Testing the Limits of Neural Sentence Alignment Models on Classical Greek and Latin Texts and Translations

Caroline Craig[1,*], Kartik Goyal[2,*], Gregory Crane[3,*], Farnoosh Shamsian[4,*] and David A. Smith[1,*]

[1]*Khoury College of Computer Sciences, Northeastern University, 360 Huntington Ave, Boston, MA 02115, United States*

[2]*School of Interactive Computing, Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332, United States*

[3]*School of Arts and Sciences & School of Engineering, Tufts University, 419 Boston Ave, Medford, MA 02155, United States*

[4]*Leipzig University, Augustuspl. 10, 04109 Leipzig, Germany*

Abstract

The Greek and Latin classics, like many other ancient texts, have been widely translated into a variety of languages over the past two millennia. Although many digital editions and libraries contain one or two translations for a given text, about one hundred translations of the *Iliad* and twenty of Herodotus, for example, exist in English alone. Aligning the corpus of classical texts and translations at the sentence and word level would provide a valuable resource for studying translation theory, digital humanities, and natural language processing (NLP). Precise and faithful sentence alignment via computational methods, however, remains a challenging problem. Current alignment methods tend to have poor coverage and recall since their primary aim is to extract single sentence pairs for training machine translation systems. This paper evaluates and examines the limits of such state-of-the-art models for cross-language sentence embedding and alignment of ancient Greek and Latin texts with translations into English, French, German, and Persian. We release evaluation data for Plato's *Crito*, manually annotated at the word and sentence level, and larger test datasets based on coarser structural metadata for Thucydides (Greek) and Lucretius (Latin). Testing LASER and LaBSE for sentence embedding and nearest-neighbor retrieval and Vecalign for sentence alignment, we found best results using LaBSE-Vecalign. LaBSE worked surprisingly well on ancient Greek, most probably because it had been merged with modern Greek data in its training. Both LASER-Vecalign and LaBSE-Vecalign did best when there were many ground-truth one-to-one alignments between source and target sentences, and when the order of sentences in the source was preserved in the translation. However, these conditions are often not present in the kinds of literary and free translation we wish to study, nor in editions with multiple translations, extensive commentary, or other paratext. We perform book-level and chapter-level error analysis to inform the development of a software pipeline that can be deployed on the vast corpus of translations of ancient texts.

Keywords

sentence alignment, multilingual embedding, machine translation, Ancient Greek, Latin

1. Introduction

Texts from the corpus of ancient Greek and Latin have been translated multiple times into many languages. Having access to sentence-level links between original texts and their translations would be useful to students and researchers alike. Without access to living speakers, being able to easily consult multiple translations at a granular level while reading an ancient text would enrich the reader's understanding of the original [1, 2]. This would also open up access to non-experts in the source language, ancient or modern [3]. Moreover, such a dataset would represent a valuable resource to the machine learning community. Ancient texts have been translated many times into multiple languages and span different styles, content, and contexts, a variety that makes for "an excellent challenge for NLP" [4]. They also provide ample source material for a large and varied dataset of multilingual parallel sentences. Finally, sentence alignment—the process of automatically matching corresponding sentences in a source text and its translation—is crucial for finer-grained computational analysis of translations. This is because a large number of models and methods for computationally processing multilingual data assume access to parallel data at the level of sentences or small chunks of text.

This paper presents our work evaluating approaches to sentence alignment and their ability to work with pre-modern and literary texts. In so doing, we aim to shed light on the features of this data that can present challenges and the types of errors that are likely to occur. In addition to the challenges associated with processing ancient languages with limited resources, we also provide evidence for additional sources of error associated with sentence alignment of ancient texts and their translations: *noisiness* in the translations due to the presence of substantial paratext including footnotes, commentaries, multiple translations, and chunks of source text itself. We also present the pipeline that achieved the best results, LaBSE - Vecalign [5, 6], and how it can be used successfully to align pre-modern and literary texts with their translations. In this pipeline, we first extract embeddings of the sentences in the source text and the translations via LaBSE, which is a state-of-the-art method to encode multilingual sentence similarity. Then, we use Vecalign, which combines alignment by dynamic programming with approximate coarse-to-fine pruning, to align source and target sentences using their similarity computed by LaBSE or another embedding model. We run experiments using Plato's *Crito* and its translations in English [7, 8], German [9] and Persian, annotated at the chunk level;[1] Thucydides' *Peloponnesian War* and its translations in English [10] and French [11]; and Lucretius's *On the Nature of Things* and its translation in English [12]. The French translation of Thucydides and

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France
*Corresponding author.
craig.car@northeastern.edu (C. Craig); kartikgo@gatech.edu (K. Goyal); Gregory.Crane@tufts.edu (G. Crane); farnoosh.shamsian@uni-leipzig.de (F. Shamsian); dasmith@ccs.neu.edu (D. A. Smith)
https://github.com/caro28/ (C. Craig); https://kartikgo.github.io/ (K. Goyal); https://facultyprofiles.tufts.edu/gregory-crane (G. Crane); https://www.khoury.northeastern.edu/home/dasmith/ (D. A. Smith)
0000-0001-7444-7495 (K. Goyal); 0000-0001-6847-2811 (G. Crane); 0000-0002-6636-6940 (D. A. Smith)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]As described in §4.2.1, Farnoosh Shamsian supervised the students performing these alignments and producing the Persian translation as part of a project at the University of Leipzig.

the English translation of Lucretius contain substantial paratext; all other translations and the original ancient Greek and Latin texts do not. In the next phase of our work, we will apply LaBSE - Vecalign to the database of 1,526 translations of Greco-Roman texts compiled by the Open Greek and Latin Project (OGL)[2] and share their sentence-level alignments.

2. Related Work

Sentence alignment has not received much attention as an end in itself, with models [6, 13, 14, 15] built primarily to create training data for machine translation systems that prioritize precision over recall [16, 17]. Previous generations of such sentence alignment models were based on sentence length [14, 13], augmented in [18] by the use of a word translation model to estimate the probability that aligned sentences are translations of each other. Crucially, they did not rely on dense low-dimensional representations of sentences for capturing semantic similarity, such as sentence embeddings. Hence, they perform poorly in general when compared to newer models like Vecalign [6], especially for high resource languages. Where accurate machine translation systems exist for the appropriate language pairs, researchers have often found it easier to translate other languages into English before performing monolingual alignment [19, 20]. In this work, we show that reliable sentence-level alignment for translations of ancient language texts is a challenging task.

Word alignment has received substantial coverage by the natural language processing community [21, 22]. However, these approaches presume access to chunk-level bitext pairs for good performance, or corresponding short spans of text (such as sentences in [21, 22]) in both the source language and its translation. Our work seeks to fill this gap by assembling a sentence alignment tool for ancient language texts and their translations.

3. Method

We describe our pipeline to align parallel sentences across source texts and their translations, which consists of the following steps: a) segment the texts into sentence-level chunks via automatic heuristics and language-specific sentence segmentation tools like stanza [23], b) compute low-dimensional representations (embeddings) for the segmented sentences that are specifically designed to be informative about semantic similarity between sentences in multiple languages, c) use the embeddings across the sentences from a source text and its translation to guide prediction of sentence alignments.

3.1. Sentence Embeddings

We tested two models for obtaining sentence representations: LASER [24] and LaBSE [5]. Both are multilingual sentence embedding models, which support 200[3] and 109 languages, respec-

[2]The Open Greek and Latin Project's collections of translations can be found on their GitHub page: https://github.com/OpenGreekAndLatin

[3]LASER was originally trained on 93 languages; this encoder is accessible as "LASER2" and is the one we used for our experiments. The authors have since released additional "LASER3" encoders that each focus on an additional language.

tively. The authors of Sentence-BERT (S-BERT), who extended S-BERT so that it could embed sentences in multiple languages in [25], performed extensive experiments on a diverse set of tasks comparing the performance of eight sentence embedding models [25]. They found that LASER and LaBSE performed best for retrieving exact translations (BUCC bitext mining task), while their S-BERT-based model performed best for retrieving semantically similar sentences that are not exact translations of each other. We determined that our task most resembled bitext mining across the tasks investigated by [25], though the presence of non-literal translations in our dataset may challenge this assumption. Therefore, we focused our comparison on LaBSE and LASER, since these two methods have been shown to be better at *finding translation pairs* as opposed to only focusing on semantic similarity.

