"Gotta catch ‘em all!": Retrieving people in Ancient Greek texts combining transformer models and domain knowledge

Marijke Beersmans[1], **Alek Keersmaekers**[1], **Evelien de Graaf**[1], **Tim Van de Cruys**[1], **Mark Depauw**[1], **Margherita Fantoli**[1], [1]KU Leuven,

Correspondence: marijke.beersmans@kuleuven.be

Abstract

In this paper, we present a study of transformer-based Named Entity Recognition (NER) as applied to Ancient Greek texts, with an emphasis on retrieving personal names. Recent research shows that, while the task remains difficult, the use of transformer models results in significant improvements. We, therefore, compare the performance of four transformer models on the task of NER for the categories of people, locations and groups, and add an out-of-domain test set to the existing datasets. Results on this set highlight the shortcomings of the models when confronted with a random sample of sentences. To be able to more straightforwardly integrate domain and linguistic knowledge to improve performance, we narrow down our approach to the category of people. The task is simplified to a binary PERS/MISC classification on the token level, starting from capitalised words. Next, we test the use of domain and linguistic knowledge to improve the results. We find that including simple gazetteer information as a binary mask has a marginally positive effect on newly annotated data and that treebanks can be used to help identify multi-word individuals if they are scarcely or inconsistently annotated in the available training data. The qualitative error analysis identifies the potential for improvement in both manual annotation and the inclusion of domain and linguistic knowledge in the transformer models.

1 Introduction

Identifying the mentions of people in texts is one of the goals of the broader task of Named Entity Recognition (NER). For scholars working on historical texts, accurately finding and identifying people is particularly valuable for studying the representation of individuals, both in qualitative and data-driven studies. The present research, for instance, is embedded in a broader project aiming at performing large-scale analysis on the mentions of individuals in Ancient Greek and Latin texts (NIKAW, Networks of Ideas and Knowledge in the Ancient World).

For classical languages, and Ancient Greek in particular, the task remains challenging to automate. This study capitalises on recent advancements in transformer models, which have shown promising improvements over previous approaches. After introducing the available methods and data for NER on Ancient Greek (Sections 2 and 3), in Section 4, we compare four recent transformer models of Ancient Greek and their performance for NER, with a focus on identifying mentions of people. This comparison allows the selection of a model for further exploration. Since the Ancient World has a wealth of domain-specific resources on offer, in Sections 5, we focus on the specific task of predicting PERS entities by simplifying the NER task, and we explore how integrating gazetteers (Section 5.2) and syntactic annotations (Section 5.3) can impact the process of pinpointing individuals in texts. In the qualitative error analysis in Section 5.4, we identify several shortcomings of the reduced transformer method and discuss how domain knowledge and linguistic information impact the performance. With this, we contribute to advancing NER for Ancient Greek, identifying the strengths and limitations of currently available models and data and offering concrete suggestions for the way forward.

2 Related Work

The task of NER for historical languages presents several challenges, which can be traced back to four main factors (Ehrmann et al., 2023): diversity of sources, noisiness of data, language change, and lack of resources. These challenges are transferable to Ancient Greek and Latin corpora. However, the use of transformer models yields promising results: this is demonstrated for Latin by Torres Aguilar (2022); Beersmans et al. (2023), and for Ancient Greek by Yousef et al. (2023); Pal-

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 152–164
August 15, 2024 ©2024 Association for Computational Linguistics

ladino and Yousef (2024). Palladino and Yousef (2024) present two transformer models finetuned for the task of Ancient Greek NER. This model was created by training a XLM-RoBERTa-based multilingual model that was previously fine-tuned on the word alignment task for ancient languages, including Ancient Greek (Yousef et al., 2022a,b) and an Ancient-Greek-BERT model (Singh et al., 2021) respectively.

