Word Sense Disambiguation for Ancient Greek: Sourcing a training corpus through translation alignment

Alek Keersmaekers, Wouter Mercelis, Toon Van Hal
University of Leuven, Belgium
{alek.keersmaekers,wouter.mercelis,toon.vanhal}@kuleuven.be

Abstract

This paper seeks to leverage translations of Ancient Greek texts to enhance the performance of automatic word sense disambiguation (WSD). Satisfactory WSD in Ancient Greek is achievable, provided that the system can rely on annotated data. This study, acknowledging the challenges of manually assigning meanings to every Greek lemma, explores strategies to derive WSD data from parallel texts using sentence and word alignment. Our results suggest that, assuming the condition of high word frequency is met, this technique permits us to automatically produce a significant volume of annotated data, although there are still significant obstacles when trying to automate this process.

1 Aims

This contribution aims at making active use of translations of Ancient Greek texts in order to improve results in automatic word sense disambiguation (WSD). Section 2 outlines the general research context, showing that decent WSD in Ancient Greek is, in the current stage, feasible if the system can be trained on annotated data. Given the impracticality of manually annotating word meanings to all Greek lemmas, this paper explores the possibility of generating a significant volume of annotations automatically. Section 3 surveys related work both at the level of our general aim – word-sense disambiguation of Ancient Greek – and at the level of the methodology we adopt for attaining automatically annotated data for word-sense disambiguation, viz. sourcing from parallel texts via sentence and word alignment. After detailing the methodology adopted (Section 4), we subsequently discuss the results obtained, possible avenues for improvement and perspectives for applications (Sections 5-7).

2 Research context: towards onomasiological searches

It is generally known that in natural languages there is not a one-to-one mapping between form and meaning: one form or term can express various meanings or concepts (e.g. ‘bright’ can refer to light or intelligence) and vice versa (e.g. there are various ways to express that a person is intelligent, including ‘bright’, ‘clever’, ‘smart’ etc.). In semantic theory, studying the various meanings that a specific form expresses is called the ‘semasiological’ perspective, while studying the various forms that can be used to express a certain meaning is called the ‘onomasiological’ perspective (see Geeraerts, 2010).

This has important practical consequences: while it is straightforward to query most annotated corpora for specific terms, querying it for specific concepts is usually far less straightforward (see, for instance, Goossens, 2013). Most corpora have not been annotated semantically, given that the annotation is labor-intensive and often subjective, and semantics is multifaceted. However, to avoid manual annotation, one could make use of so-called ‘vector-based models of meaning’ or ‘word embeddings’, which retrieve computational representations of meaning in a bottom-up manner from a large, unannotated dataset (Lenci, 2018).

In the context of Ancient Greek, exploratory studies of vector-based models for detecting onomasiology have begun to emerge, starting from the premise that these models can be harnessed to identify words bearing a similar or related meaning to a given target word (Keersmaekers and Van Hal, 2021 & 2022). In this case, if the researcher already knows some terms that can express a particular concept (say γλῶσσα and φωνή for the concept ‘language’), they can use these models to look for terms that are similar to these target words and by

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 148–159,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_018

doing so fully map the onomasiology of this concept.

However, one complication is polysemy. When using a vector-based model[1] to find the ten nearest neighbors of the term γλῶσσα, for example, the results are all body parts, such as οὖς 'ear', ὀδούς 'tooth', ὀφθαλμός 'eye', χεῖλος 'lip' and φάρυγξ 'throat'. The explanation for this is predominantly linked to the polysemy of γλῶσσα, which can denote both 'language' and 'tongue'. The latter meaning is particularly prominent due to the corpus's extensive inclusion of medical data, which constitutes 14% of all training data, in which 'tongue' is more frequently referred to.

One possible solution is WSD: if we could separate all tokens of γλῶσσα that mean 'tongue' from those meaning 'language', we could look for the nearest neighbors of γλῶσσα when only the tokens meaning 'language' are taken into account. Again, vector-base models can be employed for this: indeed, several transformer-based embedding models such as BERT (Devlin et al., 2019) do no longer model the 'general' meaning of a word but the meaning of a word in context. Such an approach for Ancient Greek is discussed in Mercelis et al. (Forthc.), using ELECTRA (Clark et al., 2020) as a language model. When this model was used in an unsupervised way, the results were disappointing, possibly due to data sparsity. However, when used in a supervised way, by finetuning the transformer network, decent results could be achieved with only about 150 training examples (for binary meaning distinctions) or 300 (for ternary meaning distinctions).

To overcome the problems related to the acquisition bottleneck in obtaining annotated data (Lefever et al., 2011: 320; Pasini, 2020), this paper will discuss an automated way of creating datasets for WSD, by exploiting parallel texts (Greek original texts and English translations). This approach initially involves aligning sentences. Subsequently, within the aligned sentences, individual words are aligned. This two-step process will enable us to annotate polysemous words in Greek with English labels, thus trying to get a hold of their polysemy.