Both LASER and LaBSE leverage vast amounts of sentence-level parallel data to map sentences that are translations of each other to a shared low-dimensional manifold such that their representations are close to each other. LASER is an encoder-decoder LSTM model trained using a translation task with publicly available parallel data across numerous language pairs. The encoder is shared across the languages and the datasets. In contrast, LaBSE is a BERT-based [26] dual encoder model that is trained via a simpler translation-ranking task, which aims to increase similarity between sentences that are translations of each other and decrease similarity between the pairs of sentences that do not translate to each other. Hence, LaBSE yields embeddings (or vectors) for the parallel source and target sentences that are each encoded separately through a 12-layer transformer embedding network. For a translation pair, these cross-lingual source-target embeddings are trained to be similar to each other. In addition, LaBSE also makes use of extensive *masked language modeling pretraining* over both parallel bitext sentences and large amounts of unpaired monolingual data in numerous languages.

We tested both LASER and LaBSE without further fine-tuning. Surprisingly, our experiments suggested that LaBSE was able to generalize well to *unsupported languages* like Ancient Greek and Latin that we focus on in this paper. In addition, as we suggest in the Discussion section, we believe that a larger share of errors, especially when working with noisy, ancient texts, would be mitigated by improvements to the alignment algorithm. We note this in our Conclusion as a possible direction for future work.

3.2. Sentence Alignment

For sentence alignment, we relied on Vecalign [6], which remains the state-of-the-art model though it was published in 2019. Vecalign computes affinity scores between sentences or groups of sentences from a source and target text. The algorithm takes in sentence embeddings as input and then uses these embeddings to assess the similarity of sentences. Then, it reduces the problem of alignment to enumeration over all possible pairs of groups of sentences across the source and target texts. This process yields highest scoring pairs, resulting in either one-to-one, one-to-many, many-to-one, or many-to-many aligned sentence pairs. This enumeration is exponentially expensive in terms of the size of the bitext (pair of source and target texts). Therefore, a dynamic program is used to perform it efficiently. Further approximations are made to reduce the runtime by incorporating inductive biases and modeling assumptions, such as the largely monotonic nature of sentence alignment across the bitext. Another source of drastic reduction in runtime is a coarse-to-fine approach employed for alignment, which prunes the search space of sentence pairs aggressively by making severe contiguity and monotonicity assumptions. These approximations and assumptions reduce the runtime to being asymptotically linear in terms of the size of the corpus. However, as we show in our experiments, these assumptions might not necessarily hold for our task. The texts we are interested in exhibit significant non-monotonicity, discontiguousness, and noise in the form of extraneous material (paratext) that is interspersed throughout the content sentences that actually align across the bitext. Our findings point toward future research on better alignment algorithms which make fewer of these unrealistic assumptions for our task while remaining practical to execute.

4. Dataset

All ancient Greek and Latin texts were extracted from The Perseus Project [27] (Perseus) and contain no paratext. The translations varied in terms of language, sources, formats, available annotations, and often included paratext as detailed in Table 1. The texts are further described by annotation level, since this impacted their use in different experiments. We aimed to test our pipeline on a varied set of texts, including different styles (dialogue, poetry, prose) and level of noisiness (texts with and without paratext). Therefore, we used available annotations, leading some experiments to be evaluated at the chunk level (Table 2), sentence level (Table 3), or at coarser levels (chapter or book: Table 4). The one exception is the small test set (Table 3) that the authors of this paper manually annotated at the sentence level. We used this test set at the outset of our project for rapid testing in order to inform next steps. Details on the paratext present in the two noisy texts can be found in Table 4.

4.1. Preprocessing: Chunk-Level and Sentence Segmentation

When we refer to "chunk-level," we mean the most fine-grained citation structure available on Perseus, such as chapters, sections, or Stephanus pages.[4] By "sentence-level," we mean the phrases obtained after sentence segmentation, explained below. Any additional pre- and post-processing is listed in the appendix.

Table 2 lists all texts which were previously annotated at the chunk level. Preprocessing on these texts was limited to concatenating the chunks into a continuous string. For the sentence-level experiments (Tables 3, 4), we applied standard preprocessing to all texts: concatenated the raw text into one string, then segmented into sentences. For languages supported by Stanza [28] (Latin, English, French), we first split the text into Stanza's sentences, then split further on semi-colons and colons. For unsupported languages (Ancient Greek), we segmented ourselves by splitting on periods, semi-colons, and colons.

[4]Stephanus pagination refers to the page breaks used in modern editions and translations of the works of Plato. They were first established by a 1578 edition published by Henri Estienne, also known as Henricus Stephanus.

Table 1
Summary of Texts Used

Work	Translator	Language	Paratext	Annotation	Annotator
Crito	(Source)	Greek	no	section	Perseus
	Fowler	English	no	section	student
	Jowett	English	no	section	student
	Schleiermacher	German	no	section	student
	Mohammadi (student)	Farsi	no	section	student
Thucydides	(Source)	Greek	no	section	Perseus
	Crawley	English	no	section	Perseus
	Bétant	French	yes	chapter	OGL
Lucretius	(Source)	Latin	no	Perseus card	Perseus
	Watson, Good *(2 translations in 1 volume)*	English	yes	book	OGL

Table 2
Datasets for Chunk-Level Evaluations: Chunks are sections of Stephanus pages (*Crito*) or sections (Thucydides). We report space-separated tokens for comparability across languages.

Work	Translator	# Chunks	Avg. Tok./Chunk	Std. Dev.
Crito	(Source)	268	15.96	14.73
	Fowler	267	21.29	19.99
	Jowett	259	20.59	18.85
	Schleiermacher	267	20.70	19.72
	Mohammadi	267	21.42	20.57
Thucydides	(Source)	3575	41.96	19.77
	Crawley	3575	56.52	27.96

Table 3
Small Test Set: Dataset for Sentence-Level Evaluation

Work	Translator	# Sentences	Avg. Tok./Sent.	Std. Dev.	Annotator
Crito	(Source)	60	18.08	16.65	Paper authors
	Fowler	66	22.05	18.10	Paper authors
	Jowett	79	17.61	14.66	Paper authors

4.2. Annotations

4.2.1. *Crito*

Students involved in a project at Lepizig University annotated thirteen translations of *Crito* (including two in English, one in German, and five in Persian), to match the chunk-level annotation of the Greek text in Perseus. In our experiments, we used both English translations (by Harold North Fowler [7] and Benjamin Jowett [8]), the one German translation (by Schleier-

Table 4
Noisy Data: Dataset for Sentence-Level Experiments with Coarse-Level Evaluation

Work	Translator	# Sent.	Avg.Tok./Sent.	Std. Dev.	Text Sent.	Paratext Sent.
Thucydides	(Source)	6097	24.63	16.77	6097	0
	Bétant	17203	14.23	10.51	10958	6245
Lucretius	(Source)	2428	20.20	13.14	3575	0
	Watson & Good	14648	13.83	11.61	8521	6127
	Watson only	9815	15.82	13.40	3936	5879
	Good only	4833	11.00	7.59	4584	248

macher [9]), and one of the Persian translations (by Mohammadi, published on Zenodo [29]). This Persian translation was done by one of the student annotators.[5]

4.2.2. The Open Greek and Latin Project (OGL)

The two noisy translations in our dataset were annotated by The Open Greek and Latin Project. These texts are available to the public in XML files that include varying levels of annotation.[6] For the Thucydides French translation by Bétant [11], annotations include tags for paratext and book and chapter boundaries. For the Lucretius edition, which contains two English translations by Rev. John Selby Watson and John Mason Good [12], annotations include tags for paratext and book boundaries.

4.3. Noisy Data Features

Both the Thucydides (fr) and Lucretius translations include paratext. In Thucydides, paratext consists of a foreword; commentary and summary of contents preceding each of the eight books; 33 footnotes; and an index. In the Lucretius edition, paratext consists of a foreword, a commentary before the prose translation, 1426 notes including 955 footnotes interspersed through the text (70 in the foreword, 861 in the Watson translation, 24 in the Good translation), and an index. The Lucretius edition contains additional noise in the form of two translations included in one edition.

