Classical Philology in the Time of AI: Exploring the Potential of Parallel Corpora in Ancient Languages

Tariq Yousef
University of Southern Denmark
yousef@imada.sdu.dk

Chiara Palladino
Furman University
chiara.palladino@furman.edu

Farnoosh Shamasian
Leipzig University
farnoosh.shamsian@uni-leipzig.de

Abstract

This contribution presents an overview of Parallel Text Processing, particularly Translation Alignment, and illustrates the current status of this task in ancient languages. In the first part, we provide the fundamental principles of Parallel Texts and give an overview of their applications for the study of ancient texts. In the second part, we indicate how Parallel Texts can be leveraged to perform other NLP tasks, including automatic alignment, dynamic lexica induction, and Named Entity Recognition. In the conclusion, we emphasize current limitations and future work.

1 Introduction

Parallel Text Processing refers to various computational tasks based on parallel corpora (Véronis, 2000). Parallel corpora are collections of texts that show some level of equivalence between them: for example, a text and its translations, or different versions of the same text.

The most important task in Parallel Text Processing is Text Alignment, that is, the automatic establishment of equivalences across various types of units: document, chunk, sentence, and word (Kay and Röscheisen, 1993). The task of aligning a text against its translation(s) is called Text-Translation Alignment (from now on, TA). The output of TA is defined as Translation Pairs (TPs), which correspond to pairs of the various units aligned (chunks, sentences, words, etc.).

TA can be considered a subfield of Text Alignment: however, it has very unique challenges attributed to the complex dynamics underlying the relationship between texts and their translations. In particular, word-level TA poses considerable complexity due to the inherent uncertainty in establishing individual equivalences: translations are not perfect transpositions of the originals, and tend to alter, normalize, expand or simplify parts of the text. Moreover, structural differences across languages, such as morphology and word order, contribute to additional difficulties.

The goal of this paper is to offer a programmatic survey of the current status of TA research in the specific domain of ancient languages, particularly Ancient Greek. As such, we will cover many different applications, both in Philology and Computer Science, with the intent of demonstrating the potential of this method in the study of ancient languages. Our aim is to illustrate how TA and parallel corpora can be used for a wide range of research, to contribute to existing debates and to inspire new questions.

2 Design and Concept of Translation Alignment Tools

Since TA was established, several tools have been designed to collect TPs, with or without integrated reading environments for visualizing the alignments (overviews are provided in our previous works Yousef and Janicke 2020; Yousef 2023). In the context of ancient languages, a limited number of tools have been developed. These include Alpheios (Almas and Beaulieu, 2013), DUCAT Citation Alignment Tool (Blackwell et al., 2020), Benner's tool for aligning the Bible (Benner, 2014), and UGARIT[1], designed to enable word-level alignments in low-resourced languages (Yousef et al., 2022c). Currently, UGARIT counts about 50 aligned languages, 700 users, and more than a million TPs[2], establishing itself as the most popular tool in this area.

UGARIT was designed as a crowd-sourcing project to collect training data for automatic alignment methods for ancient languages, but it expanded into a range of diverse applications, mostly thanks to its global community of scholars and

[1] https://ugarit.ialigner.com/

[2] Of this number, about 240,000 TPs are automatically generated through traditional statistical automatic alignment tool (Giza++).

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 179–192,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_021

students. The alignment workflow, which allows bilingual and trilingual alignments, is simple and intuitive.UGARIT allows different types of TPs: word-to-word (1-1), word-to-phrase (1-N), phrase-to-word (N-1) and phrase-to-phrase (N-N).

Alignments are immediately published online. The visualization of published alignments allows the user to compare aligned texts token by token, providing a transliteration service for non-Latin alphabets, statistical information about the percentage of aligned and not-aligned tokens, types of links, a downloadable list of TPs, and an embedding option (Figure A.1).

The tool integrates a dynamic lexicon, which can be triggered through the search function or simply by clicking on a word in an aligned text. The results are visualized as a radial cluster dendrogram, a tree view, and as a list of words with frequency (Figure A.2). The lexicon extracts all the translation equivalents of a given word across the whole database, providing a list of all languages in which that word has been translated.

3 Applications of Translation Alignment

In many modern languages, TA is successfully employed in a wide range of NLP tasks. For example, it is essential in word- and phrase-based Statistical Machine Translation (SMT) pipelines (Brown et al., 1993; Koehn et al., 2003); it can be used to analyze the output of Neural Machine Translation models (NMT) and assess their performance quality (Neubig et al., 2019); to filter and clean noisy parallel corpora (Kurfalı and Östling, 2019; Zariņa et al., 2015); to transfer linguistic annotation from one text to its translation, such as Semantic Role labels, POS tags, Named Entity tags (Yousef, 2015; Ni et al., 2017; Huck et al., 2019). Parallel Corpora aligned at word-level can support the work of professional translators (Liu, 2020), bilingual lexicon induction (Marchisio et al., 2021), and word sense disambiguation (Procopio et al., 2021). Moreover, they provide extremely useful information for vocabulary assimilation and language teaching (Vyatkina and Boulton, 2017), and for the study of the history of transmission of a corpus (Laviosa, 2021).

In the following sections, we will survey the current state of TA research for ancient languages, illustrating how the parallel corpora created on UGARIT are used for qualitative and quantitative research.

3.1 Qualitative Studies: Pedagogy and Translation Studies

Manual or supervised TA is essential for the creation of high-quality Gold Standards and training datasets. However, it can also be configured as a close reading task for translation study and language learning. In recent years, efforts have been undertaken in the realm of Digital Philology, in the context of a major emphasis on the development of open resources for innovative approaches to learning Classical languages (Crane et al., 2023).

3.1.1 Pedagogy and User Behavior

Parallel corpora on UGARIT are currently being used to teach Ancient Greek, Latin, and Persian in several universities, including Leipzig, Furman, Sâo Paulo, Tufts, University of Zagreb, Göttingen, Cattolica University, but also in schools across Europe, such as the Liceo G. Peano Tortona in Italy.