In this paper, we compare the NER performance of four transformer models for Ancient Greek, described in detail in Section 4 and 5. In addition, recent studies highlight the advantages of incorporating domain knowledge, in particular gazetteers, in the training of NER models, especially for low-resource languages (Zafarian and Asghari, 2019; Fetahu et al., 2022; Song et al., 2020). Gazetteers are external resources that often take the form of name dictionaries, grouped by a specific entity type (e.g. location or person). To leverage the advantage of domain knowledge, we incorporate the Trismegistos Gazetteers of names and name variants (TM NamVar) (Broux and Depauw, 2015)[1] and of places (TM GeoVar)[2] in two approaches described in Section 5.2. This rejoins the efforts of exploiting available knowledge bases for annotating Ancient Greek texts, as discussed in Berti et al. (2019). Finally, we address the problem of multi-token entities, which are particularly difficult to label automatically given their sparsity in the training data and the potential complexity added by factors such as overlap, nesting, and non-consecutiveness (Xia et al., 2019; Alshammari and Alanazi, 2021; Byrne, 2007; Crane, 2011). In Section 5.3, we explore the effectiveness of expanding single-token entities into multi-token entities using syntactical dependencies.

3 Data

3.1 Datasets for training and testing

There is currently no dedicated openly available benchmark dataset for Ancient Greek NER.[3] However, scholars have been annotating entities in Ancient Greek texts for a variety of goals, such as the mapping of places.[4] We combined four of such annotated Ancient Greek texts and harmonised their annotation through rule-based means. Our harmonised corpus contains data from the following projects (details summarised in Table 1): First: the *Odyssey* (henceforth OD) (Pelagios, 2021). Second, the EpiDoc XML of the *Deipnosophistae* of Athenaeus of Naucratis (DEIPN), retrieved from the Perseus digital library.[5]. Third, the Stepbible corpus (SB), available on GitHub (STE, 2023), which contains the full Ancient Greek New Testament (for further details, see Section 3.2). And finally: Pausanias' *Periegesis Hellados* (PH), courtesy of the Periegesis project (Foka et al., 2021). For information on originally annotated entity types per dataset, please refer to Table 12 in appendix C.

In addition, we manually annotated a random sample of 596 sentences from the GLAUx corpus (Keersmaekers, 2021) to test the generalisability of the results to all literary Greek material (GLAUx TEST). GLAUx contains most of the literature produced in Greek between the 8th century BC and the 4th century CE (about 27 million tokens). It is partly manually and partly automatically annotated for morphology, lemmas and syntax. While the predictions were made on the (tokenized) text, the morphological and syntactic annotation and the lemmas were used for further experiments (for details, see Section 5.3). The annotation process of GLAUx is described in Section 3.3.

3.2 Data Harmonisation

Since the datasets described in the previous section followed different guidelines, data harmonisation was necessary, following the steps detailed here.

- All entities were projected from their original files onto the GLAUx XML files to ensure similar Unicode character encoding, linguistic enrichment, tokenization, and capitalisation standards.
- Similarly to Palladino and Yousef (2024), we mapped the original annotated entities to a PERS, LOC, GRP scheme (Appendix C). PERS is used for identifiable individuals, LOC for geographical locations (both natural and human-built) and GRP for ethnonyms, nationalities and organisations. As the OD lacked a category suitable for conversion to GRP, this dataset was not used in the full NER

[1] https://www.trismegistos.org/ref/about_naw.php.

[2] https://www.trismegistos.org/geo/about.php.

[3] Palladino and Yousef (2024) compiled a dataset similar to this one, but it is not publicly available.

[4] See for instance the geographical visualisation available for the *Odyssey*.

[5] For Named Entity retrieval tools for this text in particular, see *The Digital Athenaeus project* (Berti, 2021).

text	# tokens	annotation method	period	genre
PH	242,433	manual	2nd century AD	travelogue
DEIPN	314,256	semi-automatic	3rd century AD	encyclopedic dialogue
OD	104,364	manual	8th century BC	epic poetry
SB	158,325	manual	1st -2nd century AD	religious

Table 1: Available datasets for Ancient Greek NER

but only in the reduced model described in Section 5.