5. Experiments

Our experiments are summarized in Table 5. To guide the direction of our research, we first ran our two candidate pipelines, LASER-Vecalign and LaBSE-Vecalign, on the *Crito* test set. We then validated these results with retrieval experiments on our full dataset, using available annotations. Lastly, we tested our best pipeline, LaBSE-Vecalign, on noisy data with a focus

[5] The student translators used treebanks, commentaries, lexicon entries, and English and German translations to aid their Persian translations. They also aligned their Persian translations at the word level to the Greek, using Ugarit (Mohammadi's can be found here: https://ugarit.ialigner.com/userProfile.php?userid=52434&tgid=9362).
[6] https://github.com/OpenGreekAndLatin

Table 5
Summary of Experiments

Goal	Experiments	Texts Used	Annotation	Evaluation
Initial rapid testing	LASER-Vecalign & LaBSE-Vecalign	Crito test set	Sentence	Scoring Functions (see 5.3)
Validate initial results	LASER-Vecalign & LaBSE-Vecalign	*Crito* (Fowler, Jowett)	Chunk	Recall
	Chunk retrieval	*Crito* (all), Thucydides (en)	Chunk	Recall
	Sentence retrieval	*Crito* test set	Sentence	Recall
	Sentence retrieval	Thucydides (fr), Lucretius (en)	Sentence	Recall (coarse)
Test best pipeline on noisy data	LaBSE-Vecalign	Thucydides (fr), Lucretius (en)	Sentence	Recall (coarse)

on error analysis to understand how the pipeline would fare on the type of unannotated data that we would like to align using our pipeline.

5.1. Experimental Set-Up

To run the retrieval experiments, we first passed pre-processed source and target texts segmented into chunks or sentences through an embedding model (LASER or LaBSE). Then we used cosine similarity to retrieve the most similar chunks or sentences across the bitext pairs. For chunk-level retrieval experiments, we left out from the source text any chunks with missing translations (hence the different number of chunks across translations in Table 7).

For LASER-Vecalign and LaBSE-Vecalign, we passed the chunk- or sentence-level embeddings into Vecalign, which outputs a set of predicted alignments. These may include one-to-one, one-to-many, many-to-one, or many-to-many alignments.

5.2. Evaluation

For chunk-level retrieval, we report the percentage of source chunks with the correct result among the top 1 (correct result has the highest similarity score) and top 10 similarity scores. For sentence-level retrieval, we modify the numerator and denominator to account for one-to-many sentence alignments in the ground truth: for every Greek sentence, we compute the number of correct sentences retrieved divided by the number of target sentences in the true alignment.

For LASER-Vecalign and LaBSE-Vecalign, we evaluated results based on available annotations. For the *Crito* test set, where we manually produced sentence-level ground truth, we used Vecalign's scoring functions and two new functions we formulated, described below. For chunk alignment experiments, where the ground truth is a straightforward list of one-to-one alignments, we report the percentage of incorrect predictions. Finally, for LaBSE-Vecalign on the two noisy texts, Thucydides (fr) and Lucretius (en), we evaluate the pipeline's predictions at the coarse level of annotations available in the Open Greek and Latin Project's database. These experiments were run on segmented sentences but we report accuracy relative to a predicted sentence belonging to the same chapter (Thucydides) or book (Lucretius) as the source text.

5.3. Scoring Functions For Vecalign Predictions

Vecalign's original scoring function reports strict and lax scores for Precision, Recall, and F1. The lax metric expands the definition of true positives to include any correct sentence alignment in a one-to-many or many-to-many prediction. A strict true positive requires exact matches between a ground-truth alignment and a prediction. When performing our initial rapid testing, we formulated two additional metrics. Both are based on post-processing Vecalign's results by merging predicted alignments to try to reconstitute Perseus sections. We used this annotation level as ground truth because Perseus sections are examples of the available annotations applied to ancient texts by editorial convention. Therefore, we sought to determine how the pipelines would perform at this challenging level.

If after merging there's a strict match, then this is a true positive under the "New Strict Scoring" function. In the "New Lax Scoring" function, we instead look for lax matches: for any sentence that appears on both sides of the reconstituted alignment that is in a Perseus section, true positives are increased by one.

6. Results

6.1. LaBSE Outperformed LASER

LaBSE outperformed LASER in our initial testing of the candidate pipelines, LASER-Vecalign and LaBSE-Vecalign. This was repeated in our validation experiments (chunk-alignment, chunk-retrieval, and sentence-retrieval).

When used in conjunction with Vecalign, LaBSE consistently outperformed LASER (Table 6). Unsurprisingly, both pipelines did best under the New Lax Scoring function. Since this metric gives credit for (counts as true positive) any correct sentence matching between source and target text, it most approximates the pipelines' ability to correctly align sentences. Therefore, at the sentence-level, the results on clean data with the *Crito* were overall very promising. LaBSE's and LASER's worst results were with the Strict Scoring function, confirming our hypothesis that existing pipelines would struggle to return correct alignments at the more challenging annotation level that we find in ancient texts and their translations. Results for both pipelines at the chunk level are in Table 12 in the appendix (LaBSE-Vecalign aligned all chunks correctly, while LASER-Vecalign aligned 98.88% of Greek - Fowler chunks and 94.98% of Greek - Jowett chunks correctly).

In the retrieval experiments, LASER struggled most when retrieving chunks from the English Thucydides translation, the longest text on which we tested LASER (Table 7). LASER's performance improved when retrieving sentences from the *Crito* test set (Table 8), in other words when retrieving shorter spans of text from a shorter document. We do not see similar difference with LaBSE's performance on shorter text spans and documents, with its scores

Table 6
LASER vs. LaBSE: Sentence alignment with Vecalign from the ancient Greek (G) of Plato's *Crito* to the English translations of Fowler (F) and Jowett (J) and between English translations

	G–F	G–F	G–J	G–J	F–J	F–J
Embedding Model Used	LabSE	LASER	LabSE	LASER	LabSE	LASER
Vecalign Strict Scoring						
Precision - Strict	0.737	0.655	0.600	0.414	0.536	0.472
Recall - Strict	0.778	0.704	0.623	0.453	0.577	0.481
F1 - Strict	0.757	0.679	0.611	0.432	0.556	0.476
New Strict Scoring						
Accuracy	0.870	0.796	0.685	0.500	0.712	0.577
New Lax Scoring						
Precision - Lax	0.984	0.944	0.950	0.791	0.972	0.966
Recall - Lax	0.827	0.778	0.805	0.647	0.727	0.714
F1 - Lax	0.899	0.853	0.871	0.712	0.832	0.821

Table 7
LASER vs. LaBSE: Chunk-Level retrieval from ancient Greek to translations of *Crito* (*Cr.*) and Thucydides (Thuc.)

	Fowler (*Cr.*, en)	Jowett (*Cr.*, en)	Schleiermacher (*Cr.*, de)	Mohammadi (*Cr.*, fa)	Crawley (Thuc., en)
Num. Chunks	267	259	267	267	3575
LaBSE					
Top 1	75.66%	57.92%	78.65%	71.91%	79.61%
Top 10	93.26%	83.40%	94.01%	87.27%	93.15%
LASER					
Top 1	33.71%	14.67%	34.46%	23.60%	3.05%
Top 10	67.42%	46.72%	68.16%	44.94%	8.98%

at the sentence level slightly lower yet still comparable to those at the chunk level (Table 7, Table 8).

6.2. Performance Differences on Clean Data

LaBSE and LASER both exhibited performance differences across translations of the *Crito* in experiments using the test set and the full text. With the test set, LASER-Vecalign and LaBSE-Vecalign did best aligning the Greek to Fowler's more literal translation (Table 6), where there are the most one-to-one alignments in the ground truth (70% compared to 48% for Greek - Jowett and Fowler - Jowett). Likewise, both LaBSE and LASER had higher scores retrieving sentences between the Greek and Fowler than between the Greek and Jowett. Interestingly, both models perform better retrieving sentences between Fowler and Jowett ("F-J" in Table 8) than when used in conjunction with Vecalign ("F-J" in Table 6). At the chunk level, LaBSE

Table 8
LASER vs. LaBSE: Sentence-Level retrieval from the ancient Greek (G) of Plato's *Crito* to the English translations of Fowler (F) and Jowett (J) and between English translations

	G–F	G–J	F–J
LaBSE			
Top 1	76.67%	51.39%	73.58%
Top 10	90.83%	82.54%	92.17%
LASER			
Top 1	41.67%	30.83%	61.24%
Top 10	67.50%	51.17%	78.34%

Table 9
Sentence-Level retrieval Thucydides, Lucretius (noisy data)

	Thucydides (el - fr) (from same chapter)	Lucretius (la - en) (from same book)
LaBSE		
Top 1	36.49%	76.52%
Top 10	60.70%	97.24%

and LASER also did better on Fowler's more literal translation in both chunk-level retrieval (Table 7) and chunk-level alignment (Table 12 in Appendix).