3 Related work

3.1 WSD for Ancient Greek

While the problem of automatic WSD has been tackled for decades already for English, interest in computational semantics has only raised recently for Ancient Greek, and the literature on this topic is therefore very limited. The only studies that we are aware of are Mercelis et al. (Forthc.), as discussed in Section 2, and McGillivray et al. (2019). While Mercelis et al. (Forthc.) directly explored supervised and unsupervised WSD using large language models, the angle of McGillivray et al. (2019) is somewhat different in that they explore how computational methods can be used for lexical semantic change detection. Focusing on three polysemous words (viz. μῦς, ἁρμονία and κόσμος), they explore their polysemy over time and genre using a Bayesian topic model, and match the results to manually annotated datasets of these words.

3.2 Word and sentence alignment

Word alignment used to be one of the key steps in the process of statistical machine translation. Statistical word alignment, represented by GIZA++ (Och and Ney, 2003) formed a strong baseline, which was only surpassed recently by large language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), based on transformer techniques. Nowadays, attention mechanisms in these large language models have made the word alignment task obsolete in machine translation pipelines. Nonetheless, in recent years word alignment made a comeback, albeit not solely in function of machine translation (Li, 2022). Our paper can be situated in this newfound interest in word alignment, as we focus on aligning words to create datasets for WSD.

Li (2022) provides a comprehensive summary of the history of word alignment, along with an overview of potential strategies for executing this task. Given that the word alignment task is inherently multilingual, most approaches employ a multilingual language model such as mBERT (Devlin et al., 2019) or XLM-RoBERTa (Conneau et al., 2020), which is then fine-tuned for the alignment task. In our case, this is more complex, given that Ancient Greek is in general not incorporated in such multilingual models. Hence,

[1] This example is retrieved from the vector models described in Keersmaekers & Van Hal 2021, which are based on word vectors created using singular value decomposition incorporating syntactic dependency features.

we used the recently released PhilBERTa model (Riemenschneider and Frank, 2023), a trilingual model trained on English, Ancient Greek, and Latin texts.

Yousef et al. (2022a) recently investigated translation alignment at the word level, with a particular focus on Ancient Greek. They utilized multilingual embeddings from which they selected the most similar pairs, signifying aligned words. They employed two alignment techniques: the approach of Jalili Sabet et al. (2016) and that of Dou and Neubig (2021). While according to Li (2022) both techniques handle the word alignment task proficiently, the highest-performing technique in Li's (2022) dataset was a span-extraction model by Nagata et al. (2020). This approach is widely recognized for its application in Question Answering, as the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) was designed with this technique in mind.

Chousa et al. (2020) released a similar model – also based on span-extraction – for sentence alignment. This model achieved state-of-the-art results on various modern language combinations (German – English, French – English, Japanese – English), beating previous approaches such as VecAlign (Thompson and Koehn, 2019).

3.3 Translation alignment for WSD

Parallel texts have a long-standing tradition in WSD, with its roots traced back to the work of Ng et al. (2003) (cf. Pasini, 2020: 4939 for more details). Our approach in this contribution is bilingual, viz. Ancient Greek – English. Over the past decade or so, there has been an emphasis on a multilingual rather than bilingual approach to parallel corpora for WSD (see e.g. Lefever et al., 2011). Most of these approaches rely on the massive European parliament corpus (see, e.g. Delli Bovi et al., 2017). Rather than concentrating solely on direct annotation transfer on the token level, certain researchers propose a more holistic approach. This involves taking into account the wider context provided by the entire parallel corpus, rather than merely focusing on parallel sentences (van der Plas and Apidianaki, 2014). More recently, scholars have proposed multilingual approaches in which translation parallels are replaced with propagation methods. Starting from contextualized word embeddings in English and relying on multilingual data from knowledge bases (such as WordNet and Wikipedia), such approaches can automatically generate training data for languages without labeled data for WSD (Barba et al., 2020). Recent research has also pointed out that the generation of translations can improve the quality of WSD (see e.g. Luan, 2020).

4 Methodology

4.1 Data

In our undertaking Ancient Greek is the source language and English the target language, given the abundance of English translations and manually aligned data. For our source language, we started from the GLAUx corpus (Keersmaekers, 2021), which encompasses approximately 32M Greek tokens, spanning roughly from the 8th century BC to the 4th century AD. As for the target language, the majority of our English data was drawn from the Perseus project (Smith et al., 2000).[2] However, we also incorporated openly accessible online editions for certain lengthy texts not available in Perseus, such as Dionysius of Halicarnassus's *Roman Antiquities*. Both the GLAUx data and the English translation data incorporate information about the texts' structure (e.g., division into books, chapters, sections, verses, etc.). This facilitated the alignment of 'paragraphs' in both languages. We use the term 'paragraph' loosely here, referring to the shortest shared structural unit between the Greek text and its translation, which can be, for instance, a section, chapter (if no sections are provided), or, in the case of poetic texts, a group of verses. In total, we were able to link around 7.2 million Greek tokens (approximately a quarter of the GLAUx corpus) to an English translation.