The active engagement with the text through the effort of establishing fine-grained equivalences stimulates a reflective approach to the text and creates an opportunity to design exercises that invite language learners to reflect upon the cultural and linguistic specificities of ancient texts through the contrastive comparison with modern translations: through specific exercises tailored to the level, students are stimulated to reflect on the depths of semantic and linguistic differences, and their impact on the very operation of translating (Palladino, 2020). Moreover, this process encourages a critical approach to translations as interpretations, rethinking their role in understanding ancient texts, and enabling the students to be part of a broader conversation about the reception and significance of a text over time. Palladino et al. 2021 provide a series of use cases showing how TA can be used for learning Ancient Greek or Latin at various levels, through a series of reflective and project-based exercises. Most importantly, the comparison of different translations of the same texts provides a tangible sense of the different strategies employed by professional translators, and gives a strong pragmatic understanding of the fluidity of translations and their (in)ability to transmit the original in its full meaning. Shamsian and Crane 2022 showed how TA can be integrated with grammar explanations and other types of annotations to create born-digital pipelines for learning ancient languages, even at beginner level. Through TA, students are able to critique existing scholarly translations and

reflect on how to create more accurate representations of the original. This process is particularly useful in linguistic contexts where available translations are mostly derivative from translations in other languages, like in the case of Persian.

3.1.2 Empirical study of translations and intertextual phenomena

While translations constitute a crucial aspect in the history of the transmission of ancient texts, very few studies have used computational approaches to investigate them. In this area, manual and automatic TA provides an extremely promising resource. Bizzoni et al. 2017 used an automatic alignment workflow based on the Needleman-Wunsch algorithm, using proper names as anchors to align selected passages of the *Odyssey* against a large corpus of French translations, to identify large-scale trends in translation practices across the 16th and 17th century. Shukhoshvili 2017 used UGARIT to support the creation of a complete translation of Plato's *Theaetetus* into Georgian and used the resulting corpus to investigate cross-linguistic dynamics between the two languages. Somewhat in the opposite direction, Xie 2023 used UGARIT to examine the Ancient Greek translation of the Latin text of the *Res Gestae*: the method applied combined close reading to inspect specific semantic phenomena, and distant reading through the consultation of the alignment statistics provided by the tool. Interestingly, while Xie 2023 found a remarkable degree of accuracy in the corpus, the trilingual alignment of the Rosetta Stone performed by Amin et al. 2023 on UGARIT demonstrated that the three versions of the text bear considerable differences and they cannot be considered one and the same text. Finally, Palladino et al. 2022a propose a workflow that combines close reading and quantitative indicators to support alignment-based evaluation of translations of Ancient Greek texts: the set of criteria includes frequency of link types, percentage of aligned and not-aligned words, intersection across translators, POS intersection, in combination with close reading of selected passages.

The ever-increasing amount of corpora in UGARIT also allows for big-data exploration scenarios. Palladino and Yousef 2023 used the UGARIT database to investigate cross-linguistic dynamics, studying how language and culture affect the establishment of word equivalents between text and translation. Their data show how different language systems influence the process of translation, creating very distinctive results for specific language pairs, but also that cultural context, text genre and modalities of transmission have an impact in determining structural differences in translations.

3.2 Quantitative Studies: AI and Parallel Corpora

The various applications described above show the importance of parallel corpora for the study of texts from different perspectives. For this reason, it is important to develop workflows for automated alignment tasks, which support the scalability of both qualitative and quantitative research. While this area is very well developed for modern languages, it is still in its infancy for ancient ones. In the following section, we will show current efforts in the improvement of automatic alignment models, and indicate how automatic TA can be used to enhance the performance of important NLP tasks.

Until the advent of transformer-based models, the state of the art of automatic TA was statistical methods, such as Giza++ (Och and Ney, 2003), fast_align (Dyer et al., 2013) and EfLomAl (Östling and Tiedemann, 2016). However, the performance of statistical alignment models relies on the presence and size of training datasets in the form of parallel sentences.

Recently, however, Neural Machine Translation (NMT) and multilingual transformer models have introduced the possibility of creating accurate alignments even with no training datasets (Jalili Sabet et al., 2020). Most notably, transformer models facilitate the creation of contextualized word embeddings, which encode information about a meaning of a word based on its context. Pre-trained multilingual transformer models, such as Multilingual Bert (mBERT) and XLM-RoBERTa, achieved significant performance improvements for numerous cross-lingual tasks (Conneau et al., 2019b; Devlin et al., 2018a).

Language models are now increasingly used for various NLP tasks in ancient languages (Sommerschield et al. 2023 provide a comprehensive survey in the field). Most current applications are developed with a strong interest in POS tagging and morphological analysis. To the best of our knowledge, we are pioneers in employing transformer models to automate TA tasks in ancient corpora, and to leverage on the resulting parallel texts to explore new possibilities in other NLP tasks. We

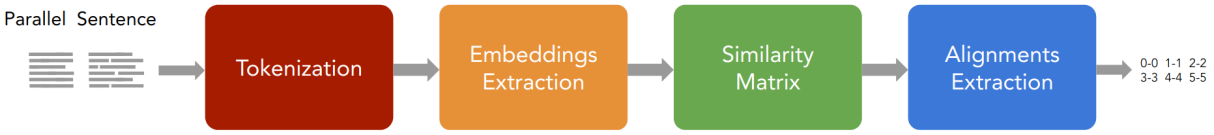

Figure 1: The alignment workflow.

use Ancient Greek as a case study, but the model we developed is multilingual and can be fine-tuned for other ancient languages.

3.2.1 Alignment Guidelines and Gold Standards

In order to evaluate the performance of automatic alignment models, it is essential to have high-quality gold standard datasets. Gold Standards are typically created by two or more annotators, whose Inter-Annotator Agreement (IAA) is measured to ensure consistency in the dataset. Guidelines are created and used to ensure that annotators are following a similar strategy.

While there is no lack of guidelines and standards for modern languages,[3] we developed the first ones specifically aimed at ancient languages: using Ancient Greek as case study, we considered translations into English, Portuguese, Persian,[4] and scholarly Latin. These guidelines can be used for the evaluation of automatic alignment tasks, but also as a general reference for students and scholars who wish to create their own parallel corpus for other purposes (Ferreira et al., 2022; Palladino et al., 2022b; Palladino and Shamsian, 2022).

The resulting Gold Standards are based on a corpus of 5,500 words from Ancient Greek epic poetry and prose (Homer, Xenophon, and Plato) and on 100 fragments of Ancient Greek translated into Latin from the Digital Fragmenta Historicorum Graecorum (DFHG)[5]. Two annotators aligned each corpus separately, after having drafted the Guidelines. The resulting IAA was measured at 86.17% for GRC-ENG and 83.31% for GRC-POR, and GRC-LAT 90.50%.