- We used the morphological tags available in the GLAUx corpus to convert all plural words annotated as a person (often Muses, Cyclopes, etc.) to GRP.
- The TITLE-category of the SB corpus also caused issues, including references to Jesus, the biblical God, and cults. To disambiguate, all capitalised singular titles (e.g. Jesus Christ) were re-annotated as PERS, all capitalised plural titles were re-annotated as GRP (e.g. Pharisees) and all non-capitalised titles (e.g. the non-capitalised word ‘god’) were discarded.
- The PH dataset contains annotated pronouns or references to entities that do not include a name (e.g. ‘the island’). We rely on capitalisation and discard all entities that do not include at least one capitalised word. For consistency, this rule was adopted in all datasets, even though non-capitalised entities were rare in the others.
- For all datasets, all entities that were not annotated with one of our final entity types (i.e. PERS, LOC, GRP), e.g. Συμποσίῳ, ‘in the Symposium’, referring to the title of a work, were dropped.

Finally, we split the data in a train, validation and test set using a 75%-12.5%-12.5% split. After harmonisation, multi-token entities were scarce (see Table 2, a total of 2,376 on 55,454 entities). In DEIPN, for example, no multi-token entities were annotated.

3.3 Annotation of GLAUx

As mentioned before, the overarching goal of our project is to conduct a large-scale analysis of the mentions of individuals in Ancient Greek (and Latin) texts. For this purpose, we start from the GLAUx corpus (Keersmaekers, 2021), introduced in Section 3.

In order to evaluate the performance of the model on GLAUx, we annotated a random sample of 596 sentences, each containing at least one capitalised word, for a total of 1,012 entities (excluding the ones annotated as O),[6] as shown in Table 2. We annotated the entity types PERS, LOC, and GRP, following the definitions described in Section 3.2. For multi-token entities such as e.g. Ἀρχαγόρας Ἀργεῖος, ‘Archagoras the Argive’, we allowed nested annotation: in this case B-PERS I-PERS for the entire string, with an additional B-GRP for Ἀργεῖος.

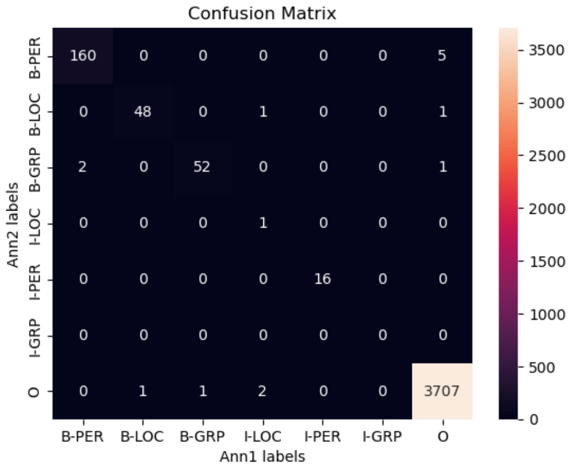

Figure 1: Confusion matrix for the IAA on GLAUx

172 sentences of the random GLAUx sample were annotated by two of the co-authors, resulting in an Inter Annotator Agreement (IAA) of 0.97 (Cohen’s kappa coefficient), calculated on word level. When excluding the O’s, the two annotators agreed on the label of 95% of the entities. The confusion matrix is shown in Figure 1. After IAA was calculated, both annotators discussed the differences to agree on a final annotation.[7] Surprisingly, the

[6] Entities annotated as O are those that do not fit the PERS, LOC, GRP scheme, such as, for example, book titles, titles of people without an actual named entity (e.g. Caesar or Pharaoh), and astronomical entities.

