Cross-linguistic annotation transfer in geoparsing experiments with Classical texts

The case of Pliny the Elder's *Natural History*

Laura Soffiantini[1]

[1]KU Leuven

The *Natural History* is an encyclopedic work written by the Latin author Pliny the Elder (first century CE). In this extensive text in 37 books, geography plays a pivotal role, with hundreds of mentions of ancient place names. In this paper, a geoparsing experiment is conducted on the *Natural History* with the scope of automatically identifying and extracting place entities. To achieve this, we take advantage of state-of-the-art NLP models to develop a multistage pipeline involving English Named Entity Recognition, English-Latin sentence alignment, and entity projection. The paper demonstrates how cross-lingual annotation transfer can be applied from a translation in a modern language back to the original text in the context of low-/medium-resource languages, such as Latin. The efficacy of the proposed pipeline is evaluated through the use of both standard metrics and a comprehensive manual error analysis. Additionally, the results are compared to those obtained by other Latin NER tools. Both analyses demonstrate that the proposed methodology achieves a superior f1-score. Finally, the majority of place entities were automatically associated with unique identifiers that enable geolocation by the projection of pre-disambiguated annotations derived from another geo-spatial project.

Keywords: Named Entity Recognition, Latin, text alignment, annotation projection, Pliny the Elder

1 Introduction

Geoparsing is the process that involves the automated identification of geographical entities within an unstructured text (Gregory et al., 2015). This task plays a significant role in the realm of Natural Language Processing (NLP) and information retrieval, and more specifically as a subtask within the framework of Named Entity Recognition (NER) and Named Entity Linking (NEL). NER consists of detecting and classifying named entities, that is, elements in texts that act as a rigid designator for a reference, typically of the types *Person*, *Location*, and *Organization*. NEL is a related but more complex task that involves disambiguating entities through a knowledge base. The kind of activity that we call "geoparsing" in a computational domain has long been of

interest to scholars, with applications in a range of research fields. From linguistics to the science of the environment through history and literature, the identification of the place entities mentioned in written sources has several applications representing the starting point of various kinds of data analysis.

In the context of geoparsing historical data, a number of additional factors come into play. One of the most significant is the often-noted underdevelopment of NLP systems tailored for languages beyond the modern ones. The scarcity of robust linguistic models, coupled with the fact that they are usually trained on specific language domains, significantly restricts their fine-tuning to limited NER tasks (Ehrmann et al., 2023; González-Gallardo et al., 2023; Won et al., 2018). Further challenges are posed by the nature of the spatial knowledge embedded in historical documents. Such knowledge often reflects a world that no longer exists, where city names may have changed, villages may have disappeared, and rivers may have changed course over time. In addition, the very paradigms used to understand space, including spatial concepts and spatial categories, are culture-specific (Geus and Thiering, 2014). The scope of the geoparsing analysis in historical data is, therefore, to address the variability in how individuals have historically described their environment by investigating the shifts that have influenced the perception and conceptualisation of space (Barker et al., 2016; Palladino, 2021).

In this study, I discuss a geoparsing case study of the *Natural History* (*NH* henceforth), an encyclopaedic work written by the Latin author Pliny the Elder in the first century CE. This massive 37-book text covers a wide range of topics, including geography, medicine, astronomy and art, among others. In particular, geography plays a pivotal role, since Pliny aims to provide a comprehensive description of the world (Doody, 2010; Murphy, 2004). As a result, significant portions of the encyclopaedia are devoted to the description of various geographical entities (Books 3-6), with an extensive catalogue of "bare names" (*NH* 3.2: *nuda nomina*) of cities, mountains, rivers, and seas. In some cases, Pliny is the only witness to a place name, making him an unparalleled source of knowledge of ancient geography. In the rest of the work, places are mentioned in order to anchor the various topics in space. For instance, in the books on botany and agriculture place names occur in relation to the origin of species[1] or the provenance of goods.[2]

The aim of this paper is to provide a pipeline for the automatic identification of place entities in the *NH*. Taking advantage of state-of-the-art NLP models and applying them in the context of low-resource languages, the paper presents a projection-based methodology for annotating named entities and specifically places in Latin texts through a multi-stage workflow that outperforms the currently available NER taggers in terms of precision and recall. The data, along with the code for the project, is available online in the GitHub repository for further review.[3]