Our guidelines considered the same general principles established for modern languages (Lambert et al., 2005), but working within the specificities of an ancient language: for example, we had to carefully address the impact of high inflection and the inconsistency shown in the translation of linguistic and rhetorical structures. As a result, while most guidelines cover 7-10 classes of phenomena, ours covered 14 main classes with several subclasses. Therefore, it is easy to understand how the alignment of an ancient text may result in higher ambiguities than modern corpora traditionally used in TA: moreover, modern corpora are usually technical texts, which leave little space for variation, but that is not the case for ancient texts, which are necessarily literary or even poetical. Although our guidelines reach and exceed the 80% threshold of optimal consistency, it is important to reflect on the origins of disagreements across annotators in order to individuate areas of improvement for both the Gold Standards and automatic TA models: factors such as the native language of the annotators, their proficiency with the language/s, their familiarity with the text and specific dialect, and the time at their disposal may all have an impact on their performance. This qualitative study is part of our future work.

3.2.2 The UGARIT Ancient Greek Alignment Model

In our previous works (Yousef et al., 2022b,d), we have trained an automatic TA model that employs the recent advances in language modelling and is able to generate accurate word-level alignments even with small amounts of training data. In this context, we adapted the pipeline illustrated in Figure 1 proposed by (Jalili Sabet et al., 2020; Dou and Neubig, 2021).

The core concept is to leverage pre-trained multilingual contextualized language models such as MBERT (Devlin et al., 2018b) and XLM-ROBERTA (Conneau et al., 2019a) or fine-tuned versions of them. A similarity matrix can be derived based on distance/similarity metrics that calculate the similarity for every two tokens based on their embeddings. Then, the word-level alignments can be predicted by employing an extraction algorithm over the similarity matrix.

The initial experiments we conducted on the pre-

[3] An overview of available resources can be found on the UGARIT website: https://ugarit.ialigner.com/guidelines.php.

[4] This set of guidelines has not yet been used for the creation of Gold Standards, therefore we did not employ it for evaluation purposes.

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

Experiment	Languages	Data Size	Source
Phase 1	GRC Monolingual	12 Millions Tokens	Perseus DL, TreeBanking, First1kGreek
Phase 2	GRC-ENG, GRC-LAT GRC-KAT	45.000 sentences	Perseus DL, DFHG, UGARIT
Phase 3	Mixed dataset	5000 sentences 190k TPs	UGARIT

Table 1: The proposed fine-tuning strategy.

		mBERT				XLM-RoBERTa			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
ENG	Softmax	80.80%	56.91%	66.78%	32.72%	92.62%	66.85%	77.65%	21.88%
	Match	65.42%	72.76%	68.90%	31.31%	79.22%	**87.26%**	**83.05%**	**17.17%**
	Argmax	84.95%	52.47%	64.87%	34.57%	**94.44%**	63.32%	75.81%	23.70%
	Itermax	78.43%	64.08%	70.53%	29.14%	91.05%	71.65%	80.19%	19.42%
LAT	Softmax	85.67%	84.64%	85.15%	14.83%	94.64%	92.39%	**93.50%**	**6.47%**
	Match	62.18%	87.97%	72.86%	27.55%	80.61%	**96.30%**	87.76%	12.50%
	Argmax	88.46%	80.80%	84.46%	15.09%	**95.52%**	91.38%	93.40%	6.55%
	Itermax	81.27%	84.78%	82.99%	17.06%	92.21%	93.33%	92.77%	7.25%
POR	Softmax	63.84%	61.27%	62.53%	37.40%	76.11%	75.61%	75.86%	24.13%
	Match	50.00%	72.61%	59.22%	41.50%	58.79%	**86.17%**	69.89%	31.01%
	Argmax	66.01%	54.92%	59.96%	39.76%	**77.25%**	71.10%	74.05%	25.81%
	Itermax	59.67%	64.06%	61.79%	38.35%	72.22%	81.02%	**76.37%**	**23.91%**

Table 2: Evaluation results of the automatic alignment model on three gold standard datasets.

trained MBERT and XLM-ROBERTA (Zero-Shot) showed significantly poor performance on Ancient Greek-English, Ancient Greek- Latin, and Ancient Greek-Portuguese datasets. Therefore, fine-tuning those models was necessary to achieve better performance. Due to the availability of parallel sentences and in order to obtain the best outcome from the training process, we conducted several experiments employing multiple training objectives (Dou and Neubig, 2021) aiming to find the best training strategy. Each experiment tested various combinations of unsupervised and supervised training. Table 1 illustrates our proposed training strategy which consists of three phases. The initial stage involved training pre-existing models using monolingual Ancient Greek corpora, which encompassed a total of 12 million tokens. Subsequently, the model underwent an unsupervised fine-tuning process utilizing a collection of 45,000 parallel sentences. This fine-tuning phase encompassed sentences in Greek-English, Greek-Latin, and Greek-Georgian. Ultimately, the model underwent supervised fine-tuning, where it was refined using precise manual alignments extracted from the UGARIT database.

The performance of the model was evaluated against the gold standard datasets using *Precision*, *Recall*, *F*1 and Alignment Error Rate *AER*.

Table 2 presents the performance evaluation of our model during phase 3, utilizing three gold standard datasets: Greek-English, Greek-Latin, and Greek-Portuguese. We evaluated the model's performance using four alignment extraction heuristics and two fine-tuned models: mBert-based model and XLM-RoBERTa-based model. Notably, the fine-tuned XLM-RoBERTa models consistently outperformed the mBERT-fine-tuned models across all cases, demonstrating their superior performance in alignment extraction. The *Match* heuristic significantly outperformed other models regarding Recall. However, it achieved always the lowest Precision. On the other hand, the Argmax heuristic consistently achieved the highest precision but the lowest recall. Both the Softmax and Itermax heuristics demonstrated balanced performance, with a relatively equal consideration given to recall and precision. Itermax showcased superior recall compared to Softmax, while Softmax displayed better precision than Itermax. Overall, the performance of these heuristics varies in terms of recall and precision, with each exhibiting strengths and weak-

nesses. The choice of the appropriate heuristic will depend on the specific requirements and priorities of the task at hand, balancing the trade-off between recall and precision based on the desired outcomes.

Our alignment model is available on HUGGING FACE [6] and can be downloaded and used locally. In order to make it more accessible, however, we implemented an online tool[7] that integrates the pre-trained model and allows users to simply paste their texts and align them automatically, with an option to visualize and download the results (Figure A.1).