6.3. Challenges of Noisy Data

Testing LaBSE and LaBSE-Vecalign on noisy data points to the challenges of the two noisy features present in our noisy dataset: the presence of paratext and multiple translations in the target document. LaBSE did worse retrieving sentences from the noisy Thucydides translation (60.70% in Top 10 in Table 9) than the English translation (93.15% in Top 10 in Table 7), even though the evaluation was done at a coarser level with the French. These experiments differ in three ways: the language of the target (French vs. English), the shorter text span in the sentence retrieval experiment, and the presence of paratext in the French. Given LaBSE's comparable results with retrieval at the chunk and sentence level on the *Crito*, the presence of paratext is the likely reason for the performance difference. At the coarser, book level we used to evaluate sentence retrieval with Lucretius, the presence of paratext no longer made a visible impact (Table 9).

When used in conjunction with Vecalign, and also evaluated at the same coarse level, LaBSE-Vecalign did very well on Thucydides (fr), despite the presence of paratext. Thus 96.72% of paratext sentences correctly aligned to null and 94.04% of French text sentences were aligned to a Greek sentence from the same chapter (Table 10). However, on Lucretius 35.03% of paratext sentences were incorrectly aligned to null and 55.23% of English text sentences incorrectly aligned to null (Table 10). We also reported results using the number of Vecalign predictions as denominator, which can be found in Table 13 in the Appendix and show a similar pattern.

Table 10
LaBSE - Vecalign: By Number of Target Sentences, Thucydides (fr - el, chapter-level evaluation) and Lucretius (en - la, book-level evaluation)

Type of Alignment	Thucydides (chapter-level)	Lucretius (book-level)
As a percent of Paratext Sentences		
Number of paratext sentences	6245	6127
Paratext sents. to null (correct)	96.72%	64.97%
Paratext sents. to source text sents. (incorrect)	3.28%	35.03%
As a Percent of Target Text Sentences		
Number of text sentences	10958	8521
Text-to-text: to source sents. from same chapter or book	94.04%	44.24%
Text-to-text: to at least 1 source sent. from same chapter or book	0.93%	0.49%
Errors, text-to-text: to no source sents. from same chapter or book	4.93%	0.04%
Errors, text-to-null	0.10%	55.23%

Table 11
LaBSE - Vecalign: Lucretius (en - la) - No Paratext, First Translation Only (book-level evaluation)

Type of Alignment	Lucretius (book-level)
As a Percent of Text Sents. from First Translation Only	
Number of Text Sentences	3648
Text-to-text: to source sents. from same book	97.97%
Text-to-text: to at least one source sent. from same book	1.67%
Errors, Text-to-text: to no source sent. from same book	0.00%
Errors, Text-to-Null	0.36%

In order to get results for Lucretius using LaBSE-Vecalign that are comparable to those on Thucydides, we had to suppress both paratext and the second translation (by Good) in the target edition. When we only suppressed paratext, 65.56% of Vecalign's predictions were (incorrect) null-to-text alignments, and 55.21% of English text sentences were incorrectly aligned to null (Table 14 in the Appendix). When we only suppressed the second translation, 96.26% of text sentences were correctly aligned to Latin sentences from the same book, but 36.26% of paratext sentences were incorrectly aligned to Latin text sentences (Table 15 in the Appendix). Finally, when we suppressed paratext and only counted text sentences from the first translation (by Watson), 97.97% of English text sentences aligned to Latin sentences from the same book, and no English text sentences aligned to Latin sentences from a different book (Table 11). In other words, only after preprocessing the English edition of Lucretius to obtain a clean dataset (no paratext, no second translation) was LaBSE - Vecalign able to achieve results comparable to those we saw with the French translation of Thucydides.

Interestingly, when we compare against retrieval results in Table 9, LaBSE did better on its own than it did with Vecalign on Lucretius, but worse on Thucydides. Thus using LaBSE embeddings only, for 97.24% of the 2428 Latin sentences we retrieved English text sentences from the same book among the top 10 most similar sentences, while LaBSE-Vecalign aligned only 44.24% of the 8521 English text sentences to Latin sentences from the same book. On the other hand with Thucydides, we retrieved French text sentences from the same chapter among the top 10 most similar sentences for 60.70% of the 6097 Greek sentences using LaBSE embeddings only, while LaBSE-Vecalign aligned 94.04% the 10958 French text sentences to Greek sentences from the same chapter. These results indicate that Vecalign's faulty assumption and inductive biases for approximate dynamic programming can override the signal from well-trained sentence representation methods and hurt the overall performance in the case of discontiguous noisy text.

7. Discussion

Our experiments led us to identify LaBSE-Vecalign as the pipeline using existing models best suited to produce sentence-level alignments of Ancient Greek and Latin texts with their translations. However, even with the clean data of the *Crito*, this pipeline evidenced differences in performance across translations. This was even the case when aligning English to English (Fowler-Jowett), with LaBSE-Vecalign showing better results aligning Greek-Fowler than Fowler-Jowett. The performance driver with *Crito* seems therefore to be the nature of the translations themselves: the number of one-to-many and many-to-many alignments in the ground truth. The differences in the two translations also emerge at the word level. We manually aligned the first half of the *Crito*'s words with Jowett's translation on Ugarit [30] and compared to alignments with Fowler prepared by an author of Ugarit's alignment guidelines [31][7]. We were able to align 52% of Greek words with Jowett's translation, compared to 80% with Fowler's (Table 16 in the Appendix). Of these aligned Greek words, with Jowett 9% crossed the predicted sentence boundaries while none did with Fowler's (Table 17 in the Appendix).

With more realistic noisy data, LaBSE-Vecalign was not able to handle both the nature of the paratext in the Lucretius edition, and its inclusion of a second translation. The Lucretius edition not only has many footnotes, these are also lengthy. Thus the first two sentences of the translation in Lucretius (following 461 sentences of the foreword and 23 sentences of commentary preceding book 1) are followed by 44 sentences of footnotes before we find the third text sentence. In contrast, Thucydides had fewer footnotes (33), also interspersed in the translated text but all short (the longest spans 2 sentences).

Figures 1 and 2 show details of the LaBSE-Vecalign results aligning Thucydides and Lucretius with their French and English translations, respectively. In the Thucydides detail, the Greek sentences on the left are much longer than the French sentences; the fifth Greek sentence (section 1.2.2) is translated over five sentences in French (last row of Figure 1). The first three rows are errors; the first two contain no overlapping sentences and the third includes the correct French sentences for section 1.1.3 as well as the French translations of about half of section 1.1.1 and all of section 1.1.2. There is no evident pattern explaining these errors. However,

[7]Word alignments with Jowett: https://ugarit.ialigner.com/userProfile.php?userid=126388&tgid=12065 and Fowler: https://ugarit.ialigner.com/userProfile.php?userid=3&tgid=8609