2 Related work

Until a few years ago, the automatic identification of place names in historical texts was carried out using more or less sophisticated string matching-techniques. This approach relied on the existence of manually curated gazetteers in digital format, specifically designed to curate authority lists of place names from the ancient Greek and Latin

1 *NH* 12.53: *in **Aegypto** nascitur et maron*, "Cat-thyme also grows in Egypt."
2 *NH* 13.2: *Et macir ex **India** advehitur*, "Also macir is imported from India."
3 `https://github.com/laurensmile/Geoparsing_Pliny/tree/main`.

world. By including place names in original language and modern translation, these resources, such as the *Pleiades Gazetteer*[4] and *Trismegistos Place*,[5] held immense potential for conducting automated geoparsing experiments on classical texts. In 2015, Kiesling presented the *ToposText* project,[6] whose scope was the annotation of all mentions of place names in classical literature in English translation. The *ToposText* annotations were primarily derived from the *Pleiades Gazetteer* in the sense that the place names were automatically extracted by various string-matching heuristics and disambiguated by an alphanumeric unique identifier linked to an external knowledge base. Another relevant project in the field of classical geography was *Hestia*, conducted by Barker et al. (2013). *Hestia* investigated the geography of the ancient world as described by the Greek author Herodotus in his *Histories*. One of the goals of the project was to georeference all the places mentioned by Herodotus. To do this, the project relied on the annotations available in the English translation of the *Histories* from the *Perseus Digital Library* (Crane et al., 2000), which were derived from the application of string-matching techniques. A consistent limitation of this methodology, however, is that it is inherently dependent on the comprehensiveness and granularity of the gazetteers used. Place names that are not already present in the authority list cannot be retrieved from the text and unseen data is missed. Finally, string-matching methods have only been applied to modern translations of classical texts, while the annotation of the original sources has never been experimented with.

In recent years, the advent of machine learning models has ushered in a new era of technologies for the geo-spatial analysis of historical texts. This shift is exemplified by the development of various pre-trained Named Entity Recognition systems, which have transformed the automated identification of place names in unstructured texts. The strength of these models lies in their ability to generalize from training data, operating independently of the completeness of the reference list, and allowing the expansion of gazetteers with unseen, new data. In the context of ancient languages, Erdmann et al. (2016, 2019) developed a neural BiLSTM-CRF (Bidirectional Long-Short Term Memory) entity classifier trained on Latin texts (*Herodotos-Project-Latin-NER-Tagger*) with the scope of annotating ancient ethnic groups. More recently, Torres Aguilar (2022) fine-tuned the transformer-based BERT and RoBERTa models for Medieval Latin and the *SpaCy* pipeline for Latin (*LatinCy*) became available, including a NER toolkit (Burns, 2023). In addition, Beersmans et al. (2023) fine-tuned two new transformer-based *LatinBERT* models for the task of NER. In their conclusions, the authors pointed out that the identification of places achieved lower performances in terms of accuracy when compared to the annotation of persons, and that multi-tokens place entities were rarely recognized. Similar conclusions were reached by Palladino and Yousef (2024), who found that place names are still the most challenging entity type for Ancient Greek NER due to the lack of a consistent annotation strategy in the available training datasets. Finally, in the context of classical geography, it is worth mentioning the development of *Recogito*[7], a *JavaScript* library specifically designed for geo-spatial tasks. The tool integrates the use of the *Stanford CoreNLP* models for NER tagging with the default language model for English only, but it has also experimented with Latin NER through the *Herodotos-Project-Latin-NER-Tagger* plugin.

One potential solution to the challenge of performing NLP tasks on languages

[4] `https://pleiades.stoa.org`.
[5] `https://www.trismegistos.org/geo/`.
[6] `https://topostext.org`.
[7] `https://recogito.pelagios.org`.

with limited resources is the cross-lingual projection, which involves transferring linguistic annotations over parallel texts from a high-resource language to a low- or medium-resource language. Applying cutting-edge machine learning technologies to poorly annotated corpora, this approach has proven to be adaptable across various language pairs (Jain et al., 2019). In the context of classical languages, Yousef et al. (2023) recently fine-tuned a multi-lingual XML-RoBERTa-based language model (*UGARIT/gr-alignment*)[8] for the word alignment, and transferred NER annotations across parallel corpora in Ancient Greek, Latin and English. The language model was trained on a substantial corpus of parallel sentences, primarily in Ancient Greek-English and Ancient Greek-Latin, and then tested in conjunction with a variety of alignment heuristics (Yousef et al., 2022a,b).

