tuned on 12,000,000 monolingual Ancient Greek tokens, 8,000 Greek–Latin, and 32,500 English–Greek parallel sentences. SPhilBERTa is based on PhilBERTa (Riemenschneider and Frank, 2023a) and fine-tuned via knowledge distillation using English–Latin and English–Greek parallel sentences, with English as the anchor language.

4.3 Isotropy-Enhancing Transformations

Anisotropy analysis We analyze the anisotropy of the best three models from the base benchmark. Following Hämmerl et al. (2023), we measure three properties: First, we examine the total anisotropy, defined as the mean cosine similarity over all considered vectors. Second, we assess the contribution of individual dimensions to the total anisotropy. It is defined by a partial cosine similarity, as shown in Equation (3).

$$CC(f_i) = \frac{1}{n} \sum_{\{x,y\} \in C} CC_i(f(x), f(y)) \quad (3)$$

where $CC_i(u, v)$ equals the contribution of the individual dimension i of two vectors u and v to the total cosine similarity and is defined by $CC_i = \frac{u_i v_i}{\|u\| \|v\|}$. Third, we count the number of outlier dimensions, i.e., those with values greater than 3 times the standard deviation of all dimensions. We apply these measures both before and after our transformations.

Whitening We apply ZCA whitening (Huang et al., 2021) to the three best models of the base benchmark. It is defined by Equation (4):

$$\hat{E} = (E - m)UD^{-\frac{1}{2}} \quad (4)$$

where $E \in \mathbb{R}^{n \times d}$ is the embedding matrix to be whitened, subtracted by the mean vector m. The covariance matrix $cov(E)$ is decomposed into $E = UDU^T$, with D as a diagonal matrix of eigenvalues and U as a matrix of eigenvectors.

CBIE Additionally, we apply CBIE (Rajaee and Pilehvar, 2021) to the same model representations to compare it to whitening, closely following their approach. First, we cluster the embeddings using k-means into 7 partitions. Then, we detect the top 12 principal components using PCA and zero-mean the clusters. Then, the identified principal components are removed for each cluster.

5 Results without Additional Data

5.1 Off-the-shelf Results

Model	Precision	Recall	F-score
XLM-R	0.2	0.1	0.1
Glot500	8.3	4.7	6.0
Qwen	44.9	35.3	39.6
SPhilBERTa	46.4	39.0	42.4
Ugarit	77.4	59.4	67.2
LaBSE	**79.9**	**64.6**	**71.4**

Table 2: Mining results on our synthetic benchmark for six sentence encoding approaches.

Table 2 displays the results of the six sentence encoding approaches on our mining benchmark.

Glot500 and XLM-R We can observe that Glot500 and XLM-R both show poor performance. This was expected for XLM-R, as it also performed poorly in Riemenschneider and Frank (2023b). Glot500 is based on XLM-R and might therefore yield poor results. However, the Ugarit model is based on XLM-R as well. Consequently, we see that the massively multilingual pre-training of Glot500 is of little help for our task, while Ugarit achieves far superior results thanks to its specific training for Greek–Latin similarity detection.

Qwen embeddings Qwen achieved a reasonable result of 39.6. Due to its lower parameter count (0.6 billion) compared to other Qwen versions with up to 30 billion parameters, LLMs with more parameters might perform better. However, 0.6 billion parameters are more than all the other language models we tested have, as LaBSE, the model with the second-most parameters, has only 0.5 billion.

Model	Base	Whitened	CBIE
LaBSE	0.20	$2.58 \cdot 10^{-5}$	$9.29 \cdot 10^{-6}$
SPhilBERTa	0.40	$3.05 \cdot 10^{-5}$	$2.93 \cdot 10^{-5}$
Ugarit	0.76	$2.42 \cdot 10^{-5}$	$1.10 \cdot 10^{-5}$

Table 3: Anisotropy of selected models out-of-the-box and after applying whitening and CBIE.

The embedding dimension is also higher for Qwen, compared to LaBSE (768).

SPhilBERTa and Ugarit SPhilBERTa and Ugarit show vastly different performance despite both being trained specifically for cross-lingual similarity detection in Greek and Latin. The main difference lies in the training method. SPhilBERTa relies on knowledge distillation, while Ugarit uses parallel sentences for training, inferring that fine-tuning with parallel sentences might be better suited for this task.