[7] Detailed information, both concerning the original annotations used to compute the IAA and the final annotation after discussion, can be found in the document *final_glaux_sample_iaa.csv* on our GitHub repository: https://github.com/NER-AncientLanguages/

	TRAIN	TRAIN$_{ody}$	VAL	VAL$_{ody}$	Held out TEST	TEST$_{ody}$	GLAUx TEST
B-PERS	21,307	2,033	4,054	381	3,090	400	578
I-PERS	290	25	122	1	83	0	51
B-LOC	8,261	699	1,345	85	1,105	76	233
I-LOC	1,061	18	278	0	196	0	11
B-GRP	8,884	41	1,291	2	1,384	4	201
I-GRP	182	0	9	0	49	0	0
O	494,668	75,248	81,968	12,547	83,182	12,519	13,454

Table 2: Entities annotated in the train, validation and tests sets. The *ody* datasets are exclusively used for the models predicting PERS/MISC. The GLAUx TEST dataset was annotated for this project to evaluate performance on data representative of all Ancient Greek literature.

main source of confusion was the attribution of the B-PER label, where one of the two annotators assigned O five times. This mostly concerned names mentioned as names or nicknames, that serve as additional specifications for a different, already mentioned entity. For instance, in the sentence "and they call his name ‘the Emmanuel’", ‘Emmanuel’ was not considered an entity by one of the annotators. After discussion, these cases were considered entities in the final annotation. The annotators also disagreed twice on the annotation of a standalone ethnonym, here referring to a specific individual: the "Samaritan" was annotated by one annotator as B-PERS and by the other as B-GRP. The annotators agreed on B-GRP, to be consistent with the plural occurrences of ethnonyms. Concerning differences in boundaries, in the case of sequences such as Φᾶσιν ποταμὸν, ‘river Phasis’, only one of the two annotators included ποταμὸν, ‘river’, in the entity. The final annotation includes both words.

4 Models for normal NER

In this section, we compare the performance of four transformer-based models for NER. We have a twofold objective: determine the best-performing model for the general NER task,[8] and determine to what extent the inclusion of domain knowledge can improve the results of the best-performing transformer models.

4.1 Trained models

We trained a total of four models and tested them on both the Held out TEST and GLAUx TEST datasets. Two of these models are also included in Palladino and Yousef (2024): the first is Ancient Greek BERT (henceforth AG_BERT), a modern Greek BERT model fine-tuned on Ancient Greek text data from the Perseus Digital Library and the First1KGreek project (Singh et al., 2021). The second is a multilingual XLM-RoBERTa model fine-tuned on Perseus data, the First1KGreek project, and various treebank datasets for Ancient Greek translation alignment, developed in the context of the UGARIT project (henceforth UGARIT). Because our training data differ from theirs, we retrained the two models instead of comparing metrics for the fine-tuned models directly. In all cases, we used a random 10-fold hyperparameter search to optimise the weight decay, the learning rate, and the number of epochs to maximise the F1 score on the validation dataset. The search space and final hyperparameters are detailed in Tables 7 and 8 in Appendix A.

We added two other models for comparison. Firstly, Ancient Greek ELECTRA-small (henceforth ELECTRA) (Mercelis and Keersmaekers, 2022), trained on Ancient Greek texts from Homer up until the 4th century CE. It is smaller than the other models and significantly faster to train. Secondly, GrεBerta (Riemenschneider and Frank, 2023), an XLM-RoBERTa model trained on a corpus of 200 million Ancient Greek tokens. The texts are partially sourced from digitisation projects such as the Perseus Digital Library and First1KGreek and partially from OCRed text from the Internet Archive.

4.2 Results on test sets

Table 3 shows the results of the four models on the ‘Held out TEST’ and the ‘GLAUx TEST’ sets. Metrics are calculated on the entity level (e.g. for multi-word entities, all comprising words of said

`NERAncientGreekML4AL`.