Greek: sentence index and section	Greek text	French: sentence index and section	LaBSE - Vecalign results: French sentences aligned to the Greek
Index: 0 Section 1.1.1	Θουκυδίδης Ἀθηναῖος ξυνέγραψε τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων, ὡς ἐπολέμησαν πρὸς ἀλλήλους, ἀρξάμενος εὐθὺς καθισταμένου καὶ ἐλπίσας μέγαν τε ἔσεσθαι καὶ ἀξιολογώτατον τῶν προγεγενημένων, τεκμαιρόμενος ὅτι ἀκμάζοντές τε ᾖσαν ἐς αὐτὸν ἀμφότεροι παρασκευῇ τῇ πάσῃ καὶ τὸ ἄλλο Ἑλληνικὸν ὁρῶν ξυνιστάμενον πρὸς ἑκατέρους, τὸ μὲν εὐθύς, τὸ δὲ καὶ διανοούμενον.	Index: 124-130 Book 1 table of contents	— Épidamne et Potidée. Événements qui provoquèrent la guerre du Péloponèse. Affaire d'Épidamne·, guerre entre Corcyre et Corinthe; premier combat naval, chap. xxiv-xxxi. Les Gorcyréens obtiennent l'alliance d'Athènes. Discours des Corcyréens et des Corinthiens, chap. xxxiixliii. Second combat naval entre les Corcyréens et les Corinthiens;
Index: 1 Section 1.1.2	κίνησις γὰρ αὕτη μεγίστη δὴ τοῖς Ἕλλησιν ἐγένετο καὶ μέρει τινὶ τῶν βαρβάρων, ὡς δὲ εἰπεῖν καὶ ἐπὶ πλεῖστον ἀνθρώπων.	Index: 132-133 Book 1 table of contents	Défection de Potidée; combat livré sous les murs de cette ville et siége commencé par les Athéniens, chap. lvi-lxvi.
Index: 2 Section 1.1.3	τὰ γὰρ πρὸ αὐτῶν καὶ τὰ ἔτι παλαίτερα σαφῶς μὲν εὑρεῖν διὰ χρόνου πλῆθος ἀδύνατα ἦν, ἐκ δὲ τεκμηρίων ὧν ἐπὶ μακρότατον σκοποῦντί μοι πιστεῦσαι ξυμβαίνει οὐ μεγάλα νομίζω γενέσθαι οὔτε κατὰ τοὺς πολέμους οὔτε ἐς τὰ ἄλλα.	Index: 153-158 Sections 1.1.1 - 1.1.3	Il s'est mis à l'œuvre dès le début de cette guerre, en présumant qu'elle surpasserait en grandeur et en importance toutes celles qui ont précédé. Ce qui le lui faisait croire, c'est que ces deux nations étaient alors au faîte de leur puissance, et qu'il voyait le reste des Grecs ou prendre parti dès l'origine pour l'une ou pour l'autre, ou en former le projet. C'est en effet le plus vaste conflit qui ait jamais ébranlé la Grèce, une partie des pays barbares et, pour ainsi dire, le monde entier. La distance des temps ne permet pas de discerner bien clairement les événements antérieurs ou d'une époque plus reculée; néanmoins, d'après les indices que mes investigations m'ont mis à même de recueillir en remontant jusqu'à la plus haute antiquité, j'ai lieu de croire que ces événements furent peu considérables sous le rapport militaire, comme à tout autre égard. II.
Index: 3 Section 1.2.1	φαίνεται γὰρ ἡ νῦν Ἑλλὰς καλουμένη οὐ πάλαι βεβαίως οἰκουμένη, ἀλλὰ μεταναστάσεις τε οὖσαι τὰ πρότερα καὶ ῥᾳδίως ἕκαστοι τὴν ἑαυτῶν ἀπολείποντες βιαζόμενοι ὑπό τινων αἰεὶ πλειόνων.	Index: 159 Section 1.2.1	Le pays qui porte aujourd'hui le nom de Grèce, ne fut pas primitivement habité d'une manière stable, mais il fut le théâtre de fréquentes migrations.
Index: 4 Section 1.2.2	τῆς γὰρ ἐμπορίας οὐκ οὔσης, οὐδ' ἐπιμειγνύντες ἀδεῶς ἀλλήλοις οὔτε κατὰ γῆν οὔτε διὰ θαλάσσης, νεμόμενοί τε τὰ αὑτῶν ἕκαστοι ὅσον ἀποζῆν καὶ περιουσίαν χρημάτων οὐκ ἔχοντες οὐδὲ γῆν φυτεύοντες, ἄδηλον ὂν ὁπότε τις ἐπελθὼν καὶ ἀτειχίστων ἅμα ὄντων ἄλλος ἀφαιρήσεται, τῆς τε καθ' ἡμέραν ἀναγκαίου τροφῆς πανταχοῦ ἂν ἡγούμενοι ἐπικρατεῖν, οὐ χαλεπῶς ἀπανίσταντο, καὶ δι' αὐτὸ οὔτε μεγέθει πόλεων ἴσχυον οὔτε τῇ ἄλλῃ παρασκευῇ.	Index 160-164 Section 1.2.2	On abandonnait sans peine ses demeures, pour faire place à de nouveaux flots d'arrivants. Comme il n'y avait aucun commerce, aucune communication assurée ni par terre ni par mer; que chacun exploitait le sol uniquement dans la mesure de ses besoins, sans penser à s'enrichir, sans même faire de plantations (car avec des villes ouvertes, on ne savait jamais si les récoltes ne seraient pas enlevées par des ravisseurs étrangers); enfin, comme on espérait trouver partout la subsistance journalière, on émigrait sans difficulté. Aussi la Grèce n'avait-elle ni grandes villes, ni aucun des éléments essentiels de la puissance.

Figure 1: Behavior of LaBSE - Vecalign aligning first 5 sentences of the Thucydides Greek to the French: rows shaded in green are correct alignments; remaining errors are shown through color-coded text

with Lucretius, some errors can be explained (and excused), for example the first alignment in Figure 2: the footnotes that are also aligned to the first sentence of the Latin include part of the translated text ("O Bountiful Venus") and Latin original ("Alma Venus"). As with Thucydides, the remaining errors occurred across sentences covering similar subjects.

Figure 2 suggests that the errors with Lucretius can be attributed to Vecalign: to correctly align all blue English sentences to the first Latin sentence, the algorithm would have to skip sentences 485-527 and then recognize the fragment in sentence 528 as part of the phrase begun in sentence 484. Vecalign was not designed to handle this case; its approximation to run in linear time averages the embeddings of consecutive sentences using a relatively small window (10 sentences in the default settings that we used). The blue sentences are too far apart in the English edition for Vecalign to capture them in one alignment. This is reminiscent of an error we encountered in the LASER-Vecalign results aligning the *Crito* with Jowett's English translation when a sentence in Jowett (in red) was out of order relative to the Greek (Figure 3).

8. Conclusion and Future Work

The extensive corpus of translations of the Greek and Latin classics holds tremendous promise for the study of translation, natural language processing, and variation and change in cultural assumptions. By studying both close translations and those full of literary license, paraphrase, censorship, and misunderstanding, we hope to enable scholars and students to understand this corpus better and translation-studies and NLP researchers to perform empirical studies of variation in translation. Our experiments demonstrate that a pipeline composed of state-of-the-art NLP systems for performing automatic sentence alignment on literary text in ancient languages is useful but leaves a lot of room for improvement. Even our best-performing con-

Latin: sentence index and section	Latin text	English: sentence index and section	LaBSE - Vecalign results: English sentences aligned to the Latin
Index: 0 Book 1	Aeneadum genetrix, hominum divomque voluptas, alma Venus, caeli subter labentia signa quae mare navigerum, quae terras frugiferentis concelebras, per te quoniam genus omne animantum concipitur visitque exortum lumina solis:	Index: 481-482 Foreword Index: 483-484 English prose translation Index: 485-487 Book 1 notes	and concludes with a praise of philosophy, which assists mankind to penetrate the mysteries of nature. a O BOUN riFUL Venus,1 mother of the race of Æneas,2 delight of gods and men, who, beneath the gliding constellations of heaven,3 fillest with life4 the ship-bearing sea and the fruit-producing earth; 5 since by thy influence every kind of 1 O bountiful Venus.] Ver. 2. Alma Venus. The word means kind, bountiful, benignant, nourishing, from alo, to nourish. "
Index: 1 Book 1	te, dea, te fugiunt venti, te nubila caeli adventumque tuum, tibi suavis daedala tellus summittit flores, tibi rident aequora ponti placatumque nitet diffuso lumine caelum.	Index: 525-527 Book 1 notes Index: 528-531 English prose translation	—Comfort bring and mirth To the ship-bearing seas, corn-bearing earth. Fruges, however, means fruits of the earth in general. living creature is conceived, and, springing forth, hails the light of the sun. Thee, 0 goddess, thee the winds flee ; be fore thee, and thy approach, the clouds of heaven disperse; for thee the variegated earth2 puts forth3 her fragrant flowers;
Index: 2 Book 1	nam simul ac species patefactast verna diei et reserata viget genitabilis aura favoni, aeriae primum volucris te, diva, tuumque significant initum perculsae corda tua vi.	Index: 532-533 English prose translation	on thee the waters of ocean smile, and the calmed heaven beams with effulgent4 light. For, as soon as the vernal face of day5 is unveiled, and the genial gale of Favonius exerts its power unconfined, the birds of the air first, 0 goddess, testify of thee and thy coming, smitten in heart by thy influ ence.
Index: 3 Book 1	inde ferae pecudes persultant pabula laeta et rapidos tranant amnis:	Index: 534 English prose translation	Next, the wild herds bound over the joyous pastures, and swim across the rapid streams.
Index: 4 Book 1	ita capta lepore te sequitur cupide quo quamque inducere pergis.	Index: 535 English prose translation	So all kinds of living creatures, captivated by thy charms and thy allurements, eagerly follow thee whithersoever thou proceedest to lead them.