We trained word alignment models using data from the Alpheios project[3] and from the UGARIT project (Yousef et al., 2022b) as training data (66929 tokens). For the sentence alignment task, we used the same data sources, supplemented with Pedalion data (Keersmaekers et al., 2019), as well as a parallel New Testament corpus and data from the Greek Learner Texts Project.[4] In addition to this, we also annotated data ourselves. In total, this amounted to 15178 training sentences.

[2] Data taken from https://github.com/PerseusDL/canonical-greekLit.
[3] https://alpheios.net/pages/tools/
[4] See https://greek-learner-texts.org and https://github.com/jtauber/plato-texts.

During the development stage, we assessed the word alignment task using the same gold standard data (5076 tokens) employed by Yousef et al. (2022a). This facilitated a direct comparison of our results with their work. We evaluated the sentence alignment model using our own held-out data (879 sentences). This dataset was the most appropriate for evaluating performance as it consisted of parallel paragraphs, whereas for other datasets, we were forced to artificially combine sentences into existing or sometimes even entirely new paragraphs (fixed at a length of 10 sentences), since they did not provide paragraph data. Given the length of some of these paragraphs in our evaluation dataset, this dataset posed a significant challenge for the model in accurately predicting sentence alignments.

4.2 Sentence alignment

Our target corpus, GLAUx, is paragraph-aligned, requiring us to first conduct sentence alignment to enable word alignment within these sentences.

Segmenting Ancient Greek paragraphs into sentences is a straightforward process, given the existence of meticulous editions of the available texts and the general lack of abbreviations that might complicate splitting at full stops. Thus, our aim is to extract the English sentences that correspond to a particular Ancient Greek sentence from an entire English paragraph.

To achieve this, we employ a span-extraction approach, based on the work of Chousa et al. (2020), as discussed in Section 3.2. This method represents the state-of-the-art approach and is methodologically quite similar to the word alignment model. The key distinction lies in the focus of extraction: tokens from sentences in the case of word alignment, and sentences from paragraphs for sentence alignment.

4.3 Word alignment

As noted earlier, there are several strategies for word alignment. For this task, we selected the span-extraction approach as well. This method was the top performer in the study by Li (2022),[5] and it utilizes annotated data, to which we had access. Additionally, choosing a different approach to word alignment than Yousef et al. (2022a) allowed us to compare the outcomes.

For each pair of parallel sentences, we used the English sentence as the context. Then, for every token in the Ancient Greek sentence, we treated this sentence as a 'question', similar to the terminology used in SQuAD. In this sentence, the current token was demarcated with a special separation token. Both the context and the 'question' were processed by a PhilBerta model (Section 3.2), fine-tuned for the span-extraction task. The model then predicted the start and end indices of the corresponding English token in the context, or the English sentence, thereby aligning the Ancient Greek and English tokens.

Upon completing this process, we secured a corpus that was aligned at the word level.

Lemma's	**Frequency band**
γλῶσσα; λόγος; φωνή	1
ῥῆμα	2

Table 1: Linguistic terms.

Lemma	**Frequency band**
αἴσθησις; καταλύω; ἀλλότριος	1
βίοτος; ἀπαντάω; μιαρός	2
ἱστός; ἀνύω; ξηρός	3

Table 2: Randomly selected terms.

4.4 From translation alignment to WSD

To investigate how useful the word-aligned results are for WSD, we created two test sets of (a) words referring to metalinguistic concepts and (b) randomly selected polysemous words, as shown in Tables 1 and 2 respectively. The first set of words was handpicked by our team, as this work was initiated within the framework of a project focused on the onomasiology of linguistic concepts. The second set was chosen to extend the validation of our approach beyond the confines of this specific project. To be precise, we utilized the word list by Van Hal (2013), which provides information on the frequency (in four frequency bands) and polysemy of various Greek words, excluding those that are extremely common. From the first three frequency bands of Van Hal (2013), we randomly selected one noun, one adjective, and one verb.

[5] Note, however, that the target languages of these studies are all modern languages that are less inflectional than Greek and can utilize larger language models.