[8] The best model will be published on HuggingFace upon acceptance, while the code for training the models is available on GitHub (ibid.)

entities must be correctly annotated by the model to be considered a true positive). Unless otherwise specified, we indicate the F1 score per category. The evaluation focuses on the assignment of entity type to every token and thus I-tags are not explicitly shown in the table because of the inconsistency of the annotation of these entities in the training data, as done by (Palladino and Yousef, 2024). However, it is important to note that Recall for I-tags of all types was low, as can be seen in Table 10 in Appendix B. This can be attributed to their relative scarcity in training and validation data, and a way to improve these results is discussed in Section 5.3.

First, it is notable that all the models perform better on the Held out TEST than on the GLAUx TEST. For PERS, the best-retrieved category, this translates into a minimum drop of 0.01 (GrεBerta) to a maximum of 0.05 (UGARIT). Secondly, while on the Held out TEST AG-BERT, ELECTRA and UGARIT have a very similar performance, on the GLAUx TEST, AG-BERT outperforms the other three.

5 Predicting PERS entities (Reduced models)

Because the overarching project in which this research is embedded is primarily interested in the mentions of people, and because, as demonstrated by Table 3, the prediction of LOC and GRP entities is more difficult than PERS, the next part of the paper focuses on adapting the NER task to predict individuals as comprehensively and consistently as possible. We propose the three following approaches:

- Simplify the task from standard NER to predicting whether a single token references a person (PERS) or not (MISC) (see 5.1).
- Incorporate information from the TM NamVar and GeoVar gazetteers as either a post-processing rule or a binary mask added to the model input (see 5.2).
- Utilise the GLAUx syntactic dependencies to (re)create multi-token entities after annotation by the models (see 5.3).

5.1 Training models to predict PERS-MISC

To create the data for the simplified NER task, which only predicts an entity label (PERS or MISC) for every capitalised token, and by default predicts O for all other tokens, we automatically re-annotated all capitalised words of the entity type PERS without B- or I- specifications: so, for example, the name ‘Simon Petrus’ was re-annotated as PERS PERS. This process causes a difference in entity count compared to the data used for the normal model, as visible in the ‘support’ columns of Tables 3 and 4. All other capitalised tokens were annotated as MISC. Non-capitalised tokens are always classified as non-entities. Critical editions of Ancient Greek text often lack a sentence-initial capital, so it is reasonable to assume that anything that is capitalised is an entity of some kind. In earlier work, capitalisation in critical editions of Ancient Greek (and Latin) texts has been similarly leveraged for NER e.g. in the Perseus Project (Crane, 2011) and Trismegistos (Broux and Depauw, 2015). We use the same base models and hyperparameter optimisation method as described above for the normal NER (details available in Table 9 in Appendix A). The results in Table 4 show that all models perform well on this task, with AG_BERT marginally outperforming the others. We thus only use this model (from now on AG_BERT_simple), for gazetteer and dependency incorporation.

5.2 Gazetteer approaches

As detailed in Section 2, including domain knowledge in the training of NER models may be advantageous. Here, in collaboration with the Trismegistos team, we explore the incorporation of the TM gazetteers NamVar and GeoVar (see Section 2), authoritative lists widely used in the field of ancient history.

TM NamVar aims at an exhaustive coverage of personal names attested in Ancient Greek (800 BCE - 800 CE), including all spelling and linguistic variants. For names outside Egypt, TM NamVar has integrated the Greek Lexicon of Personal Names (LGPN).[9] The coverage of the regional LGPN volumes varies over time, e.g. regarding the inclusion of non-Greek names. TM is in the process of adding whatever names are missing, both in epigraphic and in Greek literary texts. Currently there are 81,588 Greek name variants (out of a total of 239,201 for all languages and scripts). TM GeoVar for Ancient Greek currently focuses mainly on spelling and linguistic variants of place names found in texts from Egypt

5.2.1 Rule-based approach (AG_BERT_rule)

To create AG_BERT_rule, a post-processing rule was added to the prediction of AG_BERT_simple:

[9] https://www.lgpn.ox.ac.uk/.