3 Corpus

For this study, the parallel corpus of the *NH* in Latin and English was used. The Latin text (487,289 tokens) was obtained from the *PerseusDL/canonical-latinLit*[9] in the .XML tokenized and lemmatized form published by Clérice (2021). For the English translation (804,866 tokens), I used the *Perseus Project* .HTML text file available from *ToposText*.[10]

4 Methodology

The pipeline consisted of four main steps, which are fully described in the following sub-sections. The first step involved preprocessing the *ToposText* English text file, as described in 4.1. The second step entailed applying the state-of-the-art *flairNLP* system to the English translation of the *NH* for the automated NER prediction (4.2). This resulted in the *ToposText* annotations being integrated with new place entities detected by the NER. In parallel, an automated alignment model was employed to align the English and Latin texts at the sentence (4.3) and word levels (4.4). By extracting the translation equivalents for place names, the annotations were transferred accordingly by a direct projection heuristic. As a result, an automatically annotated Latin text for place entities was obtained. Place name attestations that were already annotated in *ToposText* were also automatically geo-identified and linked to an external knowledge base. The results obtained from each stage of the workflow are discussed in the Results.

4.1 Preprocessing

The .HTML text file from *ToposText* was tokenized using the *segtok* library.[11] In a .CSV file, each token was associated with a unique identifier and with the reference position at the book.chapter.paragraph level. For each token within a *ToposText* <place> tag (e.g., <place>Athens</place>), the *ToposText* unique identifier provided within the tag (e.g., '380237PAth' for 'Athens') was extracted and associated with the corresponding token. The original file contained 8,895 *ToposText* tags for places.

[8] `https://huggingface.co/UGARIT/grc-alignment`.

[9] `urn_cts_latinLit_phi0978.phi001.perseus-lat2.xml`.

[10] Apart from *ToposText*, no other publicly available digital annotated version of the *NH* exists. In 2022, the *Pliny the Elder's World* project conducted by B. Turner and R. Talbert published a list of geo-referenced place names from the geographical books of the *NH* (`https://https://isaw.nyu.edu/research/pliny-the-elder/`).

[11] `https://pypi.org/project/segtok/`.

As there is currently no evaluation of the quality of the *TopposText* annotation, our first objective was to assess its accuracy against a ground truth annotation. The gold standard was manually curated for Book 4 (18,664 tokens), dedicated to the geographical description of Central Europe and particularly representative of the entity types to be identified, consisting of a long list of place names. The annotations were encoded in Beginning-Inside-Outside (BIO) format. According to the BIO-format conventions, the first or only word of an entity is indicated with the B- prefix, whereas words within a multi-word entity are specified by the prefix I-, as shown in the example below (Table 1). Tokens located outside the entities are marked with an 'O'.

The	O
city	B-LOC
of	I-LOC
Athens	I-LOC
.	O

Table 1: Named entity tagging in BIO style of the phrase "The city of Athens."

In order to ensure consistency, clear guidelines were established, encompassing specific criteria for named entity types and annotation granularity. In summary, the gold standard was crafted in accordance with the conventions established by Romanello and Najem-Meyer (2022) as follows:

- **Definition of the entity type**: locations (LOC). References to named places within the text – that is, all the proper names for geographic locations – were annotated.[12] These include:
  - administrative locations including regions, cities, towns, districts, and villages;
  - physical places such as mountains, deserts, plains, islands, lands, seas, rivers, lakes, canals, and springs;
  - streets, squares, roads, and highways;
  - buildings, such as temples, sanctuaries, gates, forts, and stations.
- **Entity boundaries**: The annotation encompasses multi-word entities, including pre- and post- modifiers, such as "Mount Pindus", "gulf of Corinth" and "Red Sea". Articles are excluded from the annotation. The gold standard does not contain nested annotations.
- **Adjectives** derived from toponyms (e.g., "Corinthian") are only annotated if they pertain to places (e.g., "Corinthian gulf"), not to other entities such as people (e.g., "Perikles the Athenian"), or objects (e.g., "Corinthian bronzes").
- **Ethnics** are omitted from the annotation. In ancient sources, certain regions were named after the groups of people inhabiting them, rather than having specific place names. However, including ethnic names in the annotation would

[12] The notion of "place" is not straightforward, and there is a lively scholarly debate about the conceptualization of spatiality in antiquity. For practical reasons, an extensive discussion is beyond the scope of this paper. In this study, a broad definition of "place" will be adopted in accordance with the guidelines provided by the *Pleiades Gazetteer*.

have extended beyond the project's primary focus, which centered on physical locations.