Next, for each of the target words listed in Table 1-2, we extracted the word alignments retrieved with our automatic models. The results were quite messy, containing many one-to-many alignments (likely due to our training data): an example (γλῶσσα) is shown in Table 3. Additionally, they contain inflected forms ('tongues') as well as function words such as articles and prepositions, due to linguistic differences between the two languages (i.e. Greek uses case marking, does not have an indefinite article and uses definite articles differently from English etc.). We therefore further cleaned the data by (a) tokenizing the results, (b) removing punctuation, (c) removing stop words and (d) lemmatizing each word in the results, using the NLTK packages *stopwords* and *WordNetLemmatizer* (Bird et al. 2009). After doing so, we further removed noise by calculating the frequency of each remaining lemma and removing all the lemmas that occur less than 1% in the total results. An example of the final output for γλῶσσα is given in Table 4. Although the table still contains some noise (e.g. the adjectives 'rare', 'good' and 'ordinary'), most of the results are clear translations of the word γλῶσσα.

Nevertheless, the results contained several synonyms or very closely related words (e.g. 'lip' and 'mouth' in Table 4). To use these results for WSD, they therefore need to be clustered in some way. In order to obtain a first idea which criteria the clustering should use, we performed the clustering manually, although automatic clustering is obviously necessary if one wants to scale up this approach to the full Greek corpus. Concretely, we used both frequency and meaning relatedness as criteria: in all cases, we clustered very closely related meanings (i.e. near-synonyms) together, but also clustered meanings when they were only somewhat closely related but were infrequent. In other words, we used a pragmatic criterion: if there were too little examples of a specific meaning, it would be problematic to learn this meaning through WSD, so it would be worth it to combine them with examples of another related meaning, even if some meaning granularity was lost by doing so. We did not assign irrelevant words to a cluster (e.g. 'rare', 'good', 'ordinary' in Table 4), but simply discarded them from the dataset. The results of the manual meaning clustering can be seen in Tables 5-8 in Appendix. To create a final dataset for WSD, for each cleaned up word alignment we checked if it contained any of the words assigned to one of the clusters, and if not, the example was discarded. Next, one could use these results to train models for WSD, take the tokens from the Greek corpus that were assigned to one of the meanings that they are interested in (e.g. the linguistic meaning of γλῶσσα in our case) and calculate the nearest neighbors based on these tokens, as detailed in Section 2. However, we did not perform this step in the scope of this paper.

Alignment	#	**Alignment**	#
tongue	57	my tongue	5
the tongue	18	in a tongue	5
tongues	11	of	5
with tongues	9	the tongues	5
a	7	speech	4
of the tongue	6	a tongue	4
.	5	lips	3
his tongue	5	…	…

Table 3: Example of word alignment results: γλῶσσα.

Alignment	#	**Alignment**	#
tongue	168	language	4
word	15	good	4
speech	12	voice	3
rare	8	mouth	3
lip	6	ordinary	3

Table 4: Cleaned results of γλῶσσα

5 Results

5.1 Sentence alignment

Firstly, we evaluated the model on the held-out data described in Section 4.1. This resulted in an accuracy (exact matches) of 73% (644/879). The F1-score, which also takes into account partial matches, was 86%.

Since such a quantitative evaluation can be misleading (since the test data might not entirely match our target corpus), we also manually conducted an evaluation of sentence alignment performance using 133 sentences from the target corpus chosen at random. The accuracy was 65% (86/133), somewhat lower than the 73% of the automatic evaluation, indicating that these results

might be too rosy.[6] Out of the remaining 35% of sentence alignments that were not correct, half of them (25/47) were partially correct, i.e. the Greek sentence contained the English translation but included more text, or vice versa.

The size of the training corpus at the sentence alignment task appears to be of great importance. It was our hypothesis that non-problematic corresponding sentences (in a 1-to-1 ratio, i.e. without Greek sentences that are mapped to multiple English sentences, or vice-versa) that were combined into artificial paragraphs (cf. Section 4.1) would contribute little. This turned out not to be the case. A model trained on data without these artificial paragraphs performed significantly worse, with an accuracy of 43% and an F1-score of 41% on the held-out data.

5.2 Word alignment

In contrast with the results shown in Li (2022), the span-extraction approach implemented in our model performed worse than the approach of Jalili Sabet et al. (2016) and Dou and Neubig (2021), as used by Yousef et al. (2022a). The comparison is difficult however, as they not only used another alignment approach, but also utilized another training dataset. Their best-performing model achieved an F1-score of 81.5, and an Alignment Error Rate (AER) of 18.7. It is, however, not exactly clear how the metrics are computed, viz. how punctuation and source words that do not have an alignment (e.g. untranslatable particles) are exactly handled. In the gold dataset, tokens without alignment are not annotated. Thus, it is not clear whether they are included in the evaluation or not.

The scores including these source tokens and punctuation, are an F1 of 47.7 and an AER of 43.5 (5076 tokens in total). If we leave these out, the F1 score rises to 59.6, and the AER is 35.9. For the scope of this project, the former evaluation is the most important, as the WSD task is mainly interested in content words such as verbs, nouns and adjectives. In contrast with these part-of-speech classes, the left-out tokens are mainly punctuation marks and untranslatable particles, which are of less importance for the WSD task.