LaBSE LaBSE shows the best performance with 71.4 points, better than Ugarit or SPhilBERTa, although it has not been trained on Greek. This might be due to its training method, which separates dissimilar sentences on large corpora and has been proven valuable for sentence mining tasks, even for unseen languages.

5.2 Anisotropy Study

Table 3 presents the three anisotropy measures before and after CBIE and whitening transformations. We can observe that LaBSE shows the lowest anisotropy, while Ugarit exhibits the highest by far. The low anisotropy by LaBSE can be explained by its training method, which aims to move dissimilar sentences apart and thereby utilizes the embedding space better. Ugarit's high anisotropy is not surprising either, since Hämmerl et al. (2023) discovered that its base model, XLM-R, showed very high anisotropy, which seems to be inherited here. After applying whitening as well as CBIE, the anisotropy is reduced to almost 0. Both transformations, consequently, are effective in lowering the anisotropy of embedding spaces. Appendix B displays the full results in Table 8, and a visual illustration of the measured anisotropy is shown in Figure 2.

5.3 Whitening and CBIE

Table 4 shows the results of applying whitening and CBIE on the embeddings of the mining benchmark. As expected, LaBSE does not profit much

from these transformations, as it already had low anisotropy and better cross-lingual alignment. In contrast, SPhilBERTa shows a significant performance increase with both whitening and CBIE, more than expected from its moderate anisotropy, especially in comparison to Ugarit. Despite the similar performance of SPhilBERTa with both methods, it is clear that whitening is the superior choice for Greek–Latin similarity detection with these models. It yields better results than CBIE across all models. This is specifically pronounced when excluding SPhilBERTa. Ugarit with whitening emerges now as the best model on our benchmark, with a score of 86.5.

Model	Precision	Recall	F-score	Δ_F
Whitening				
LaBSE	71.3	68.9	70.0	– 1.4
Ugarit	**89.4**	**83.9**	**86.5**	+19.3
SPhilBERTa	84.0	68.4	75.4	+33.0
CBIE				
LaBSE	71.4	49.3	58.3	–13.1
Ugarit	80.9	57.3	67.1	– 0.1
SPhilBERTa	80.5	67.3	73.3	+30.9

Table 4: Mining results after applying whitening and CBIE. Δ_F indicates the difference with the base F-score.

6 Fine-tuning

We now explore a standard fine-tuning setting to leverage a small Greek–Latin parallel corpus.

6.1 Methodology

Dataset We use around 10,000 direct parallel sentences for fine-tuning. 6,612 sentences stem from the DFHG corpus,[3] which has already been used to fine-tune Ugarit. Another 3,587 sentences come from Thucydides' *Peloponnesian War* with a modern Latin translation. By applying a minimum word count per sentence of 3, the sentence number is reduced to 10,013.

Training Settings We employ the CachedMultipleNegativesRankingLoss as the objective function. Since the MultipleNegativesRankingLoss seemed to work quite well for LaBSE, we also use it here in its cached version to support large in-batch sizes despite limited VRAM. The in-batch size is indeed

[3] https://www.dfhg-project.org/

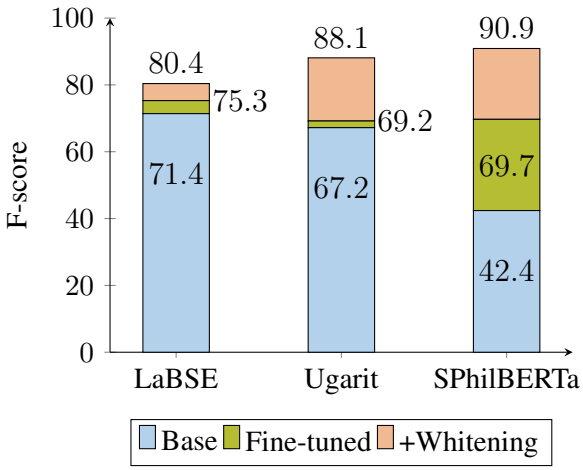


Figure 1: F-scores of the base benchmark (blue), after fine-tuning (green) and after applying whitening to the fine-tuned results (red).

crucial to the performance of the MultipleNegativesRankingLoss, since it works by pulling dissimilar sentences within a single batch apart. The batch size is set to 256, and we fine-tune with two epochs to account for the limited amount of training data. This training method is applied to the best three models of the base benchmark.

Whitening Since whitening improved the mining performance of the off-the-shelf models, we also examine whether this effect remains as pronounced *after* fine-tuning.