The pre-trained alignment model can be used to scale all the qualitative operations described above, but also for a variety of downstream tasks. In the following sections, we will describe our preliminary results in the areas of Bilingual Lexica Induction and Named Entity Recognition.

3.2.3 Bilingual Lexica Induction

The significance of aligned word-level parallel corpora as a data source for terminology banks and bilingual dictionaries is emphasized by Véronis 2000. These resources are highly valuable to improve the performance of professional translators, to enrich and train translation memory software, to retrieve terminology lists for technical texts, or in lexicographic studies. However, it is worth noting that not all language pairs can be easily aligned, especially when dealing with ancient and low-resourced languages. In this proof of concept, we applied automatic dictionary induction to produce high-quality translation pairs for languages that do not share parallel texts. Additionally, we represented the acquired translation pairs within a graph-based data structure. This approach allows us to integrate manual alignments and dictionary entries and facilitate performing clustering or pivoting to generate translation pairs of languages with no direct connections.

Corpora: we used 400,000 parallel sentences in 6 languages (Ancient Greek, Arabic, English, Hebrew, Latin, and Persian). Our corpus derives from the Bible[8], the Perseus Digital Library[9], and the DFHG corpus.

Alignment: we used our fine-tuned alignment model for Ancient Greek to perform the word/phrase alignments. We employed *Itermax* heuristic to extract the most accurate translation pairs from the similarity matrix since it achieved the highest Phrase Alignment Accuracy (Yousef et al., 2023; Yousef, 2023).

Graph Generation: Figure 2 illustrates the proposed graph structure. We model every translation pair as two nodes connected with an edge. Additional relations can be added to indicate different linguistic features if they are available. For example, connecting a phrase with its constituent words or linking a word with its lemma. These relations can be beneficial for running sophisticated queries. Shi et al. (2021) proposed a matching ratio that considers the alignment frequency and how frequently the two words co-occurred in the corpus. However, this ratio works only with one-to-one alignments. Therefore we proposed an alignment score that considers phrases as well:

$$score(s,t) = \frac{2 * A(s,t)}{A(s|L_t) + A(t|L_s)} \quad (1)$$

Where $A(s,t)$ indicates how many times the two words/phrases are aligned together, $A(s|L_t)$ indicates how many times s is aligned in total to words/phrases in the same language as t, and $A(t|L_s)$ indicates how many times t is aligned in total to words/phrases in the same language as s.

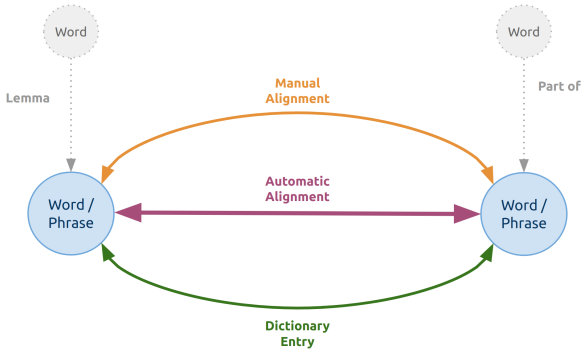

Figure 2: The graph structure of the induced TPs.

The resulting graph contains over 614k nodes and 1,620k edges from Automatic Alignment, and an additional 193k TPs collected from UGARIT as Manual Alignment. Moreover, graph clustering algorithms such as CHINESE WHISPER (Biemann, 2006), a hard partitioning and flat clustering algorithm, can be applied to cluster graph entries into sets containing words/phrases that are semantically related or share the same meaning. Figure 3 shows a cluster of aligned words/phrases in various languages. This cluster is one of 7300 clusters

[6] https://huggingface.co/UGARIT/grc-alignment
[7] http://ugarit-aligner.com
[8] https://github.com/christos-c/bible-corpus.
[9] http://www.perseus.tufts.edu/hopper/.

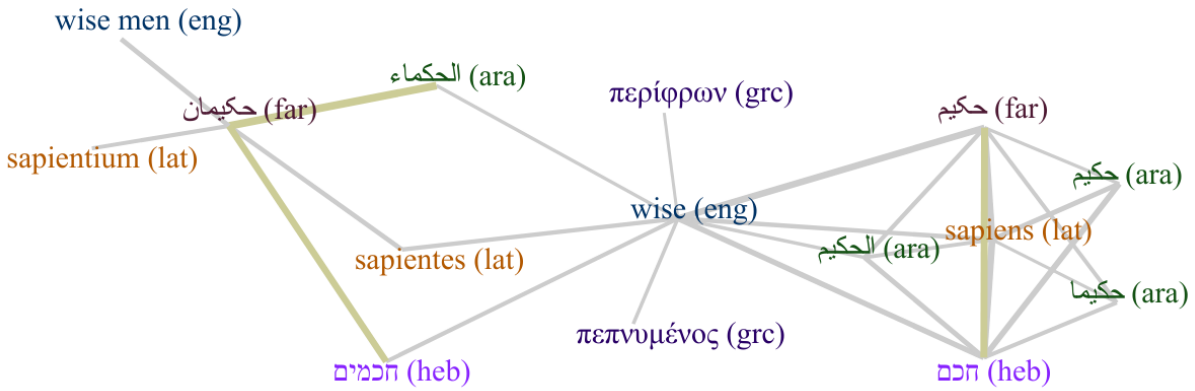

Figure 3: An example from graph clustering results.

obtained after filtering out the relations with frequency less than 5, alignment score less than 0.25, and running CHINESE WHISPER clustering algorithm for 20 iterations. Figure A.4 shows another cluster with an extended visualization, in which manual alignments and PART_OF relations are displayed.

The results of this work are available on our GitHub[10]: The resulting dictionaries can provide an invaluable resource to establish equivalences across languages that are not normally translated into each other: for example, a Persian speaker studying Ancient Greek can use this resource to extract Persian equivalents of Greek words, instead of relying on English or French translations. Moreover, the dictionaries provide insights into real-world use of the words, as they derive directly from contextual usages in texts. Therefore, the development of this application will considerably improve the use of parallel corpora for teaching, translating, and language learning.

Our future work in this regard includes expanding the corpus of accurate manual alignments for other low-resourced languages, and expanding the monolingual and bilingual datasets to improve the accuracy of the model in other languages. We will also develop a user interface with various search and visualization functions.