Figure 2: Behavior of LaBSE - Vecalign aligning first 5 sentences of the Lucretius Latin to the English: rows shaded in green are correct alignments; remaining errors are shown through color-coded text

Ground truth		Greek		Jowett
[16] : [32, 33, 34]	16	"σὺ δὲ οὔτε Λακεδαίμονα προῃροῦ οὔτε Κρήτην, ἃς δὴ ἑκάστοτε φῂς εὐνομεῖσθαι, οὔτε ἄλλην οὐδεμίαν τῶν Ἑλληνίδων πόλεων οὐδὲ τῶν βαρβαρικῶν, ἀλλὰ ἐλάττω ἐξ αὐτῆς ἀπεδήμησας ἢ οἱ χωλοί τε καὶ τυφλοὶ καὶ οἱ ἄλλοι ἀνάπηροι"	32	"You had your choice, and might have gone either to Lacedaemon or Crete, which you often praise for their good government, or to some other Hellenic or foreign State."
[17] : [33]	17	οὕτω σοι διαφερόντως τῶν ἄλλων Ἀθηναίων ἤρεσκεν ἡ πόλις τε καὶ ἡμεῖς οἱ νόμοι δῆλον ὅτι	33	"Whereas you, above all other Athenians, seemed to be so fond of the State, or, in other words, of us her laws (for who would like a State that has no laws?), that you never stirred out of her"
[18] : [33]	18	τίνι γὰρ ἂν πόλις ἀρέσκοι ἄνευ νόμων	34	"the halt, the blind, the maimed, were not more stationary in her than you were."
[19] : [35]	19	νῦν δὲ δὴ οὐκ ἐμμενεῖς τοῖς ὡμολογημένοις	35	And now you run away and forsake your agreements.
[20] : [36]	20	"ἐὰν ἡμῖν γε πείθῃ, ὦ Σώκρατες"	36	"Not so, Socrates, if you will take our advice"

Figure 3: Behavior of LASER - Vecalign when a sentence is out of order in Jowett's translation relative to the Greek, *Crito* (section 52e-53a)

figuration struggles when translations exhibit significant discontiguity and non-monotonicity with respect to the source text. We expect this to be the case for the majority of extant translations that we would like to analyze computationally. The performance of our pipeline is further compromised by the presence of paratext—footnotes, commentaries, alternate translations, quotes from the source, and other extraneous material. This kind of noise in the form of paratext is also a common feature of collections of translations like the one compiled by the Open Greek and Latin Project that we aim to process, align at multiple granularities, and analyze computationally. Another example of noisy collections is the series of nineteenth-century Hachette editions of the classics, which alone contains 46 editions with more than one French translation each.[8]

Considering these factors, we perceive two directions for future work. The first is to improve Vecalign's model to handle longer paratext and multiple translations by revisiting the assump-

[8]The Hachette series can be found on HathiTrust's website: https://babel.hathitrust.org/cgi/mb?a=listis;c=1152044995

tions it makes in pruning its search space and redesigning the dynamic program for alignment. The second approach is to build classification systems fine-tuned for the genre of classical translations. Since paratext annotations are not always available, this would allow us to automatically identify paratext and detect multiple translations before running books through our LaBSE-Vecalign pipeline. Fortunately, in the collection compiled by the Open Greek and Latin Project, some of this processing has been done, and the XML files tag the paratext explicitly. It remains to build models to detect editions with multiple translations and facing source and target texts using page-level language identification to remove these violations of Vecalign's continuity and monotonicity assumptions. We are currently preprocessing several translation collections of interest for running through our sentence alignment pipeline in order to release a large dataset linking the Greek and Latin classics and their translations at the sentence level.

Acknowledgments

We would like to thank the Open Greek and Latin Project for sharing their annotations of translations of Ancient Greek and Latin texts, as well as the student annotators of the *Crito* translations.

References

[1] G. Crane, The Perseus Digital Library and the future of libraries, International Journal on Digital Libraries 24 (2023) 117–128. doi:10.1007/s00799-022-00333-2.

[2] N. Coffee, An agenda for the study of intertextuality, Transactions of the American Philological Association 148 (2018) 205–223. doi:10.1353/apa.2018.0008.

[3] G. Crane, A. Babeu, L. M. Cerrato, A. Parrish, C. Penagos, F. Shamsian, J. Tauber, J. Wegner, Beyond translation: Engaging with foreign languages in a digital library, International Journal on Digital Libraries 24 (2023) 163–176. doi:10.1007/s00799-023-00349-2.

[4] Y. Assael, T. Sommerschield, J. Prag, Restoring ancient text using deep learning: a case study on Greek epigraphy, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6368–6375. doi:10.18653/v1/D19-1668.

[5] F. Feng, Y. Yang, D. M. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 878–891. doi:10.18653/v1/2022.acl-long.62.

[6] B. Thompson, P. Koehn, Vecalign: Improved sentence alignment in linear time and space, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1342–1348. doi:10.18653/v1/D19-1136.

[7] Plato, H. N. Fowler, W. Lamb, Plato in Twelve Volumes, Vol. 1 translated by Harold North

Fowler; Introduction by W.R.M. Lamb, volume 1, Harvard University Press and William Heinemann Ltd., Cambridge, MA and London, 1966.

[8] Plato, B. Jowett, Crito, translated by Benjamin Jowett, The Internet Classics Archive, 1994-2009. URL: https://classics.mit.edu/Plato/crito.html.

[9] Plato, F. Schleiermacher, Platons Werke, In der Realschulbuchhandlung, 1809.

[10] Thucydides, R. Crawley, History of the Peloponnesian War, J.M. Dent and E.P. Dutton, London and New York, 1910.

[11] Thucydides, Élie Ami Bétant, Histoire de la Guerre du Péloponnèse de Thucydide, Librairie de L. Hachette et Cie, 1863.

[12] T. Lucretius Carus, J. M. Good, J. S. Watson, Lucretius On the nature of things, G. Bell and sons, 1893.

[13] W. A. Gale, K. W. Church, A program for aligning sentences in bilingual corpora, Computational Linguistics 19 (1993) 75–102.

[14] P. F. Brown, J. C. Lai, R. L. Mercer, Aligning sentences in parallel corpora, in: 29th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Berkeley, California, USA, 1991, pp. 169–176. doi:10.3115/981344.981366.

[15] R. Sennrich, M. Volk, MT-based sentence alignment for OCR-generated parallel texts, in: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers, Association for Machine Translation in the Americas, Denver, Colorado, USA, 2010.

[16] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. O. Rojas, L. P. Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, J. Zaragoza, Paracrawl: Web-scale acquisition of parallel corpora, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4555–4567.

[17] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, M. Post, Multilingual TEDx corpus for speech recognition and translation, in: Proc. Interspeech 2021, 2021, pp. 3655–3659. doi:10.21437/Interspeech.2021-11.

[18] R. C. Moore, Fast and accurate sentence alignment of bilingual corpora, in: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers, Springer, Tiburon, USA, 2002, pp. 135–144.

[19] K. Thai, M. Karpinska, K. Krishna, B. Ray, M. Inghilleri, J. Wieting, M. Iyyer, Exploring document-level literary machine translation with parallel paragraphs from world literature, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 9882–9902.

[20] J. Uszkoreit, J. M. Ponte, A. C. Popat, M. Dubiner, Large scale parallel document mining for machine translation, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, 2010, p. 1101–1109.

[21] T. Yousef, C. Palladino, F. Shamsian, A. d'Orange Ferreira, M. F. dos Reis, An automatic model and gold standard for translation alignment of ancient greek, in: Proceedings of the 13th Conference on Language Resources and Evaluation, LREC 2022, European Language Resources Association, 2022, pp. 5894–5905.

[22] Z.-Y. Dou, G. Neubig, Word alignment by fine-tuning embeddings on parallel corpora, in:

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2112–2128. doi:10.18653/v1/2021.eacl-main.181.
[23] J. Bauer, C. Kiddon, E. Yeh, A. Shan, C. D. Manning, Semgrex and ssurgeon, searching and manipulating dependency graphs, in: Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023), Association for Computational Linguistics, Washington, D.C., 2023, pp. 67–73.
[24] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the Association for Computational Linguistics 7 (2019) 597–610. doi:10.1162/tacl_a_00288.
[25] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 4512–4525.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
[27] D. A. Smith, J. A. Rydberg-Cox, G. R. Crane, The Perseus Project: a digital library for the humanities, Literary and Linguistic Computing 15 (2000) 15–25. doi:10.1093/llc/15.1.15.
[28] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 101–108.
[29] F. Shamsian, S. Assimi, A. M. Sarabi, N. Mohammadi, K. Nikpour, F. Rahimi, Sentence-aligned student translations of Crito (Ancient Greek, English, German, Persian), 2023. doi:10.5281/zenodo.8273374.
[30] T. Yousef, C. Palladino, F. Shamsian, M. Foradi, Translation alignment with Ugarit, Information 13 (2022). doi:10.3390/info13020065.
[31] T. Yousef, C. Palladino, F. Shamsian, A. D. Ferreira, M. Reis, An automatic model and gold standard for translation alignment of ancient greek, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 5894–5905.