6.2 Mining Results

Fine-tuned Results Figure 1 presents the results with fine-tuning. One key takeaway is that all models improve at least slightly relative to the base performance. SPhilBERTa exhibits a significant score increase by 27.3. Ugarit's small improvement can be attributed to data redundancy, since the sentences from the DFHG dataset were already used for training Ugarit. The marginal improvement by LaBSE is surprising, however, since it has not been trained on Greek–Latin sentences. Nevertheless, since it uses a similar objective function, that positive effect might lead to other models profiting more than LaBSE, which has already experienced the effect of this loss function.

Results after Whitening After whitening, both Ugarit and SPhilBERTa perform much better, while LaBSE only increases its score by a moderate margin. This is in line with our expectations, as LaBSE showed the lowest anisotropy and therefore did not benefit much from isotropy-enhancing meth-

ods such as whitening. SPhilBERTa specifically proves to be the new best model with 91 points on our benchmark. We see that fine-tuning with parallel sentences yields positive results, especially when combined with whitening, which retains its effectiveness even in already fine-tuned models.

7 Knowledge Distillation

7.1 Methodology

As Riemenschneider and Frank (2023b) explored knowledge distillation to obtain a reliable Greek–Latin model, we also explore combining the strengths of two chosen models. Distillation aims to take advantage of a teacher model M with an advanced representation of a specific anchor language. The student model $\hat{M}$ is trained to mimic the embeddings of the teacher model. This is done by using parallel sentences between the target language and the anchor language with the loss function in Equation (5):

$$MSE(B) = \frac{1}{|B|} \sum_{j \in B} \Big[(M(s_j) - \hat{M}(s_j))^2 + (M(s_j) - \hat{M}(t_j))^2 \Big] \quad (5)$$

where $M(s)$ indicates the embedding of an anchor language sentence s in batch B from model M, while t denotes a sentence from the target language. The anchor language embeddings as well as the target language embeddings are shifted into the embedding space of the teacher model, thereby replicating the semantic space of the anchor language in the teacher model. Consequently, a more robust cross-lingual representation is achieved without the need for full retraining.

Experiment Details We examine both English and Latin as anchor languages. We chose English as it is a high-resource language with much data available, while Latin is one of the target languages. With the latter, we can also compare its performance with the English-mediated approach. We selected Latin over Greek due to the higher amount of data on which models have been trained.

As a teacher model, we use LaBSE due to its better English representation comparatively. It is indeed crucial to have a teacher model that has a reliable representation of the anchor language. Since its embeddings are mimicked by the student model, their quality largely influences the result. Additionally, we use the fine-tuned version of SPhilBERTa, which we call SPhiTune, as teacher model. This provides a reference point for models with stronger abilities in one of the target languages, here, Latin. We note that it also scored second in Section 6.2.

As student models, we rely on Glot500 and SPhilBERTa, the former for its broad multilingual understanding, the latter for its already specialized training on data relevant to the Greek–Latin pair.

All experiments with English serving as an anchor language are conducted with 1 epoch, while for Latin, we train with 2 epochs due to the limited amount of training data.

Given the promising results of whitening, we also examine whether it can further improve performance. We employ it both after distillation and for the teacher embeddings. Applying it to the teacher embeddings should lower the anisotropy and pass that effect to the student models. Appendix D presents further reproducibility details.

English-based Distillation For English as an anchor language, we use Greek–English and Latin–English parallel sentences for distillation, similarly to SPhilBERTa. Since the two language pairs feature relatively larger datasets, we rely on 380,000 Greek–English sentences from Krahn et al. (2023) and 100,000 Latin–English sentences.[4]

Latin-based Distillation When we consider Latin as an anchor language, we use the same 10,000 sentences as for fine-tuning (Section 6.1). The training data, therefore, is not comparable in size to using English as an anchor language, but covers both of our languages of study.

7.2 Mining Results with Distillation

English Table 5 presents selected results of knowledge distillation with English as an anchor language. The full results are displayed in Table 9 of Appendix C. We can observe that knowledge distillation through English is indeed an effective approach for better mining performance. LaBSE performs better as a teacher model than SPhiTune. This is not surprising, since the relevant property for teacher models is their representation of the anchor language, where LaBSE excels, while SPhiTune was not trained for English understanding.