3.2.4 NER for Ancient Greek

An additional application of our alignment model pertains to enhancing the efficacy of Named Entity Recognition (NER) in the context of ancient languages through the employment of annotation projection. This workflow leverages on cross-lingual transfer: the basic principle is that, if NER models reach accurate results in one language, we can use an automatic alignment workflow to align an annotated text with another one, for which NER models do not achieve such a high performance. This principle, called annotation projection, consists in projecting NER annotations performed on English translations on an aligned text in an ancient language, so that Named Entities can be extracted and classified through the alignment process.

NER is in great demand among scholars of ancient languages. However, it comes with significant challenges including OCR-generated errors and noisy data, complexity of the sources, lack of gold standards and guidelines. The only survey on the topic for historical languages is provided by Ehrmann et al. 2021, with some recent updates in Sommerschield et al. 2023. New pipelines based on transformers have shown considerable improvement in this area, although NER remains a particularly challenging task (Palladino et al., 2020; Yousef et al., 2022a; Burns, 2023; Yoo et al., 2022).

While most of these experiments use a direct training approach with annotated datasets of Named Entities in the target language, we propose a novel workflow that integrates annotation projection and leverages on our automatic alignment model. Figure 4 illustrates our pipeline: we collect a parallel corpus of Ancient Greek and English translations; automatically annotate the text of the English translations using *AllenNLP*, an accurate off-the-shelf NER system[11]; then, we employ automatic word alignment to retrieve translation pairs,

[10] https://github.com/UgaritAlignment/

[11] We benchmarked three high-quality English NER models, namely, *spaCy*, *AllenNLP* and *flairNLP* to select the model with the highest accuracy on our corpus. The comparison revealed that *AllenNLP* and *flairNLP* significantly outperformed *spaCy*, and their performance was very close.

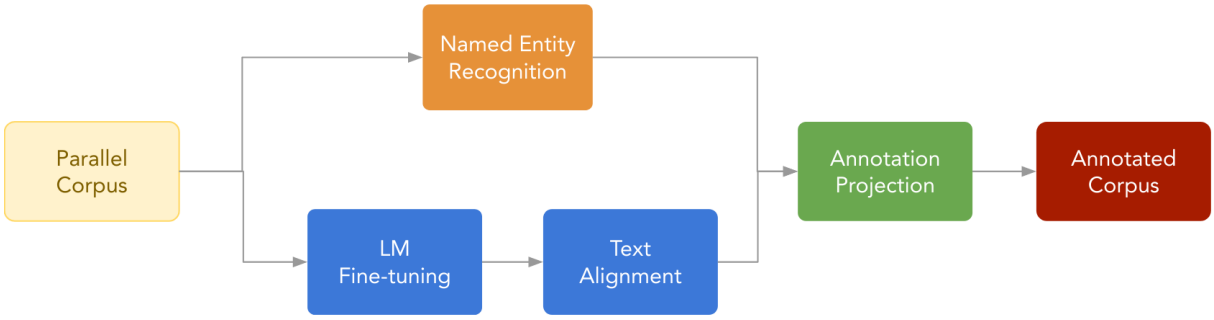

Figure 4: Named-Entity annotation projection pipeline.

and project the annotations from the English translations onto the corresponding tokens in the Ancient Greek text using a direct mapping heuristic.

While *AllenNLP* provides four entity classes (PERS, LOC, ORG, MISC), we only used PERS, LOC, and MISC, as the ORG entity label does not apply intuitively to ancient naming systems (see further on this issue Ehrmann et al. 2021; an alternative strategy for labeling is proposed for Latin by Burns 2023.).

We tested the workflow on the Bible corpus, using English for annotation and selecting versions in Ancient Greek, Latin, and Arabic[12]. We decided to expand the range of languages beyond Ancient Greek, which is still the most present in training datasets, to show the potential of the multilingual model.

Two domain experts performed qualitative evaluation on 100 random verses (about 550 entities per dataset) and assigned a score as shown in table 3. The evaluation results show an accuracy of 86.63% in Ancient Greek, 82.34% in Latin, and 75.54% in Arabic: understandably, Arabic showed the worst performance because we had much less corpora available for training. The most common errors were found in the misclassification of entities, sometimes as a consequence of the fact that English translations adopted a different entity type. Most notably, many ethnonyms (MISC in our dataset) were translated with place-names in English, and therefore classified as LOC in the ancient language. Moreover, incomplete or partial alignments were frequent in multi-token entities, such as "Jesus Christ", "Simon Zelotes", and "Pontius Pilate".

[12]All versions were taken from the Bible Corpus on GitHub, while the Ancient Greek version was retrieved from the Perseus Digital Library.

4 Conclusions and Future Work

Parallel Corpora are today's Rosetta Stones (Véronis, 2000). They can be used for a variety of philological and computational tasks, as they provide a medium between languages and cultures. This study shows the importance of parallel text processing, specifically in the context of Translation Alignment, for various activities in the study of low-resource languages. The value of TA emerges in its various applications, which include language learning, NLP development, dictionary extraction, and research on translations and cross-linguistic interactions.

Most importantly, the development of accurate TA models can significantly contribute to improve the performance of important NLP tasks in ancient languages through the mdeium of annotation projection. For this reason, we plan a significant expansion of monolingual and bilingual corpora for supervised and unsupervised training, in order to improve performance on other ancient languages. Moreover, we will test analogous workflows based on annotation projection for other NLP tasks, such as POS tagging and lemmatization. In this sense, the development of accurate sentence alignment workflows is fundamental, as it can significantly enhance the performance of word-alignment models.

Despite the great success of transformers and language models, we want to emphasize that manually annotated corpora and guidelines are still essential to ensure accurate performance and to detect patterns of error. Gold Standards and output evaluation require strong disciplinary expertise, especially in scenarios where the research questions are complex. For this reason, as already emphasized by Sommerschield et al. 2023, the best efforts in the domain of automatic text processing are achieved

Score	Ancient Greek	Latin	Arabic
Correct alignment / Correct NER	86.63%	82.34%	75.54%
Incorrect alignment / Correct NER	7.26%	12.87%	21.16%
Correct alignment / Incorrect NER	5.28%	3.96%	2.98%
Incorrect alignment / Incorrect NER	0.83%	0.83%	0.33%

Table 3: Manual evaluation of 100 randomly selected verses.

by multidisciplinary teams, where the contribution of scholars of the language and philologists can provide better information about the idiosyncrasies of the material, and crucially contribute to the evaluation of the results. High-quality philological work is essential for progress in this field, and the only way we can produce reliable tools that will be used by Digital Humanists and Humanists as well.