A. Description of Models

A.1. LASER

We installed LASER following instructions on the model's GitHub page.[9] Sentence embeddings are the output of LASER's encoder; we followed instructions for building them on the "embed" task page. Language-specific encoders are not available for the languages in our dataset, therefore we defaulted to LASER2. We used default parameters.

A.2. LaBSE

We accessed LaBSE through its HuggingFace implementation under the Sentence-Transformers class.[10] We used default paramaters.

A.3. Vecalign

We installed Vecalign following instructions on its GitHub page.[11] Running Vecalign on sentence embeddings requires a few steps. First, build "overlap" files from original text document (each row represents one sentence), or "concatenations of consecutive sentences" (each row represents one concatenation from one to *n* number of consecutive sentences). Second, embed the overlap files (each row is an embedding of a concatenation). Third, align sentences using the embeddings of concatenated consecutive sentences.

We tested the impact of two parameters using the *Crito* test set for Ancient Greek - Fowler's translation: number of overlaps (10 and 7) and maximum alignment size (8 and 5). Changing the parameters had no impact on the results, so we used Vecalign's default values (number of overlaps = 10 and max alignment size = 8).

B. Additional Pre- and Post-Processing

B.1. *Crito*

B.1.1. Data Extraction

Both the Greek and the annotated translations were extracted at the chunk level, concatenated into a continuous series for use in chunk-level experiments, and segmented into sentences following our standard preprocessing for use in sentence-level experiments. The Greek we extracted from Perseus and the annotated translations from a google sheet shared with us.

[9]Installation instructions can be found on LASER's GitHub homepage: https://github.com/facebookresearch/LASER/

[10]Documentation and instructions for using LaBSE's HuggingFace implementation can be found here: https://huggingface.co/sentence-transformers/LaBSE

[11]Installation and usage instructions can be found on Vecalign's GitHub homepage: https://github.com/thompsonb/vecalign

Table 12
LASER-Vecalign vs. LaBSE-Vecalign: Percentage of Correct Chunk Alignments from the ancient Greek (G) of Plato's *Crito* to the English translations of Fowler (F) and Jowett (J)

	G-F	G-J
Num. Chunks	267	259
LaBSE-Vecalign	100%	100%
LASER-Vecalign	98.88%	94.98%

B.2. Thucydides

B.2.1. Ancient Greek Text

Before preprocessing, we removed newline characters that appeared to be errors in the text extracted from Perseus. After segmentation, certain ending punctuation marks were split into the next sentence (e.g. right quote after speech or square bracket in footnote). We corrected this by adding the punctuation mark back to the preceding sentence.

B.2.2. French Translation by Bétant

Before segmentation, we removed the whitespace that by convention appears between text and French quotation marks (*guillemets*). Right *guillemets* were nevertheless frequently split onto their own sentence or appended to the beginning of the following sentence. We corrected this error after segmentation.

B.3. Lucretius

B.3.1. Latin Text

Before preprocessing, we also removed newline characters that appeared to be errors in the text extracted from Perseus.

C. Results of Chunk Alignment With LASER-Vecalign and LaBSE-Vecalign

Table 12 reports the results of aligning the two English translations of the *Crito* at the chunk level using LASER-Vecalign and LaBSE-Vecalign. Given the small number of errors, we have also reported them in full in Figures 4 and 5. Figure 4 shows errors from LASER - Vecalign aligning *Crito* to Fowler at the chunk level. Figure 5 shows errors from LASER - Vecalign aligning *Crito* to Jowett at the chunk level.

0059-003|220.|Cr. 52b.1.	οὐ γὰρ ἄν ποτε τῶν ἄλλων Ἀθηναίων ἁπάντων διαφερόντως ἐν αὐτῇ ἐπεδήμεις εἰ μή σοι διαφερόντως ἤρεσκεν, καὶ οὔτ᾽ ἐπὶ θεωρίαν πώποτ᾽ ἐκ τῆς πόλεως ἐξῆλθες, ὅτι μὴ ἅπαξ εἰς Ἰσθμόν, οὔτε ἄλλοσε οὐδαμόσε, εἰ μή ποι στρατευσόμενος, οὔτε ἄλλην ἀποδημίαν ἐποιήσω πώποτε ὥσπερ οἱ ἄλλοι ἄνθρωποι, οὐδ᾽ ἐπιθυμία σε ἄλλης πόλεως οὐδὲ ἄλλων νόμων ἔλαβεν εἰδέναι, ἀλλὰ ἡμεῖς σοι ἱκανοὶ ἦμεν καὶ ἡ ἡμετέρα πόλις:	for you would never have stayed in it more than all other Athenians if you had not been better pleased with it than they; you never went out from the city to a festival, or anywhere else, except on military service, and you never made any other journey, as other people do, and you had no wish to know any other city or other laws, but you were contented with us and our city.
0059-003|221.|Cr. 52b.2.	οὕτω σφόδρα ἡμᾶς ᾑροῦ καὶ ὡμολόγεις καθ᾽ ἡμᾶς πολιτεύσεσθαι, τά τε ἄλλα καὶ παῖδας ἐν αὐτῇ ἐποιήσω, ὡς ἀρεσκούσης σοι τῆς πόλεως.	So strongly did you prefer us and agree to live in accordance with us; and besides, you begat children in the city, showing that it pleased you.
0059-003|222.|Cr. 52c.1.	ἔτι τοίνυν ἐν αὐτῇ τῇ δίκῃ ἐξῆν σοι φυγῆς τιμήσασθαι εἰ ἐβούλου, καὶ ὅπερ νῦν ἀκούσης τῆς πόλεως ἐπιχειρεῖς, τότε ἑκούσης ποιῆσαι.	And moreover even at your trial you might have offered exile as your penalty, if you wished, and might have done with the state's consent what you are now undertaking to do without it.

Figure 4: Errors aligning *Crito* to Fowler's English translation

Table 13
LaBSE - Vecalign: By number of predicted alignments, Thucydides (el - fr, chapter-level evaluation) and Lucretius (la - en, book-level evaluation)

Type of Alignment	Thucydides (chapter-level)	Lucretius (book-level)
Number of Predicted Alignments	12046	11073
Text-to-text: all sentences from same chapter or book	45.29%	15.42%
Null-to-paratext	50.14%	35.95%
Total correct alignments (at chapter or book level)	*95.43%*	*51.97%*
Text-to-text: at least one sentence from same chapter or book	3.44%	4.07%
Errors: null-to-text or text-to-text from different chapter or book	0.83%	44.56%

D. LaBSE-Vecalign on Noisy Data: Additional Results

In Table 13, we report results of LaBSE-Vecalign on our noisy dataset using the number of predicted alignments as denominator.

Table 14 reports results of LaBSE-Vecalign on Lucretius after suppressing paratext only. We see continued errors in aligning English text sentences to null (55.21%).

In Table 15, we report results after suppressing only the second translation (by Good). In this experiment, we kept paratext in order to isolate the impact of multiple translations in the target text. These results show improved correct text-to-text alignments, with 96.26% of English text sentences aligned to Latin text sentences from the same book. However, the errors aligning English paratext sentences to Latin text sentences persist, with 36.26% of paratext sentences from the Watson translation aligned to Latin text sentences.

E. Word-Level Alignment of *Crito* with Fowler's and Jowett's translations

Table 16 shows the percentage of Greek and English words that are covered by word-level alignments. Of these aligned Greek words, Table 17 reports the percentage that cross the sentence boundaries predicted by LaBSE-Vecalign.