Each named location was manually annotated in the gold standard. Subsequently, the gold standard was employed as a benchmark to assess the accuracy of various systems of automated place name extraction, as described in Section 5.1.

4.2 Named Entity Recognition

In this step, a comparative evaluation was conducted of various NER systems to identify the optimal model for the translation of the *NH*. Two cutting-edge English NER tagging tools, *SpaCy*[13] and *flairNLP*,[14] were tested for the annotation of LOC labels (in the case of *Spacy*, both the LOC and GPE labels were extracted). The inspection of the results revealed that the *flairNER-large* model exhibited superior accuracy and recall capabilities, as discussed in Section 5.1. Finally, a validation process was initiated to determine whether each identified entity was either fully or partially represented in *ToposText*.

4.3 Automatic Text Alignment

The parallel texts were automatically aligned using *Vecalign* (Thompson and Koehn, 2019), a state-of-the-art method for bilingual sentence alignment. The method operates on the assumption that when parallel sentences from two different languages are represented in a vector space, their sentence embeddings become closely related. *Vecalign* was employed in conjunction with embeddings generated by the Language-agnostic BERT Sentence Embedding (*LaBSE*) (Feng et al., 2022), a multilingual transformer model fine-tuned for bitext mining. *LaBSE* has been demonstrated to be more accurate than other models, particularly in scenarios where there is limited training data, as recently evidenced by Petruccioli (2022) in the context of the alignment of Latin texts.

The *Vecalign* method is initialized with the input of the text files in .TXT format, one for each language, containing a list of the sentences to be aligned. The sentences were obtained by segmenting the text at the level of the full stop and, in the case of English, also in the presence of the punctuation marks *?* and *!*. During the process of splitting Latin sentences, certain challenges were encountered due to some specific features of the text. These include the frequent use of full stops to indicate abbreviations (e.g. abbreviated proper nouns) and the presence of noise (e.g. sequences of *? ? ?*) due to issues in the quality of OCR. Consequently, only full stops preceded by a token longer than two characters were considered for sentence splitting. In general, it was observed that the use of longer sentences was preferable when employing the *Vecalign* method.

Vecalign generated .TXT overlap files containing sentence concatenations (i.e., sentence 1, sentence 1 + sentence 2, sentence 1 + sentence 2 + sentence 3, and so on) from the input .TXT files. The overlap files were employed to generate sentence embeddings. *Vecalign* then employed the pre-processed files, the overlap files and the derived sentence embeddings to compute sentence similarity. In *Vecalign*, each sentence can be aligned to one (one-to-one alignment) or more sentences (one-to-many alignment). The output of the alignment process was a list of index pairs, corresponding to the indexes of the aligned sentences, along with a floating-point number representing the computed alignment cost. The scoring function was based on the normalized cosine

[13] `https://spacy.io/`.

[14] `https://github.com/flairNLP/flair`.

distance between the sentence embeddings, whereby the higher the cost, the more distant are the embeddings of the aligned sentences. In the event that no alignment was found, the alignment cost was set to zero.

4.4 Annotation Projection

The translation alignment is the process of establishing a correspondence between words in the source text and their equivalents in the translation. In this experiment, I used *SimAlign* (Sabet et al., 2020), an automated alignment method that utilizes cross-lingual semantic similarity between tokens based on static and contextualized embeddings. Among the various extraction algorithms proposed by Sabet et al. (2020), the iterative method *Itermax* was selected, which showed better performance particularly for distant languages. *Itermax* employs the simple *SimAlign* baseline iteratively, by modifying the similarity matrix conditioned on the alignment edges of the preceding iteration. The model also allows a token to be aligned with several other tokens, improving the identification of alignment edges when a single word corresponds to two or more words in the target language. The word embeddings were derived from the *UGARIT/gr-alignment* language model (see Related work), which had previously been tested in conjunction with *SimAlign* and *Itermax* (Yousef et al., 2023).