Acknowledgements

We are most grateful to all the people who contributed to the development of UGARIT, the Alignment Models, and who provided Gold Standards and annotated datasets: Gregory Crane, Monica Berti, Anise d'Orange Ferreira, Michel Ferreira dos Reis, Josh Kemp, David Wright, Maia Shukhoshvili, Sisi Xie, Brian Clark, and all the scholars and students who worked on Translation Alignment on Ugarit.

References

Bridget Almas and Marie-Claire Beaulieu. 2013. Developing a New Integrated Editing Platform for Source Documents in Classics. *Literary and Linguistic Computing*, 28(4):493–503.

Miriam Amin, Angelos Barmpoutis, Monica Berti, Eleni Bozia, Josephine Hensel, and Franziska Naether. 2023. The Digital Rosetta Stone Project. In Rita Lucarelli, Joshua A. Robertson, and Steve Vinson, editors, *Ancient Egypt, New Technology.*, volume 17 of *Harvard Egyptological Studies*, pages 58–84. Brill, Leiden - Boston.

Drayton Benner. 2014. A Tool for a High-Carat Gold-Standard Word Alignment. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 80–85.

Chris Biemann. 2006. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City. Association for Computational Linguistics.

Yuri Bizzoni, Marianne Reboul, and Angelo Del Grosso. 2017. Diachronic Trends in Homeric Translations. *Digital Humanities Quarterly*, (2).

Christopher W. Blackwell, Chiara Palladino, Mackense Greico, and Allie Bolton. 2020. DUCAT: Passage/Translation Alignment with the CITE Architecture. In *Proceedings of the DH2020 Digital Humanities Conference 2020, Ottawa*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311. Place: Cambridge, MA Publisher: MIT Press.

Patrick J. Burns. 2023. LatinCy: Synthetic Trained Pipelines for Latin NLP. ArXiv:2305.04365 [cs].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116. ArXiv: 1911.02116.

Gregory Crane, Alison Babeu, Lisa M. Cerrato, Amelia Parrish, Carolina Penagos, Farnoosh Shamsian, James Tauber, and Jake Wagner. 2023. Beyond translation: engaging with foreign languages in a digital library. *International Journal on Digital Libraries*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805. : 1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zi-Yi Dou and Graham Neubig. 2021. 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*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named Entity Recognition and Classification on Historical Documents: A Survey. ArXiv:2109.11406 [cs].

Anise d'Orange Ferreira, Michel Ferreira dos Reis, and Tariq Yousef. 2022. Critérios ou Convenções de Alinhamento do Grego às Traduções em Português. Publisher: Zenodo. Version Number: 1.0.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual Annotation Projection Is Effective for Neural Part-of-Speech Tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 223–233, Ann Arbor, Michigan. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Martin Kay and Martin Röscheisen. 1993. Text-translation Alignment. *Computational Linguistics*, 19(1):121–142.

Philipp Koehn, Franz J Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, University of Southern California Marina Del Rey Information Sciences Inst.

Murathan Kurfalı and Robert Östling. 2019. Noisy parallel corpus filtering through projected word embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 277–281.

Patrik Lambert, Adrià De Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation*, 39(4):267–285. Publisher: Springer.

Sara Laviosa. 2021. *Corpus-based Translation Studies: Theory, Findings, Applications*, volume 17 of *Approaches to Translation Studies*. Brill, Leiden - Boston.

Kanglong Liu. 2020. *Corpus-Assisted Translation Teaching: Issues and Challenges*, volume 7 of *Corpora and Intercultural Studies*. Springer.

Kelly Marchisio, Philipp Koehn, and Conghao Xiong. 2021. An Alignment-Based Approach to Semi-Supervised Bilingual Lexicon Induction with Small Parallel Corpora. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 293–304, Virtual. Association for Machine Translation in the Americas.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A Tool for Holistic Comparison of Language Generation Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Chiara Palladino. 2020. Reading Texts in Digital Environments: Applications of Translation Alignment for Classical Language Learning. *The Journal of Interactive Technology and Pedagogy*, (18).

Chiara Palladino, Maryam Foradi, and Tariq Yousef. 2021. Translation Alignment for Historical Language Learning: a Case Study. *Digital Humanities Quarterly*, 015(3).

Chiara Palladino, Farimah Karimi, and Brigitte Mathiak. 2020. NER on Ancient Greek with minimal annotation. *DH2020 Ottawa. Book of Abstracts*.

Chiara Palladino and Farnoosh Shamsian. 2022. Translation Alignment: Ancient Greek to English. Annotation Style Guide and Gold Standard. Publisher: Zenodo Version Number: 1.0.

Chiara Palladino, Farnoosh Shamsian, and Tariq Yousef. 2022a. Using Parallel Corpora to Evaluate Translations of Ancient Greek Literary Texts. An Application of Text Alignment for Digital Philology Research. *Journal of Computational Literary Studies*, (1).

Chiara Palladino, David J. Wright, and Tariq Yousef. 2022b. Translation Alignment: Ancient Greek to Latin. Annotation Style Guide and Gold Standard. Publisher: Zenodo Version Number: 1.0.

Chiara Palladino and Tariq Yousef. 2023. To say almost the same thing? A study on cross-linguistic variation in ancient texts and their translations. *Digital Scholarship in the Humanities*.

Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. 2021. MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. volume 4, pages 3915–3921. ISSN: 1045-0823.

Farnoosh Shamsian and Gregory R. Crane. 2022. Open Resources for Corpus-Based Learning of Ancient Greek in Persian. *Journal of Interactive Technology and Pedagogy*, (21).

Haoyue Shi, Luke Zettlemoyer, and Sida I Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. *arXiv preprint arXiv:2101.00148*.

Maia Shukhoshvili. 2017. Methodology of Translation Alignment of Georgian Text of Plato's "Theaetetus". *International Journal of Language and Linguistics*, 4(4).

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*.

Nina Vyatkina and Alex Boulton. 2017. Corpora in Language Teaching and Learning. *Language Learning and Technology*, (3).

Jean Véronis, editor. 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Text, Speech and Language Technology. Springer Netherlands, Dordrecht-Boston-London.