5.3 Manual clustering of the results

The results derived from applying word alignment and subsequently manually clustering them, as outlined in Section 4.4, are presented in Tables 5-8 (found in the Appendix). A notable observation is that a considerable proportion of the data, accounting for 49% of all aligned tokens on average across all target words, included many translations that could not be neatly clustered (labelled as 'other' in these tables). This percentage varied from 24% (for μιαρός) to as high as 84% (for ἰστός). These typically fall into two categories: (a) words that were excluded by the frequency filter (see Section 4.4) or (b) incorrect word alignments. Concerning category (a), there are instances where the frequency filter eliminates relevant terms. A case in point is 'Latin' for γλῶσσα, which was filtered out despite clearly referring to the linguistic sense of γλῶσσα (contextually appearing in 'λέγειν ἱκανῶς ἑκατέραν γλῶτταν' which was roughly translated to 'to speak both Latin and Greek fluently'). Conversely, when the frequency filter is not used, the data evidently becomes cluttered with irrelevant results. For instance, some of the single-occurrence results for γλῶσσα include 'she-bear', 'of frigidity', and 'power of lubricating', which are unquestionably incorrect translations for γλῶσσα. Given that translation alignment at both the sentence and word levels only reaches a respective F1-score of 86 and 60 percent, it is inevitable that the data will contain numerous errors, resulting from either inaccurate sentence or word alignment.

Since the frequency threshold was relatively low, for less frequent words (viz. βίοτος, μιαρός, ἰστός, and ἀνύω) no words were filtered out, allowing us to assess how many alignment pairs were relevant for the task described in this paper. As can be deduced from Tables 5-8, for βίοτος 40% of all alignments were irrelevant, for μιαρός 24%, for ἰστός 84% and for ἀνύω 67%. This averages out to 54%, meaning that only half of the alignments were relevant for compiling a WSD dataset.

This has serious consequences for the possibilities of automating this approach. On the one hand, the frequency filter was absolutely necessary, given the amount of noise present in the data, which would make automatic clustering problematic. On the other hand, if an absolute frequency filter would have been used (e.g.

[6] Although the differences are barely statistically significant, with p=0.05 with Fisher's exact test.

filtering out translations that occur less than 3 times), this would lead to data sparsity for less frequent words. Therefore an obvious solution would be expanding the data, either by improving the alignment results or by adding more parallel English translations to the data.

On a brighter note, this method is clearly capable of retrieving a sufficient number of relevant examples for more frequent terms, thus creating a useful dataset for WSD. Nevertheless, there are several important considerations. Firstly, it is worth noting that the manual clustering was highly subjective: another researcher may well have grouped the words differently than we did. In such instances, an automatic clustering method might offer greater objectivity, even though automatic methods carry their own inherent biases. Generally speaking, the use of parallel translations is more effective when meanings can be more clearly differentiated (e.g., in the case of ἱστός, where there is a stark difference between 'mast' and 'loom'), rather than when the differences are somewhat vague (for instance, for λόγος, the distinctions between 'word', 'statement', and 'report' are not always easily discernible).

Secondly, the level of granularity that is possible to distinguish is dependent on the number of examples for a specific sense, especially when taking into account that some senses are more present in the data that we are using than other senses. While for λόγος many fine-grained distinctions can be made, for γλῶσσα only a general 'linguistic' sense can be distinguished, conflating the translations 'voice', 'speech', 'language' and 'word'. Meanwhile, for some WSD is not possible at all: for ξηρός all translations pointed to 'dry' (while the word also has other meanings in Greek, such as 'slim' and 'harsh').

Finally, one obvious issue is that this method assumes that the English translation equivalents do not have the exact same sense ambiguity as the Greek words. This does not always hold true. In the γλῶσσα-case, for instance, the English term 'tongue' can occasionally signify 'language', as exemplified in phrases like 'mother tongue'. This interpretation is also found in some of the more antiquated translations within our corpus. Another example is αἴσθησις, where 'sense' in English is similarly ambiguous between the meaning 'sensation, perception' and 'faculty for experiencing the outside world'. This issue could be solved in multiple ways, e.g. by using parallel translations from other languages that do not have this sense ambiguity. Alternatively, WSD could be conducted on the English data. However, this adds another automated step, which may potentially compromise the quality of the final results.

6 Avenues for better results

6.1 Improving alignment

Clearly, as the previous section demonstrates, inaccurate alignment results significantly curtail the volume of data that can be employed for WSD. Therefore, enhancing automatic alignments is a vital step towards further improvements.