The aligned sentences obtained in the previous step (4.3) were used as input for the word alignment process. Initially, each sentence was tokenized, and the resulting tokens underwent vectorization. The cosine similarity was then computed as a similarity measure, and the *Itermax* algorithm extracted the translation pairs from the similarity matrix. The outcome of the word alignment process was a list of word index pairs for each sentence pair. Finally, the annotations were projected from the English tokens to their Latin equivalents. Given that the word order may differ between English and Latin, and tokens representing the same entity may not be consecutive, the projection process did not account for the distinction between B(eginning)- and I(nside)- labels. Each English attestation of a place name was associated with an identifier, which was linked to all the token(s) pertaining to the entity (e.g., 'id1' and 'id2' in Figure 1). Once the equivalent Latin token(s) had been aligned, the identifier was transferred. Finally, new BIO labels were assigned according to the relative position of the tokens.

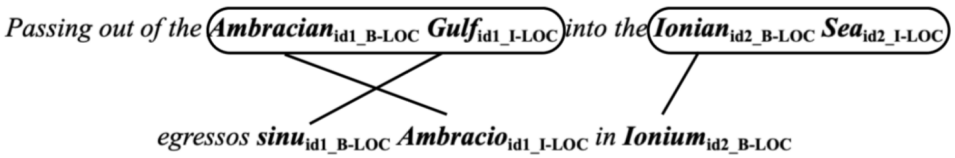

Figure 1: Example of the annotation projection between two sentences.

5 Results

The combination of the *ToposText* annotations and the output of the *flairNER-large* resulted in the annotation of 13,964 place entities in the English text of the *NH*. The automated projection yielded 14,391 annotated Latin tokens, comprising 13,099 B-LOC and 1,292 I-LOC annotations. A total of 8,404 place entities were automatically

disambiguated in the Latin text. The following sections present an assessment of the accuracy of the output of the different steps of the workflow when applied out-of-the-box to Book 4, for which gold data was manually curated. The quantity and type of errors occurring at each stage were examined, and their impact on the final projection was analysed. Finally, the competing NER methods *LatinCy* and *Herodotos-Project-Latin-NER-Tagger* were applied directly to the Latin text of Book 4, and the output was compared with the result of the projection. With regard to *LatinCy*, I tested the large (*la-core-web-lg*) pipeline and the *la-core-web-trf* model, which is backed by the multilingual BERT transformer architecture.

5.1 Named Entity Recognition

The evaluation of the annotation task was conducted using three standard classification metrics: precision, recall, and f1-score. Evaluated against the gold standard, the *ToposText* annotation in Book 4 exhibited remarkably high precision (0.991), indicating that the majority of the annotations are correct (Table 2). However, it should be noted that *ToposText* is substantially incomplete, with a considerable number of place names missing, which as resulted in a low recall value (0.671). Given the absence of guidelines for entity boundaries in *ToposText*, the quality assessment was conducted at the named entity level, instead of the token level, and partial matches with the gold data were considered as true positives as long as they contained at least one capitalized word. To illustrate, if the gold standard included the annotation "Mount [B-LOC], Pindus [I-LOC]" and, at the same position, *ToposText* only annotated "Pindus [B-LOC]" (but not the pre-modifier), this would be counted as correct, despite differing labels for tokens at the same position.

	Place Entities	Precision	Recall	F1-score
ToposText	1,309	**0.991**	0.671	0.800
flairNER	1,338	0.905	0.637	0.748
flairNER-large	1,912	0.952	0.938	0.945
SpaCy-trf	1,937	0.938	0.938	0.938
ToposText + flairNER-large	1,983	0.949	**0.968**	**0.958**
Gold standard	1,931			

Table 2: Precision, recall and f1-scores of *ToposText*, *flairNLP*, *SpaCy* and *ToposText* combined with *flairNER-large* on Book 4.

When applied on Book 4, the NER systems *flairNER*, *flairNER-large*, and *SpaCy-trf* produced different outputs, as illustrated in Table 2. *FlairNER-large* outperformed the other systems, achieving a high-quality annotation with an f1-score of 0.945, along with strong precision and recall values. The model failed to identify 119 place entities and while it did manage to detect 115 of these, they were not correctly labeled as places. Notably, *flairNER-large* and *SpaCy-trf* exhibited similarities in their annotation styles, as evidenced by the annotation agreement (PAA = 81.7) computed by counting label matches at the token level.