	AG-BERT	**Electra**	**GrεBerta**	**UGARIT**	**support**
			Held out TEST		
PERS	**0.87**	0.86	0.76	0.86	3,090
LOC	**0.73**	0.71	0.57	0.73	1,105
GRP	0.81	0.80	0.68	**0.83**	1,384
Macro F1	0.80	0.79	0.67	**0.81**	5,579
			GLAUx TEST		
PERS	0.78	0.76	0.73	**0.79**	578
LOC	**0.75**	0.71	0.60	0.66	233
GRP	0.78	**0.78**	0.73	0.76	201
Macro F1	**0.77**	0.75	0.68	0.74	1,012

Table 3: Results (F1 score) for NER per label on in-domain (Held out TEST) and out-of-domain (GLAUx TEST) data

	AG_BERT	**Electra**	**GrεBerta**	**UGARIT**	**support**
			Held out TEST		
PERS	**0.90**	0.87	0.83	0.89	3,539
MISC	**0.90**	0.88	0.84	0.89	3,706
macro F1	**0.90**	0.88	0.83	0.89	7,245
			GLAUx TEST		
PERS	**0.88**	0.84	0.81	0.85	605
MISC	**0.88**	0.87	0.83	0.86	699
macro F1	**0.88**	0.86	0.82	0.86	1,304

Table 4: Results (F1 score) for the prediction of PERS and MISC labels on in-domain (Held out TEST) and out-of-domain (GLAUx TEST) data

if the lemma of a capitalised token appears in TM NamVar, but not in TM GeoVar, it is always classified as a person. Both on Held out TEST and on GLAUx TEST, this approach increases Recall (by ca. 0.03 points) but has a detrimental effect on Precision (drop of more than 0.06 points) (see Table 5).

5.2.2 Machine Learning approach (AG_BERT_mask)

For AG_BERT_mask, we incorporated the rule described in Section 5.2.1 as input for the model. A binary mask was added to the training data where 1 indicated the rule applied and 0 that it did not. This mask was provided as additional input information to the model. We retrained AG_BERT_simple with the same final hyperparameters as described in Section 5. The results in Table 5 show that while no effect is visible on Held out TEST, this approach improved Precision on GLAUx TEST from 0.84 to 0.90, with a slight drop in Recall (from 0.92 to 0.91). We thus conclude that the ML approach yields better results than the rule-based approach, and we integrate the syntax on the top of AG_BERT_mask.

5.3 Incorporating syntax for the retrieval of multi-token entities (AG_BERT_syntax)

In the training data, names with ethnonyms and patronyms are rarely annotated as multi-token entities. They are frequently annotated as two separate entities, as is the case DEIPN (e.g. Λεωνίδης ὁ ῾Ηλεἵος, ‘Leonides of Elis’, annotated as B-PERS O B-GRP) and PH (e.g. Δεϰελεύς Σωφάνης, ‘Sophanes of Decelea’, annotated as B-GRP B-PERS), although there are exceptions (e.g. in PH Θεοδώρου τοῦ Σαμίου, ‘Theodorus of Samos’, annotated as a B-PERS I-PERS I-PERS). However, for the disambiguation and linking of people retrieving the full name is crucial.

	AG_BERT_simple			AG_BERT_rule			AG_BERT_mask			support
	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
	Held out TEST									
PERS	0.88	**0.93**	**0.90**	0.79	**0.96**	0.87	0.88	**0.93**	**0.90**	3,539
MISC	**0.93**	0.88	**0.90**	**0.96**	0.75	0.84	**0.93**	0.88	**0.90**	3,706
Macro	**0.90**	**0.90**	**0.90**	0.87	0.86	0.86	**0.90**	**0.90**	**0.90**	7,245
	GLAUx TEST									
PERS	0.84	0.92	0.88	0.78	**0.95**	0.86	**0.90**	0.91	**0.90**	605
MISC	0.92	0.84	0.88	**0.95**	0.76	0.85	0.92	**0.91**	**0.91**	699
Macro	0.88	0.88	0.88	0.86	0.86	0.85	**0.91**	**0.91**	**0.91**	1,304

Table 5: Results (Precision, Recall and F1 score) for the prediction of PERS and MISC labels on in-domain (Held out TEST) and newly annotated (GLAUx TEST) data, by not including the Gazetteer (AG_BERT_simple), including the Gazetteer with a rule-based approach (AG_BERT_rule) and with a mask (AG_BERT_mask).