Applying whitening afterwards leads to another notable performance increase. SPhilBERTa fine-tuned with knowledge distillation can profit sim-

[4] https://huggingface.co/datasets/grosenthal/latin_english_translation

Teacher model	Precision	Recall	F-score	Δ_F
Base Result				
LaBSE	**91.2**	**83.3**	**87.1**	+44.7
SPhiTune	68.6	61.5	64.9	+22.5
Whitening applied afterwards (to "Base Result")				
LaBSE	**98.0**	**97.1**	**97.6**	+55.2
SPhiTune	91.3	74.6	82.1	+39.7

Table 5: Results of knowledge distillation with **English** as an anchor language, before and after whitening. We compare two *teacher* models (LaBSE and SPhiTune) for the same student model (SPhilBERTa). Δ_F indicates the difference to the base benchmark score of the *student* model.

ilarly from whitening as its base version and receives the best result yet. The same experiment has been conducted with Glot500 as student model. However, it exhibited low performance. This is probably due to its worse representation of Greek and Latin, as already seen in the base benchmark. Hence, we left it out for this analysis and report it in Appendix C. We additionally report the results with whitening applied to the teacher embeddings there, due to its small influence on mining performance.

Teacher model	Precision	Recall	F-score	Δ_F
Base Result				
LaBSE	42.6	36.8	39.5	– 2.9
SPhiTune	**72.5**	**63.1**	**67.5**	+25.1
Whitening applied afterwards (to "Base Result")				
LaBSE	**92.0**	81.5	86.4	+44.0
SPhiTune	91.8	**89.0**	**90.4**	+48.0

Table 6: Results of knowledge distillation with **Latin** as an anchor language, before and after whitening. Δ_F indicates the difference to the base benchmark score of the *student* model (SPhilBERTa).

Latin We show selected results of knowledge distillation with Latin as an anchor language in Table 6. The full results are displayed in Table 10 of Appendix C. We note that Latin as anchor language leads to worse results in almost every case. This is in line with our expectations, however, because the sentence representations of LaBSE and SPhiTune turn out worse than those for English as a high-resource language, due to the limited amount of training data. Another factor is the difference in the size of the parallel datasets, with 10,000 sentences for Latin as anchor language and 480,000 for English.

We also observe that here SPhiTune seems to perform better than LaBSE as teacher model. This is due to the specific training SPhiTune received regarding the Latin language. This allows it to create better Latin sentence embeddings than LaBSE, which remains multilingual with no language-pair-specific pre-training. Here also, Glot500 shows worse scores than SPhilBERTa on our benchmark (cf. results in Table 10 of Appendix C). Since applying whitening to the teacher embeddings led to negligible differences, we also report those results in the Appendix. The smaller amount of available parallel sentences, combined with poorer language representation than English, makes Latin less suitable as an anchor language for distillation.

In short, knowledge distillation in combination with whitening appears to be a promising approach to improve cross-lingual alignment. Here, English seems to be a better anchor language, whereas using Latin is less effective in our case.

Both fine-tuning with parallel sentences and knowledge distillation do enhance the model performance on our synthetic benchmark. However, the results of knowledge distillation through English are much better than those of direct fine-tuning. This is not surprising, since the data available for knowledge distillation with English as anchor language is also much more abundant than Greek–Latin parallel sentences.

8 Conclusion

This article assessed Latin and Greek sentence alignment through parallel sentence mining and compared several enhancement methods based on differing data requirements. First, we created a synthetic benchmark to evaluate the performance of language models in cross-lingual Greek–Latin similarity detection. Using this benchmark, we analyzed a diverse set of six models regarding their suitability for mining. LaBSE, Ugarit, and SPhilBERTa achieved here the best off-the-shelf performance. We then compared post-processing methods without additional data to improve cross-lingual alignment, namely CBIE and whitening. We found that whitening consistently outperforms CBIE and strongly improves mining performance. Fine-tuning and knowledge distillation are both valuable methods for improving models to leverage additional available parallel data. In our setting, knowledge distillation with English as an anchor

language appeared as more efficient, mainly thanks to larger available datasets and better language representation. Finally, combining knowledge distillation with whitening leads to the best results. Specifically, SPhilBERTa with LaBSE as a teacher achieves a significantly high F-score of 97.6 on our synthetic benchmark.

Future work will extend the limited set of language models evaluated in this article. Additionally, our best mining approaches can be used to find additional parallel sentences from monolingual corpora. Another problem that remains to be tackled is to find a method to evaluate the performance of models for detecting actual semantic similarity as in paraphrases rather than translations.