The combination of *ToposText* with *flairNER* led to a further improvement in results with total of 795 new annotations. In comparison to using *ToposText* alone, there was a significant increase in recall (0.968), which was primarily due to a substantial drop in

false negatives. Although the resulting precision decreased from 0.991 to 0.949, due to an increase in false positives, the overall annotation quality showed a considerable improvement. This enhancement is highlighted by the f1-score, which reached the highest value obtained (0.958).

5.2 Automatic Text Alignment

The *Vecalign* method produced 395 alignments between 510 English sentences and 457 Latin sentences in Book 4. Manual inspection of a sample of the alignments revealed that out of 200 English sentences, 191 were correctly aligned, 8 were partially aligned and 1 was misaligned. In our evaluation, we considered an alignment to be strictly correct (SC) if the English sentence was completely contained within the aligned Latin sentence(s). Conversely, if the English sentence was not found within the aligned Latin sentence(s), we categorised it as misalignment (MIS). Alignments where the English sentence only partially overlapped within the aligned Latin sentence(s) were labelled as partially correct (PC).

In the qualitative assessment of the results, we found that the most influential factor on the quality of the alignment was the closeness of the translation to the parallel text, in particular similarity in sentence length and syntactic distribution. In cases where syntactic differences were present, *Vecalign* tried to align the sentences through one-to-many alignments, which occasionally led to errors. For example, consider the following alignments:

> [185] At this spot there is another Isthmus, similar in name to the other, and of about equal width; and, in a manner by no means dissimilar, two cities formerly stood on the shore, one on either side, Pactye on the side of the Propontis, and Cardia on that of the Gulf of Melas, the latter deriving its name from the shape which the land assumes.
>
> [169] *Alius namque ibi Isthmos angustias similes eodem nomine et pari latitudine inlustrat.* [170] *Duae urbes utrimque litora haut dissimili mode tenuere, Pactye a Propontide, Cardia a Melane sinu, haec ex facie loci nomine accepto,* ***utraeque conprehensae postea Lysimachea V p. a Longis Muris****.*

> [186] **These, however, were afterwards united with Lysimachia, which stands at a distance of five miles from Macron Tichos**. [187] The Chersonesus formerly had, on the side of the Propontis, the towns of Tiristasis, Crithotes, and Cissa, on the banks of the river Aegos; it now has, at a distance of twenty-two miles from the colony of Apros, Resistos, which stands opposite to the colony of Parium.
>
> [171] *Cherronesos a Propontide habuit Tiristasin, Crithoten, Cissam flumini Aegos adpositam; nunc habet a colonia Apro XXII p. Resisthon, ex adverso coloniae Parianae*.

The English sentence with index [186]: "These, however, were afterwards united with Lysimachia, *etc.*" was incorrectly aligned with the Latin sentence [171], rather than the corresponding Latin sentence [170]: *utraeque conprehensae postea Lysimachea* etc. This discrepancy can be attributed to the fact that the English translation does not accurately reflect the syntactic structure of the Latin text. In this instance, the Latin sentence [170] was translated into two separate English sentences [185] and [186], which resulted in the alignment method failing to correctly associate one of them with its corresponding Latin counterpart.

A comparable observation can be made when examining partial alignments, as illustrated by the following example.

> [14] Passing out of the Ambracian Gulf into the Ionian Sea, we come to the coast of Leucadia, with the Promontory of Leucate, and then the Gulf and the peninsula of Leucadia, which

last was formerly called Neritis. [15] By the exertions of the inhabitants it was once cut off from the mainland, but was again joined to it by the vast bodies of sand accumulated through the action of the winds.

[12]*Egressos sinu Ambracio in Ionium excipit Leucadium litus, promunturium Leucates, dein sinus et Leucadia ipsa paeninsula, quondam Neritis appellata , opere accolarum abscisa continenti ac reddita ventorum flatu congeriem harenae adtumulantium,* ***qui locus vocatur Dioryctos stadiorum longitudine trium***.

[16] **This spot is called Dioryctos, and is three stadia in length: on the peninsula is the town of Leucas, formerly called Neritus**.