0059-003|220.|Cr. 52b.1.	οὐ γὰρ ἄν ποτε τῶν ἄλλων Ἀθηναίων ἁπάντων διαφερόντως ἐν αὐτῇ ἐπεδήμεις εἰ μή σοι διαφερόντως ἤρεσκεν, καὶ οὔτ᾽ ἐπὶ θεωρίαν πώποτ᾽ ἐκ τῆς πόλεως ἐξῆλθες, ὅτι μὴ ἅπαξ εἰς Ἰσθμόν, οὔτε ἄλλοσε οὐδαμόσε, εἰ μή ποι στρατευσόμενος, οὔτε ἄλλην ἀποδημίαν ἐποιήσω πώποτε ὥσπερ οἱ ἄλλοι ἄνθρωποι, οὐδ᾽ ἐπιθυμία σε ἄλλης πόλεως οὐδὲ ἄλλων νόμων ἔλαβεν εἰδέναι, ἀλλὰ ἡμεῖς σοι ἱκανοὶ ἦμεν καὶ ἡ ἡμετέρα πόλις:	for you would never have stayed in it more than all other Athenians if you had not been better pleased with it than they; you never went out from the city to a festival, or anywhere else, except on military service, and you never made any other journey, as other people do, and you had no wish to know any other city or other laws, but you were contented with us and our city.
0059-003|221.|Cr. 52b.2.	οὕτω σφόδρα ἡμᾶς ᾑροῦ καὶ ὡμολόγεις καθ᾽ ἡμᾶς πολιτεύσεσθαι, τά τε ἄλλα καὶ παῖδας ἐν αὐτῇ ἐποιήσω, ὡς ἀρεσκούσης σοι τῆς πόλεως.	So strongly did you prefer us and agree to live in accordance with us; and besides, you begat children in the city, showing that it pleased you.
0059-003|222.|Cr. 52c.1.	ἔτι τοίνυν ἐν αὐτῇ τῇ δίκῃ ἐξῆν σοι φυγῆς τιμήσασθαι εἰ ἐβούλου, καὶ ὅπερ νῦν ἀκούσης τῆς πόλεως ἐπιχειρεῖς, τότε ἑκούσης ποιῆσαι.	And moreover even at your trial you might have offered exile as your penalty, if you wished, and might have done with the state's consent what you are now undertaking to do without it.
0059-003|223.|Cr. 52c.2.	σὺ δὲ τότε μὲν ἐκαλλωπίζου ὡς οὐκ ἀγανακτῶν εἰ δέοι τεθνάναι σε, ἀλλὰ ᾑροῦ, ὡς ἔφησθα, πρὸ τῆς φυγῆς θάνατον:	But you then put on airs and said you were not disturbed if you must die, and you preferred, as you said, death to exile.
0059-003|232.|Cr. 53a.1.	τίνι γὰρ ἂν πόλις ἀρέσκοι ἄνευ νόμων;	for who would be pleased with a city apart from its laws?
0059-003|233.|Cr. 53a.2.	νῦν δὲ δὴ οὐκ ἐμμενεῖς τοῖς ὡμολογημένοις;	And now will you not abide by your agreement?
0059-003|234.|Cr. 53a.3.	ἐὰν ἡμῖν γε πείθῃ, ὦ Σώκρατες:	You will if you take our advice, Socrates;
0059-003|235.|Cr. 53a.4.	καὶ οὐ καταγέλαστός γε ἔσῃ ἐκ τῆς πόλεως ἐξελθών. σκόπει γὰρ δή, ταῦτα παραβὰς καὶ ἐξαμαρτάνων τι τούτων τί ἀγαθὸν ἐργάσῃ σαυτὸν ἢ τοὺς ἐπιτηδείους τοὺς σαυτοῦ.	and you will not make yourself ridiculous by going away from the city. “For consider. By transgressing in this way and committing these errors, what good will you do to yourself or any of your friends?
0059-003|261.|Cr. 54b.4.	ἐὰν δὲ ἐξέλθῃς οὕτως αἰσχρῶς ἀνταδικήσας τε καὶ ἀντικακουργήσας, τὰς σαυτοῦ ὁμολογίας τε καὶ συνθήκας τὰς πρὸς ἡμᾶς παραβὰς καὶ κακὰ ἐργασάμενος τούτους οὓς ἥκιστα ἔδει, σαυτόν τε καὶ φίλους καὶ πατρίδα καὶ ἡμᾶς, ἡμεῖς τέ σοι χαλεπανοῦμεν ζῶντι, καὶ ἐκεῖ οἱ ἡμέτεροι ἀδελφοὶ οἱ ἐν Ἅιδου νόμοι οὐκ εὐμενῶς σε ὑποδέξονται, εἰδότες ὅτι καὶ ἡμᾶς ἐπεχείρησας ἀπολέσαι τὸ σὸν μέρος.	but if you escape after so disgracefully requiting wrong with wrong and evil with evil, breaking your compacts and agreements with us, and injuring those whom you least ought to injure — yourself, your friends, your country and us — we shall be angry with you while you live, and there our brothers, the laws in Hades' realm, will not receive you graciously;
0059-003|262.|Cr. 54c.1.	ἀλλὰ μή σε πείσῃ Κρίτων ποιεῖν ἃ λέγει μᾶλλον ἢ ἡμεῖς.’ ταῦτα, ὦ φίλε ἑταῖρε Κρίτων, εὖ ἴσθι ὅτι ἐγὼ δοκῶ ἀκούειν, ὥσπερ οἱ κορυβαντιῶντες τῶν αὐλῶν δοκοῦσιν ἀκούειν, καὶ ἐν ἐμοὶ αὕτη ἡ ἠχὴ τούτων τῶν λόγων βομβεῖ καὶ ποιεῖ μὴ δύνασθαι τῶν ἄλλων ἀκούειν:	Be well assured, my dear friend, Crito, that this is what I seem to hear, as the frenzied dervishes of Cybele seem to hear the flutes, and this sound of these words re-echoes within me and prevents my hearing any other words.
0059-003|263.|Cr. 54c.2.	ἀλλὰ ἴσθι, ὅσα γε τὰ νῦν ἐμοὶ δοκοῦντα, ἐὰν λέγῃς παρὰ ταῦτα, μάτην ἐρεῖς.	And be assured that, so far as I now believe, if you argue against these words you will speak in vain.
0059-003|264.|Cr. 54c.3.	ὅμως μέντοι εἴ τι οἴει πλέον ποιήσειν, λέγε.	Nevertheless, if you think you can accomplish anything, speak.
0059-003|265.|Cr. 54c.4.	Κρίτων. ἀλλ᾽, ὦ Σώκρατες, οὐκ ἔχω λέγειν.	Crito. No, Socrates, I have nothing to say.

Figure 5: Errors aligning *Crito* to Jowett’s English translation

Table 14
LaBSE - Vecalign: Lucretius (en - la) - No Paratext, Both Translations (book-level evaluation)

Type of Alignment	Lucretius (book-level)
As a Percent of Predicted Alignments	
Number of Predicted Alignments	6876
Text-to-text: to source sentences from same book	34.42%
Text-to-text: to at least one source sentence from same book	0.00%
Errors, text-to-text: to no source sentence from same book	0.01%
Errors, text-to-null	65.56%
As a Percent of English Text Sentences	
Number of Text Sentences	8165
Text-to-text: to source sentences from same book	43.99%
Text-to-text: to at least one source sentence from same book	0.75%
Errors, text-to-text: to no source sentence from same book	0.05%
Errors, text-to-null	55.21%

Table 15
LaBSE - Vecalign: Lucretius (en - la) - With Paratext, First Translation Only (book-level evaluation)

Type of Alignment	Lucretius (book-level)
As a Percent of English Paratext Sentences From First Translation Only	
Number of Paratext Sentences	5913
Paratext sentences to null (correct)	63.74%
Paratext to Latin text (incorrect)	36.26%
As a Percent of English Text Sentences From First Translation Only	
Number of Text Sentences	3902
Text-to-text: to sources sentences from same book	96.26%
Text-to-text: to at least one source sentence from same book	1.08%
Errors, text-to-text: to no source sentence from same book	0.00%
Errors, text-to-null	2.67%

Table 16
Word Alignments of First 1,822 Words of the *Crito* with Fowler and Jowett Translations (through section 48a.4)

	Fowler (en)	Jowett (en)
Ancient Greek Words Covered by Alignments	80%	52%
English Words Covered by Alignments	83%	53%

Table 17
Percent of Aligned Greek Words that Cross Vecalign's Predicted Sentence Boundaries, First 1,822 Words of *Crito* (through section 48a.4)

Embedding Model Used	Fowler (en)	Jowett (en)
LASER	2%	20%
LaBSE	0%	9%