Sisi Xie. 2023. Textual Alignment of Res Gestae: Translation in Historical Languages. *The Stoa: a Review for Digital Classics*.

Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. HUE: Pretrained Model and Dataset for Understanding Hanja Documents of Ancient Korea. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1832–1844, Seattle, United States. Association for Computational Linguistics.

Tariq Yousef. 2015. Word alignment and named-entity recognition applied to greek text reuse. *MSc's Thesis. Alexander von Humboldt Lehrstuhl für Digital Humanities, Universität Leipzig*.

Tariq Yousef. 2023. *Translation Alignment Applied to Historical Languages*. Ph.D. thesis.

Tariq Yousef, Gerhard Heyer, and Stefan Jänicke. 2023. Evalign: Visual evaluation of translation alignment models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 277–297.

Tariq Yousef and Stefan Janicke. 2020. A Survey of Text Alignment Visualization. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2022a. Transformer-based named entity recognition for ancient greek.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022b. 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.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, and Maryam Foradi. 2022c. Translation Alignment with Ugarit. *Information*, 13(2):65. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022d. Automatic Translation Alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Ieva Zariņa, Pēteris ikiforovs, and Raivis Skadiņš. 2015. Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 185–192.

Robert Östling and Jörg Tiedemann. 2016. Efficient Word Alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*, 106.

A Appendix

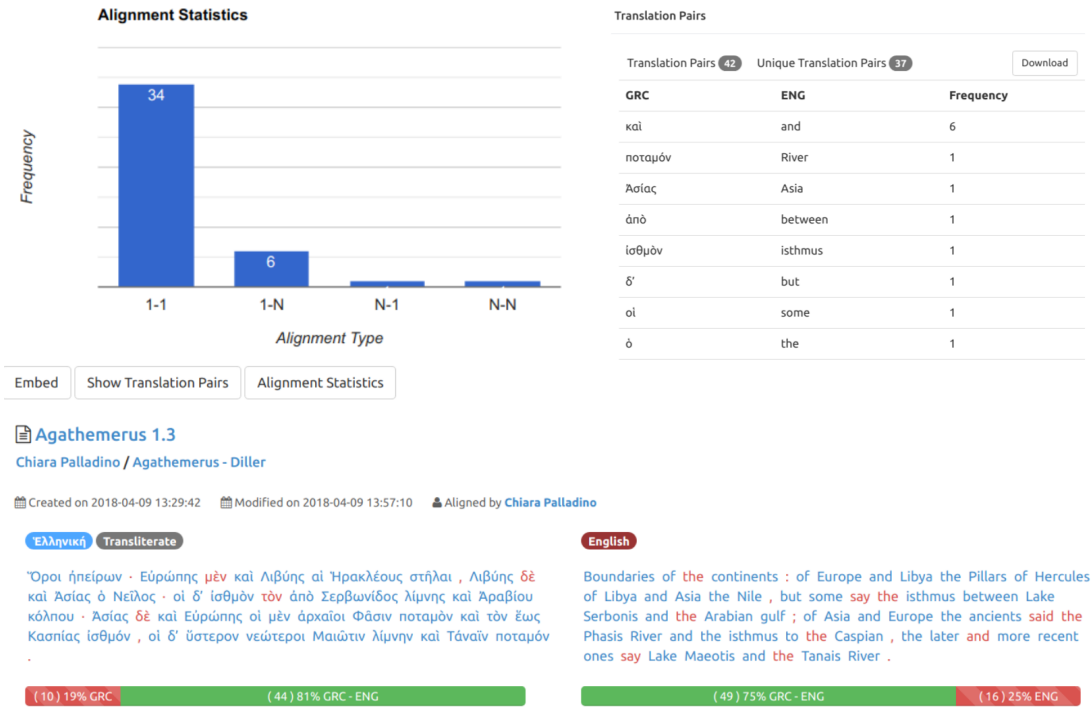

Figure A.1: Ugarit manual alignment tool, Side-by-side visualization of bilingual aligned texts.

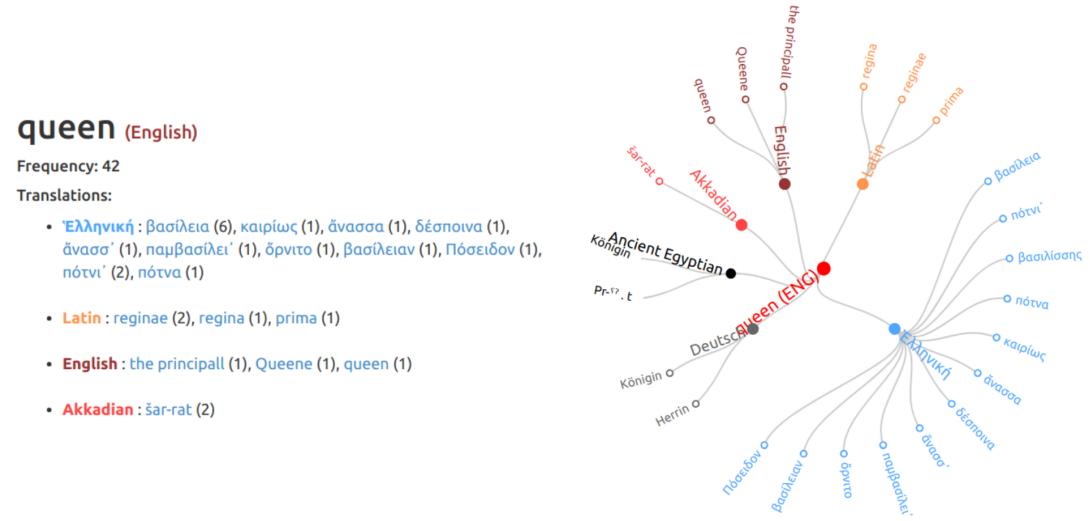

Figure A.2: Visualization of translation pairs search results.