Limitations

The main limitation of this study comes from the bias when selecting the datasets to create our synthetic benchmark. As suitable corpora are scarce, we focused on selected authors and time periods for both monolingual and parallel sentences. This choice, however, guarantees that the created benchmark is consistent and that no clear divergence can be found between the two types of sentences. As writer style, language period, and domain are all influential factors in mining, we report all the sources used for our benchmark creation.

Our benchmark also relies on a synthetic approach, instead of a true mining setting, with comparable corpora. Nevertheless, the methodology is established through the BUCC Shared Task and subsequent endeavours for bitext mining evaluation.

Acknowledgments

We thank the anonymous reviewers for their comments. This work was partly funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *Preprint*, arXiv:2009.10053.

Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson. 2013. The tesserae project: intertextual analysis of latin poetry. *Literary and Linguistic Computing*, 28(2):221–228.

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 *International Conference on Learning Representations (ICLR)*.

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*.

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.

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.

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.

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.

Kevin Krahn, Derrick Tate, and Andrew C. Lamicela. 2023. Sentence embedding models for Ancient Greek using multilingual knowledge distillation. In *Proceedings of the Ancient Language Processing Workshop*, pages 13–22, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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 Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 11–19, Honolulu, Hawaii, USA. Association for Computational Linguistics.

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.

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. 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.

Frederick Riemenschneider and Anette Frank. 2023a. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023b. Graecia capta ferum victorem cepit. detecting Latin allusions to Ancient Greek literature. In *Proceedings of the Ancient Language Processing Workshop*, pages 30–38, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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).

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022. An automatic model and gold standard for translation alignment of Ancient Greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France. European Language Resources Association.

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 Dataset sources details

Table 7 displays the detailed *monolingual* sources used to create our synthetic Latin and Ancient Greek benchmark.

B Anisotropy results

Figure 2 presents the t-SNE representations of Latin and Ancient Greek parallel sentences for three models: Ugarit, LaBSE, and SPhilBERTa. Moreover, Table 8 details our anisotropy study before and after transformations.

C Knowledge Distillation Results

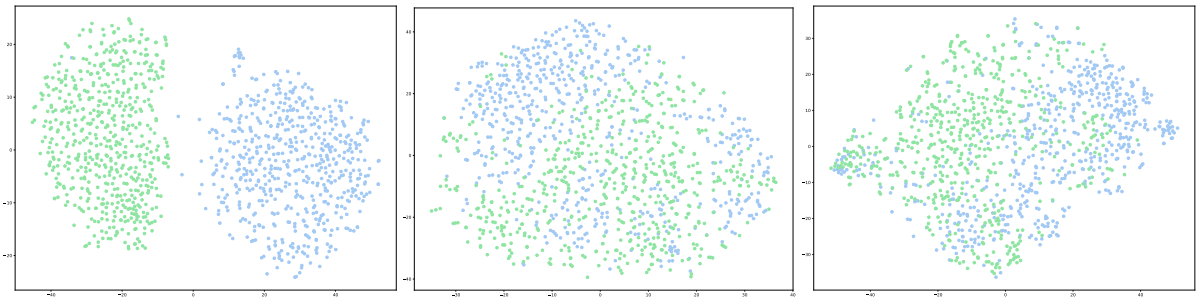
Tables 9 and 10 present the full results for our knowledge distillation experiments with English or Latin as an anchor language.

D Reproducibility

For training, we used one H100 GPU with 94GB HBM2 VRAM. Fine-tuning with Greek–Latin parallel sentences took only minutes to execute (all specifications for one model). It took a similar length of time to apply knowledge distillation with the same sentence pairs. For knowledge distillation with English as the anchor language, much more time was needed due to the larger dataset. The training took around 2 hours on the same GPU.