On a foundational level, our work relied on an existing multilingual RoBERTa model, namely PhilBerta. However, given potential mismatches between the data format of PhilBerta and GLAUx data (for instance, in terms of Unicode encoding of accents or tokenization), it might prove beneficial to adopt an English-Greek model that is more closely attuned to the GLAUx data.

Regarding sentence alignment, potential improvements could be realized by augmenting the training data. Considering our current training set is rather limited (comprising 15,178 sentences), expanding it is one possible avenue for enhancing results (a step we are presently exploring; cf. Section 6.4). However, this inevitably entails a significant amount of manual work. An alternative strategy is to refine the alignment method itself. Our current method relies solely on word embedding information. While this might function effectively for language models with extensive data, Greek embedding models could be too sparse to be effectively deployed in isolation. Supplemental information might thus bolster the results, such as sentence position within a paragraph (naturally, Greek and English sentences tend to occupy similar positions within identical paragraphs) and the frequency of matches between the English translation of a Greek word, using a bilingual dictionary, and the English sentence. Moreover, the word alignment task could inform sentence alignment: very low probabilities in word alignment might signal that sentence alignment has misidentified a sentence. Lastly, an entirely different approach than the one employed in this study could also be considered. Adopting an unsupervised approach like VecAlign (Thompson and Koehn, 2019) could address the problem of

having to depend excessively on annotated examples.

Given that our method for word alignment is based on the same technique as sentence alignment, all the above considerations hold true for the former task as well. However, manually annotating word alignment proves to be even more labor-intensive than sentence alignment. Hence, unsupervised models may prove particularly advantageous for this task.

6.2 Improving the clustering

While the alignment results could be improved further, the task is inherently challenging and it is therefore likely that a significant amount of noise will always persist in the data. Thus, it is vital to implement effective techniques for filtering this noise. The simple frequency filter used in this study could potentially be too restrictive in some instances, such as with the Greek word μιαρός, which has several one-time translations for the concept 'miserable'. To address this, we might consider semantic similarity (operationalized through language models) as an additional criterion, specifically by including low-frequency translations if they show substantial semantic similarity to a higher-frequency translation.

For this study, we manually performed the clustering, but naturally, automatic clustering is necessary if we aim to extend this approach to the entire Greek corpus. A feasible method might involve clustering words with similar static embeddings in English.

6.3 Alternative methods

The applicability of new techniques for WSD and translation alignment, as discussed in Section 3.3, to Ancient Greek remains uncertain. When it comes to multilingual approaches, there is a scarcity of multilingual parallel corpora featuring Ancient Greek, with the exception of Biblical texts. However, repositories like <remacle.org> and *hodoi elektronikai* <hodoi.fltr.ucl.ac.be> could facilitate the creation of a trilingual Greek, English, and French corpus. The potential of propagation methods (which necessitate knowledge bases) and automatic translations in enhancing WSD in Ancient Greek is unclear.

One reviewer commented that instead of the method proposed in this paper, one could collect training data from dictionaries, as was done by Bamman and Burns (2020). Indeed, this was the strategy we initially pursued, using a digital version of the Liddell-Scott-Jones (LSJ) dictionary that we automatically linked with the GLAUx corpus. However, it soon became clear that relying exclusively on this dataset was only possible for a few highly frequent words with many examples in the dictionary: even when excluding the irrelevant word alignments (classified as 'other' in Table 5-8), the amount of data we could retrieve from word alignment was ten times larger than from the LSJ dictionary, and there was only one word (λόγος) for which we could retrieve more than 100 examples from LSJ (243 in total, which is still much smaller than the 3128 examples from word alignment). However, these dictionaries might still provide supplementary data, or provide a solid base for clustering the word alignments (i.e. by showing which English translations 'group together' for one specific meaning).

6.4 Progress made after peer review

While the results discussed in this paper might not seem too promising initially, we found that we were able to substantially improve the results by expanding and cleaning up the training data for both sentence and word alignment, and expanding our parallel Greek-English corpus with some other openly available translations.[7] For example, for the word γλῶσσα we now obtained 829 relevant results, after removing 187 results by applying the frequency filter, while previously we only had 210 relevant results after removing 152 results (see Table 8).

7 Conclusions and perspectives

In view of the extensive research conducted on WSD for modern languages, the comparative neglect of classical languages is striking. However, significant progress can be made in the near future to rectify this disparity, thanks in part to the comprehensive philological studies conducted in the past. With a robust lexicographical tradition replete with translated example sentences, and a prolific translation history, classical language resources, once available in a digital shape, have

[7] For sentence alignment, the accuracy rose from 73% to 85%, while the F1 score increased from 86% to 92%. For word alignment, the F1 score improved from 59.6 to 70.6, while the AER dropped from 35.9 to 24.0.

the potential to unlock promising possibilities for WSD applications.