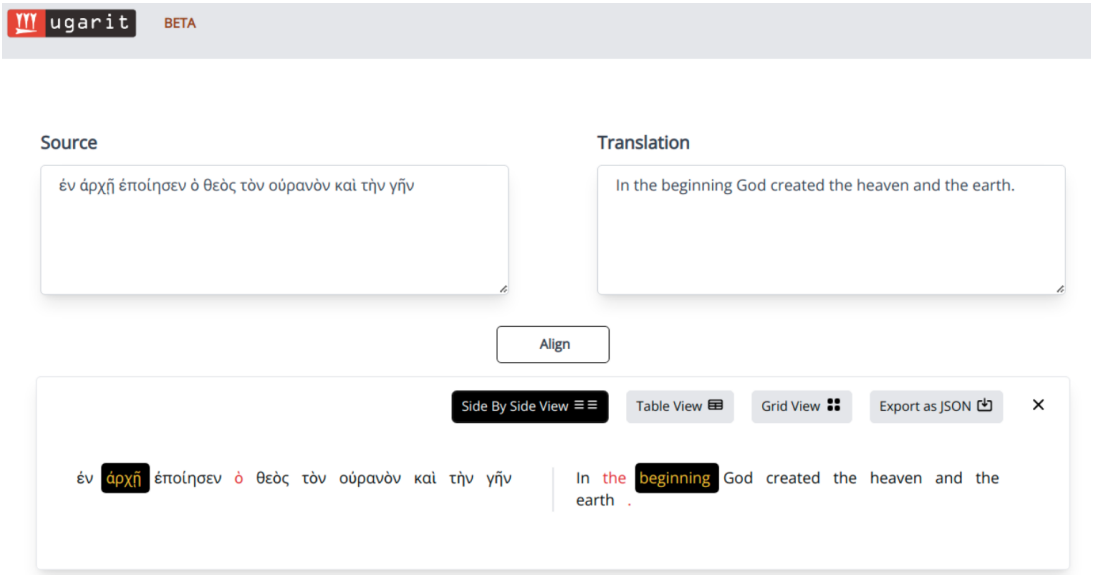

Figure A.3: Ugarit automatic alignment tool.

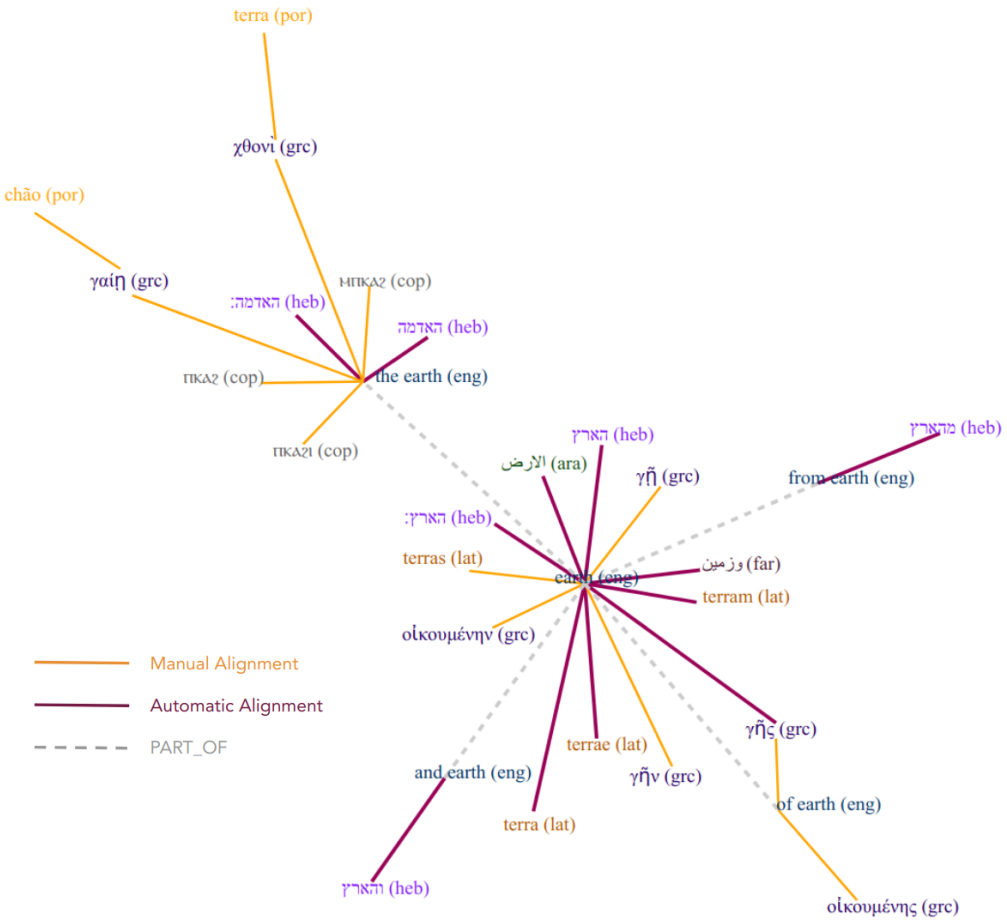

Figure A.4: An example from graph clustering results with extended relations.

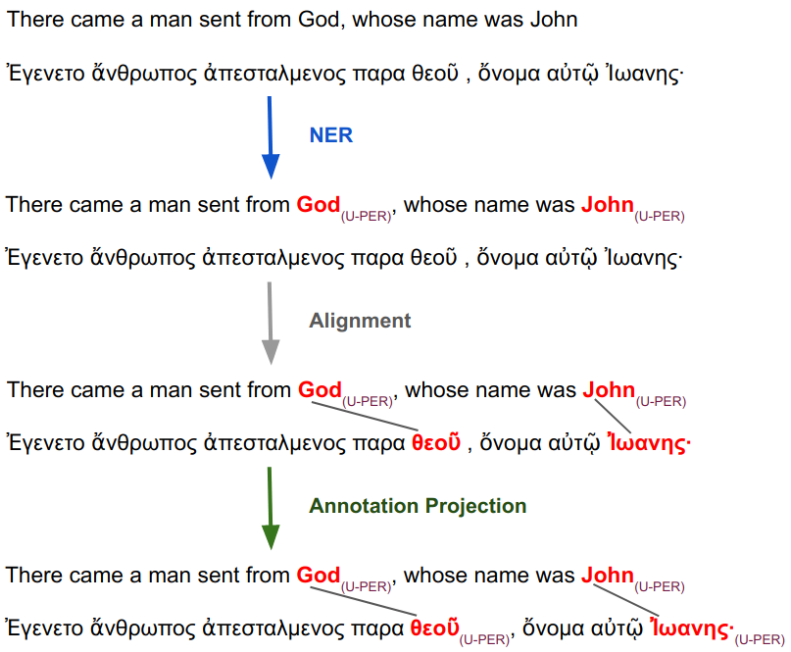

Figure A.5: An example of the annotation projection using the proposed pipeline.