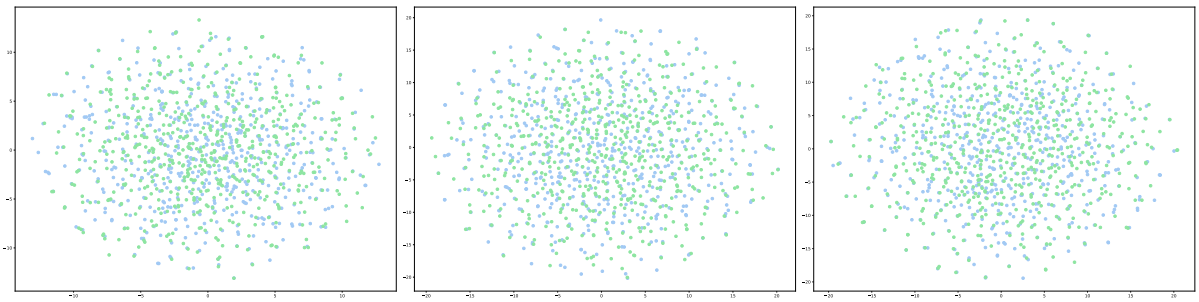
Author	Work	Domain	N_{sent}	$N_{words/sent}$
		Latin		
Ammianus Marcellinus	*Res gestae*	Historiography	3,621	26.38
Augustinus	*Confessiones*	Theology	2,413	22.33
	De civitate dei	Theology	9,898	
	De fide et symbolo	Theology	174	
Boethius	*De fide catholica*	Theology	55	22.30
	Liber de persona et duabus naturis	Theology	184	
	De trinitate	Theology	103	
Cassiodorus	*Variae*	Administrative Literature	4,483	20.99
	De anima	Theology	438	
Jordanes	*Getica*	Historiography	665	22.09
	Romana	Historiography	693	
Total			22,727	22.67
		Ancient Greek		
Xenophon	*Anabasis*	Historiography	2,311	20.19
	Apologia	Philosophy	88	
	Oikonomikos	Economy	661	
	Hellenika	Historiography	2,679	
	Cyropaedia	Biography	3,110	
	Memorabilia	Philosophy	1,274	
	Symposion	Philosophy	380	
Demosthenes	*Phillipika*	Speech	96	22.23
Polybios	*Historia*	Historiography	9,706	24.40
Isokrates	*Panathenaikos*	Speech	321	26.76
Galen	*Peri Physikon Dynameon*	Medicine	1,015	24.00
Total			21,641	22.36

Table 7: Summary of the most important properties of the **monolingual** data selected, Latin at the top, Ancient Greek at the bottom. $N_{words/sent}$ describes the average number of words per sentence. For comparison, 2,000 parallel sentences are used.



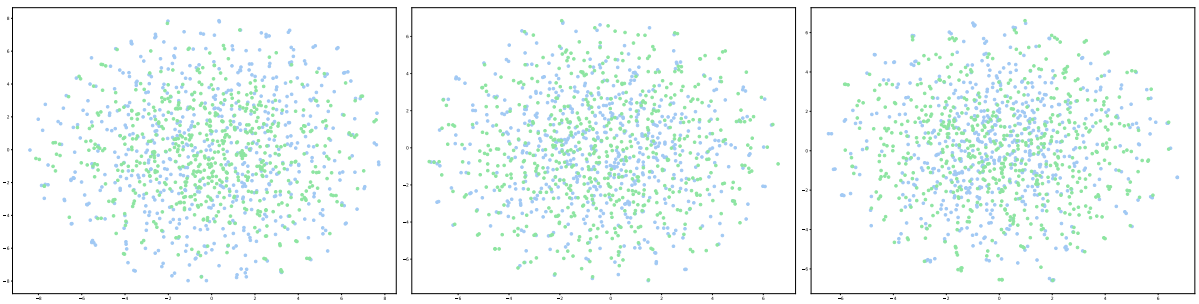
(a) Ugarit (b) LaBSE (c) SPhilBERTa

Untransformed Results



(d) Ugarit (e) LaBSE (f) SPhilBERTa

Transformed through whitening



(g) Ugarit (h) LaBSE (i) SPhilBERTa

Transformed through CBIE

Figure 2: Graphical representation of embeddings of Ancient Greek and Latin sentences from three models (Ugarit, LaBSE, and SPhilBERTa) using t-SNE, before and after applying whitening and CBIE. Each color represents a specific language.