The methodology presented in this paper appears to be a promising means to achieve our goals – coming to an onomasiological disclosure of the Ancient Greek corpus. A critical prerequisite, however, is the availability of a substantial volume of data, suggesting that the approach is effective predominantly for frequently used words.

Apart from this, we believe that this approach holds intrinsic value. For texts that have digital English translations available, we can make educated predictions regarding the meanings of the individual tokens. Additionally, this approach provides insights into the distribution of word senses as distinguished by lexicographers in Ancient Greek.

Acknowledgments

We would like to thank the reviewers for their valuable comments and suggestions. The research in this article was made possible by grant numbers HBC.2021.0210/KRIS:0508000001361 (VLAIO) and G052021N (FWO, Research Council – Flanders).

References

Bamman, David, and Patrick J. Burns. 2020. "Latin BERT: A Contextual Language Model for Classical Philology." ArXiv:2009.10053 [Cs], September. http://arxiv.org/abs/2009.10053.

Barba, Edoardo, Luigi Procopio, Niccolò Cam polungo, Tommaso Pasini, and Roberto Navigli. 2020. "MuLaN: Multilingual Label Propagation for Word Sense Disambiguation." In Proceedings of IJCAI, 3837–44.

Bird, Steven, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Chousa, Katsuki, Masaaki Nagata, and Masaaki Nishino. 2020. "SpanAlign: Sentence Alignment Method Based on Cross-Language Span Prediction and ILP." In Proceedings of the 28th International Conference on Computational Linguistics, 4750–61. Barcelona (Online): International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.418.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. "ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators." CoRR abs/2003.10555. https://arxiv.org/abs/2003.10555.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "Unsupervised Cross-Lingual Representation Learning at Scale." arXiv. https://doi.org/10.48550/arXiv.1911.02116.

Delli Bovi, Claudio, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. "EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 594–600. Vancouver: Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2094.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "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), 4171–86. Minneapolis).

Dou, Zi-Yi, and Graham Neubig. 2021. "Word Alignment by Fine-Tuning Embeddings on Parallel Corpora." arXiv. http://arxiv.org/abs/2101.08231.

Geeraerts, Dirk. 2010. Theories of Lexical Semantics. Oxford & New York: Oxford University Press.

Goossens, Diane. 2013. "Assessing Corpus Search Methods in Onomasiological Investigations." Corpus Perspectives on Patterns of Lexis 57: 271.

Jalili Sabet, Masoud, Philipp Dufter, François Yvon, and Hinrich Schütze. 2021. "SimAlign: High Quality Word Alignments without Parallel Training Data Using Static and Contextualized Embeddings." arXiv. http://arxiv.org/abs/2004.08728.

Keersmaekers, Alek. 2021. "The GLAUx Corpus: Methodological Issues in Designing a Long-Term, Diverse, Multi-Layered Corpus of Ancient Greek." In Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, 39–50. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.lchange-1.6.

Keersmaekers, Alek, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. "Creating, Enriching and Valorizing Treebanks of Ancient Greek." In Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019), 109–17. Paris: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7812.

Keersmaekers, Alek, and Toon Van Hal. 2021. “A Corpus-Based Approach to Conceptual History of Ancient Greek.” In Cognitive Sociolinguistics Revisited, edited by Gitte Kristiansen, Karlien Franco, Stefano De Pascale, Laura Rosseel, and Weiwei Zhang, 213–25. Berlin & Boston: Walter de Gruyter.

———. 2022. “In Search of the Flocks: How to Perform Onomasiological Queries in an Ancient Greek Corpus?” In Proceedings of the LREC 2022 Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), 73–83. Marseille.

Lefever, Els, Véronique Hoste, and Martine De Cock. 2011. “ParaSense or How to Use Parallel Corpora for Word Sense Disambiguation.” In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 317–22. Portland: Association for Computational Linguistics. https://aclanthology.org/P11-2055.

Lenci, Alessandro. 2018. “Distributional Models of Word Meaning.” Annual Review of Linguistics 4 (1): 151–71. https://doi.org/10.1146/annurev-linguistics-030514-125254.

Li, Bryan. 2022. “Word Alignment in the Era of Deep Learning: A Tutorial.” arXiv. https://doi.org/10.48550/arXiv.2212.00138.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv. https://doi.org/10.48550/arXiv.1907.11692.

Luan, Yixing, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. “Improving Word Sense Disambiguation with Translations.” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4055–65. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.332.

McGillivray, Barbara, Simon Hengchen, Viivi Lähteenoja, Marco Palma, and Alessandro Vatri. 2019. “A Computational Approach to Lexical Polysemy in Ancient Greek.” Digital Scholarship in the Humanities 34 (4): 893–907. https://doi.org/10.1093/llc/fqz036.