[13] *Oppidum in ea Leucas, quondam Neritum dictum.*

In this case, the English sentence [16] was aligned with the Latin sentence [13], which only partially overlaps it. In particular, only the latter part of the sentence, "On the peninsula is the town of Leucas, formerly called Neritus", corresponds to the Latin phrasing. Conversely, the first part of the English sentence, "This spot is called Dioryctos, and is three stadia in length", aligns with the Latin sentence [12]: *qui locus vocatur Dioryctos stadiorum longitudine trium*. Since the information distribution differs between the two texts, this resulted in a partial alignment. Finally, no discernible patterns emerged between the alignment scores of strictly correct, partially correct and mis- alignments.

5.3 Annotation Projection

The utilization of *SimAlign* on Book 4 resulted in the projection of 1,850 entities on a total of 1,983, with 133 entities missed. The quality assessment of the annotation transfer was performed on a sample of 100 sentences for a total of 2,107 tokens and 485 manual annotated place entities. As for the English gold standard, the quality assessment of the Latin NER was conducted at the entity level, with at least one capitalized word annotated.

As illustrated in Table 3, our annotation projection achieved the best overall performance, attaining the highest f1-score (0.923). This value is significantly higher than the other f1 scores, in particular the *LatinCy-trf* system. This result can be attributed to a significantly higher recall value of our method (0.956), which can be explained by the fact that the initial English *ToposText+flairNER* annotation was almost complete.

	Precision	Recall	F1-score
LatinCy-lg	0.762	0.550	0.639
LatinCy-trf	**0.949**	0.232	0.374
Herdotos-Project	0.877	0.340	0.490
Annotation projection	0.892	**0.956**	**0.923**

Table 3: Precision, recall and f1-scores of *LatinCy* models, *Herodotos-Project-Latin-NER-Tagger* and our projection method on 100 sentences of Book 4.

The qualitative analysis of the results showed that several factors influenced the alignment quality. First, false negative and false positive annotations in the English NER led to projection errors. Second, correct sentence alignment proved critical for word alignment. In fact, partially aligned or misaligned sentences consistently caused word misalignments. Third, the quality of the alignment depended on the accuracy of

the translation. In our case study, English place names are generally very similar to the Latin original, which contributed to correct alignment. In addition, Book 4 contains long lists of names with relatively simple syntactic structures, which facilitated token alignment. In other cases, however, a place name was the translation of a Latin ethnic name ("Thessalians" for *Thessalia*), which resulted in an incorrect annotation of the Latin text, demonstrating a consistent limitation of the transfer approach. Finally, the projection of proper names was found to be more accurate than the projection of common nouns in multi-token entities. In certain instances, annotated English common nouns lacked a Latin equivalent, but were added by the translators for the sake of clarity or due to English grammatical constraints. When transferred, the projection resulted in an erroneous annotation of the Latin text, as no suitable equivalent could be identified. In other instances, despite the presence of an equivalent Latin word, the model failed to align it correctly. This is exemplified by the sentence "Here is also Mount Taygetus, the river Eurotas, the town of Psamathus, the gulf of Gytheum" (*NH* 4.8), where the word "gulf" was aligned with the word *oppidum* instead of the correct word *sinus*.

6 Conclusions

In this study, we conducted a geoparsing experiment on Pliny the Elder's *NH*. To address the challenges of performing NER tasks on languages with limited resources, a cross-lingual multi-stage approach was used to automatically align sentence pairs and transfer annotations from the English translation back to the original Latin text. The methodology proved to be more effective for this task than other NER baselines applied directly to the Latin text, showing better performance, especially in terms of recall. Moreover, the approach can be generalized to other Latin texts, exploiting on the potential of translated parallel texts to expand Latin NER datasets without the necessity for additional training data. However, many factors can influence the performance of the annotation. The quality achieved in each automated component of the workflow (English NER, sentence alignment, word alignment) significantly influences the final output and errors propagate across the pipeline. Accurate translations are required to achieve good results and various factors, including syntactic complexity, text genre, and language style, can influence the efficacy of the alignment process. In addition, the resulting Latin annotation lacks consistency in annotation style, especially in the case of multi-token entities, due to the absence of guidelines in *TopposText* and the differing annotation styles of *TopposText* and *flairNER*. Finally, the projection method is not context-aware, unlike NER systems, which attempt to capture the meaning of words in the source language. In future work, the pipeline will be tested on documents that are not geographically specific and that exhibit a greater degree of syntactic complexity, without the benefit of prior annotation. This will permit further evaluation of the effectiveness of the NER and the annotation transfer across parallel corpora of ancient sources and their modern translations.