Model	Anisotropy	Outlier Dimensions	Average Contribution of Dimension to Total Anisotropy
Base Results			
LaBSE	0.20	Dim 316 Dim 242 Dim 404 Dim 263	0.012 0.004 0.004 0.004
SPhilBERTa	0.40	Dim 461 Dim 513	0.023 0.007
Ugarit	0.76	Dim 588	0.380
Whitening			
LaBSE	$2.58 \cdot 10^{-5}$	Dim 216 Dim 591	$8.27 \cdot 10^{-7}$ $3.64 \cdot 10^{-7}$
SPhilBERTa	$3.05 \cdot 10^{-5}$	Dim 355	$4.85 \cdot 10^{-7}$
Ugarit	$2.42 \cdot 10^{-5}$	Dim 459 Dim 267 Dim 184 Dim 239 Dim 72	$1.74 \cdot 10^{-5}$ $1.57 \cdot 10^{-6}$ $1.40 \cdot 10^{-6}$ $9.78 \cdot 10^{-7}$ $7.96 \cdot 10^{-7}$
CBIE			
LaBSE	$9.29 \cdot 10^{-6}$	Dim 136	$1.33 \cdot 10^{-7}$
SPhilBERTa	$2.93 \cdot 10^{-5}$	Dim 535 Dim 555	$4.13 \cdot 10^{-7}$ $3.99 \cdot 10^{-7}$
Ugarit	$1.10 \cdot 10^{-5}$	Dim 588 Dim 459	$2.20 \cdot 10^{-6}$ $6.77 \cdot 10^{-7}$

Table 8: Total anisotropy of selected models combined with outlier dimensions and their respective contribution to the total anisotropy, out-of-the-box and after applying whitening and CBIE.

Model	Precision	Recall	F-score	Change to base benchmark
Base Result				
LaBSE-Glot500	80.8	74.6	77.6	+71.6
LaBSE-SPhilBERTa	**91.2**	**83.3**	**87.1**	+44.7
SPhiTune-Glot500	49.2	38.6	43.3	+37.3
SPhiTune-SPhilBERTa	68.6	61.5	64.9	+22.5
With Whitening on the Teacher Embeddings				
LaBSE-Glot500	**91.5**	73.6	81.6	+75.6
LaBSE-SPhilBERTa	89.6	**83.7**	**86.6**	+44.2
SPhiTune-Glot500	71.5	65.3	68.2	+62.2
SPhiTune-SPhilBERTa	85.9	78.6	82.1	+39.7
Whitening applied afterwards (to "Base Result")				
LaBSE-Glot500	95.1	93.9	94.5	+88.5
LaBSE-SPhilBERTa	**98.0**	**97.1**	**97.6**	+55.2
SPhiTune-Glot500	84.3	77.2	80.6	+74.6
SPhiTune-SPhilBERTa	91.3	74.6	82.1	+39.7
Whitening applied afterwards (to "With Whitening")				
LaBSE-Glot500	94.7	92.5	93.6	+87.6
LaBSE-SPhilBERTa	**98.5**	**96.5**	**97.5**	+55.1
SPhiTune-Glot500	89.5	76.5	82.5	+76.5
SPhiTune-SPhilBERTa	93.9	93.4	93.6	+51.2

Table 9: Results of knowledge distillation with **English** as anchor language, with comparison to base benchmark score of student model.

Model	Precision	Recall	F-score	Change to base benchmark
Base Result				
LaBSE-Glot500	0.9	0.9	0.9	-5.1
LaBSE-SPhilBERTa	42.6	36.8	39.5	-2.9
SPhiTune-Glot500	1.4	1.1	1.2	-4.8
SPhiTune-SPhilBERTa	**72.5**	**63.1**	**67.5**	+25.1
With Whitening (Teacher Embeddings)				
LaBSE-Glot500	1.8	2.1	2.0	-4.0
LaBSE-SPhilBERTa	48.0	37.7	42.3	-0.1
SPhiTune-Glot500	3.3	3.9	3.6	-2.4
SPhiTune-SPhilBERTa	**74.0**	**69.3**	**71.6**	+29.2
Whitening applied afterwards (to "Base Result")				
LaBSE-Glot500	7.2	3.8	5.0	-1.0
LaBSE-SPhilBERTa	**92.0**	81.5	86.4	+44
SPhiTune-Glot500	28.8	19.9	23.5	+17.5
SPhiTune-SPhilBERTa	91.8	**89.0**	**90.4**	+48.0
Whitening applied afterwards (to "With Whitening")				
LaBSE-Glot500	18.3	18.2	18.2	+12.2
LaBSE-SPhilBERTa	85.6	75.9	80.5	+38.1
SPhiTune-Glot500	26.2	23.4	24.7	+18.7
SPhiTune-SPhilBERTa	**90.7**	**90.5**	**90.6**	+48.2

Table 10: Results of knowledge distillation with **Latin** as anchor language, with comparison to base benchmark score of student model.