Mercelis, Wouter, Toon Van Hal, and Alek Keersmaekers. Forthcoming. “Tongue, language or noise? Word Sense Disambiguation in Ancient Greek with corpus-based methods.” In International Colloquium of Ancient Greek Linguistics.

Nagata, Masaaki, Katsuki Chousa, and Masaaki Nishino. 2020. “A Supervised Word Alignment Method Based on Cross-Language Span Prediction Using Multilingual BERT.” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 555–65. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.41.

Ng, Hwee Tou, Bin Wang, and Yee Seng Chan. 2003. “Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study.” In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 455–62. Sapporo: Association for Computational Linguistics. https://doi.org/10.3115/1075096.1075154.

Och, Franz Josef, and Hermann Ney. 2003. “A Systematic Comparison of Various Statistical Alignment Models.” Computational Linguistics 29 (1): 19–51.

Pasini, Tommaso. 2020. “The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation.” In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 4936–42. Yokohama: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/687.

van der Plas, Lonneke, and Marianna Apidianaki. 2014. “Cross-Lingual Word Sense Disambiguation for Predicate Labelling of French.” In Proceedings of TALN 2014 (Volume 1: Long Papers), 46–55. Marseille: Association pour le Traitement Automatique des Langues. https://aclanthology.org/F14-1005.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. “SQuAD: 100,000+ Questions for Machine Comprehension of Text.” arXiv. https://doi.org/10.48550/arXiv.1606.05250.

Riemenschneider, Frederick, and Anette Frank. 2023. “Exploring Large Language Models for Classical Philology.” arXiv. https://doi.org/10.48550/arXiv.2305.13698.

Smith, David A., Jeffrey A. Rydberg-Cox, and Gregory R. Crane. 2000. “The Perseus Project: A Digital Library for the Humanities.” Literary and Linguistic Computing 15 (1): 15–25.

Thompson, Brian, and Philipp Koehn. 2019. “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), 1342–48. Hong Kong: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1136.

Van Hal, Toon. 2013. Ankura. Basiswoordenlijst Oudgrieks. Antwerpen & Apeldoorn: Garant.

Yousef, Tariq, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022. "An Automatic Model and Gold Standard for Translation Alignment of Ancient Greek." In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 5894–5905. Marseille: European Language Resources Association. https://aclanthology.org/2022.lrec-1.634.

Yousef, Tariq, Chiara Palladino, Farnoosh Shamsian, and Maryam Foradi. 2022. "Translation Alignment with Ugarit." Information 13 (2): 65. https://doi.org/10.3390/info13020065.

A Appendices

Lemma	Translations	N
αἴσθησις	perception, sensation	118
	sense, faculty	43
	memory, knowledge, consciousness, opinion	20
	other (unclassified)	173
καταλύω	subvert, overthrow, undermine, suppress, destroy, abolish, depose, dissolution	109
	end, break, cease, stop	33
	lodge	7
	other (unclassified)	234
ἀλλότριος	another, property, others, belongs, belonging, else, possessions	148
	alien, foreign, strange, strangers	77
	other (unclassified)	210

Table 5: Results for randomly chosen terms (frequency band 1).

Lemma	Translations	N
βίοτος	life, age	40
	mean, victual, property, substance, gold, wealth, livelihood	16
	estate, house	6
	food	4
	other (unclassified)	44
ἀπαντάω	meet, encounter	96
	come, go	42
	confront	5
	other (unclassified)	238
μιαρός	infamous, wretch, bad, cruel, abominably, wretched, abominable, rogue, foul, trouble, …	54
	polluted, pestilential, stain, blood, defiled, unclean, pestilent, filthy	14
	other (unclassified)	21

Table 6: Results for randomly chosen terms (frequency band 2).

Lemma	Translations	N
ἱστός	raft, keel, ship, mast	7
	weaving, loom, tambour	6
	other (unclassified)	68
ἀνύω	attain, prove, accomplish, gain, achieve, finish, obtain, complete, reach, stop, fulfil	22
	haste, proceed, renew	4
	other (unclassified)	53
ξηρός	dry, withered, arid, wet, dried, barren, liquid, moist, dessicant, watery	126
	other (unclassified)	78

Table 7: Results for randomly chosen terms (frequency band 3).

Lemma	Translations	N
γλῶσσα	tongue, mouth, lip	177
	voice, speech, language, word	33
	other (unclassified)	152
φωνή	voice, cry, vocal	355
	speech, language, utterance, word, tongue	112
	sound	60
	other (unclassified)	294
λόγος	say, talk, speech, statement, said, saying, speak	1032
	word	850
	argument, reason	520
	story, report, discourse	433
	formula	92
	discussion	82
	account	66
	eloquence	53
	other (unclassified)	3616
ῥῆμα	name, saying, word, expression, term	108
	verb	17
	sentence, phrase	16
	speech	4
	other (unclassified)	92

Table 8: Results for linguistic terms.