References

Elton Barker, Stefan Bouzarovski, Christopher Pelling, and Leif Isaksen. On using a digital text in modern humanities research: the case of Herodotus' Histories. *Bulletin of the Institute of Classical Studies. Supplement*, 9(122):45–62, 2013. URL `http://www.jstor.org/stable/44216322`.

Elton Barker, Stefan Bouzarovski, Christopher Pelling, and Leif Isaksen. *New Worlds from Old Texts: Revisiting Ancient Space and Place*. Oxford University Press, 2016.

Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. Training and Evaluation of Named Entity Recognition Models for Classical Latin. In *Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023*, pages 1–12, 2023. URL https://aclanthology.org/2023.alp-1.1/.

Patrick J. Burns. LatinCy: Synthetic Trained Pipelines for Latin NLP. (*arXiv:2305.04365*). *ArXiv:2305.04365* [*cs*], 2023. URL http://arxiv.org/abs/2305.04365.

Thibault Clérice. lascivaroma/latin-lemmatized-texts: 0.1.2. 2021. URL https://doi.org/10.5281/zenodo.4731513.

Gregory Crane, David A. Smith, and Jeffrey A. Rydberg-Cox. The Perseus Project: a digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000. URL https://doi.org/10.1093/llc/15.1.15.

Aude Doody. *Pliny's Encyclopedia: The Reception of the Natural History*. Cambridge University Press, 2010.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named entity recognition and classification on historical documents: A survey. *Association for Computing Machinery*, 56(2):1–47, 2023. URL https://doi.org/10.1145/3604931.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities* (*LT4DH*), pages 85–93, 2016. URL https://aclanthology.org/W16-4012/.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2223–2234, 2019. URL https://aclanthology.org/N19-1231/.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891, 2022. URL https://aclanthology.org/2022.acl-long.62.

Klaus Geus and Martin Thiering. *Feature of Common-Sense Geography: Implicit Knowledge Structures in Ancient Geographical Texts*. LIT Verlag, 2014.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G. Moreno, and Antoine Doucet. Yes but.. Can ChatGPT Identify Entities in Historical Documents? In *2023 ACM/IEEE Joint Conference on Digital Libraries* (*JCDL*), pages 184–189, Los Alamitos, CA, USA, 2023. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/JCDL57899.2023.00034.

Ian Gregory, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. GIS, Geoparsing, and Textual Analysis: New Trends in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, 9(1), 2015. URL https://www.dedoose.com/publications/geoparsing%20gis%20and%20textual%20analysis%20current%20developments%20in%20spatial%20humanities%20research.pdf.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. Entity projection via machine translation for cross-lingual NER. 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)*, pages 1083–1092, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1100.

Trevor M. Murphy. *Pliny the Elder's Natural History: The Empire in the Encyclopedia*. Oxford University Press, 2004.

Chiara Palladino. Mapping the unmapped: Transmedial representations of premodern geographies. *BGL Berichte Geographie und Landeskunde*, 94(2):139–160, 2021. URL https://doi.org/10.1093/llc/15.1.15.

Chiara Palladino and Tariq Yousef. Development of robust NER models and named entity tagsets for Ancient Greek. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 89–97, Torino, Italia, May 2024. URL https://aclanthology.org/2024.lt4hala-1.11.

Giulia Petruccioli. Automatic sentence alignment of latin texts and their translations in vernacular french using multilingual sentence embeddings. exploring different neural network approaches., 2022.

Matteo Romanello and Seven Najem-Meyer. Guidelines for the annotation of named entities in the domain of classics. 2022. URL https://doi.org/10.5281/zenodo.6368101.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, 2020. URL https://aclanthology.org/2020.findings-emnlp.147.

Brian Thompson and Philipp Koehn. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1136.

Sergio Torres Aguilar. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lt4hala-1.17.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5, 2018. URL https://www.frontiersin.org/articles/10.3389/fdigh.2018.00002.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 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, 2022a. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.634.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. Automatic translation alignment for Ancient Greek and Latin. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France, June 2022b. European Language Resources Association. URL https://aclanthology.org/2022.lt4hala-1.14.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. Named entity annotation projection applied to classical languages. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 175–182, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.latechclfl-1.19/.