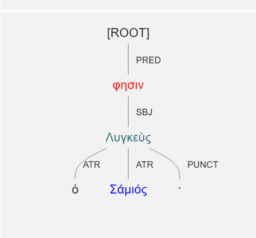

Figure 2: Dependency representation of sentence 1793 in DEIPN. https://perseids-publications.github.io/glaux-trees/0008-001/2066

In this approach, we rely on a dependency-based representation of Ancient Greek sentences as shown in Figure 2. If a capitalised word (in this case Λυγκεὺς) is annotated as a PERS by AG_BERT_mask, we check whether any of the direct children of said word is capitalised and re-annotate the entity as a multi-token. Thus, in this example, Λυγκεὺς ὁ Σάμιος, 'Lynceus of Samos', is re-annotated as B-PERS O I-PERS. Table 6 shows the results of dependency incorporation (AG_BERT_syntax) compared to the performances of the AG_BERT trained on the available data with respect to B-PERS, I-PERS. Only capitalised words are taken into account for calculating the metrics. For AG_BERT, the MISC category is created by grouping together all predictions of non-PERS tags. As shown in Table 6, dependency information greatly improves results for I-PERS tokens.

5.4 Qualitative error analysis

We performed a qualitative error analysis on the predictions of the models described in sections 5.2 and 5.3. We first describe the errors of AG_BERT_simple compared to AG_BERT_rule and AG_BERT_mask (as seen in **??**), and second, evaluate the improvement on multi-token entities with AG_BERT_syntax (as seen in 6).

5.4.1 Difficult categories

Several entity categories can be identified where AG_BERT_simple failed to predict correctly and neither AG_BERT_rule nor AG_BERT_mask offered any improvement. First, all predict MISC for nicknames such as Κακεργέτης, 'the Evildoer', or for tokens that frequently appear as non-capitalised common nouns in the training data, e.g. the PERS entity Λύχνος, the name of a deity, identical to the non-entity λύχνος, with the meaning of 'candle'.

Second, PERS is predicted for many of the MISC entities that are capitalised tokens annotated by experts as O: for example, capitalised tokens such as mathematical notations to designate geometrical entities such as points, lines, circles, etc. in texts such as Euclid's *Elementa* (GLAUx ID: 1799-001). Other examples are capitalised tokens that are entities that do not fit the PERS, LOC, GRP scheme (see 3.3) such as titles of books (e.g. Γραφή, 'the Scripture', i.e. the Bible) and titles for people (e.g Φαραώ, 'Pharaoh', and Καίσαρα, 'Caesar'). Overall, these issues stem from mismatches between training and testing data: some, such as mathematical entities were not present in the training data, others, such as titles for people, were annotated differently.

	AG_BERT_syntax			**AG_BERT**			**support**
	Pr	Rc	F1	Pr	Rc	F1	
B-PERS	0.88	0.89	0.89	0.90	0.81	0.85	581
I-PERS	0.70	0.60	0.65	0.50	0.02	0.04	50
MISC	0.91	0.91	0.91	0.82	0.95	0.88	673
macro avg	0.83	0.80	0.81	0.74	0.59	0.59	1,304

Table 6: Results for retrieving multi-word PERS entities using the syntax approach compared to training on available data, on the newly annotated GLAUx_test

5.4.2 Difference between AG_BERT_rule and AG_BERT_mask

The predictions of the two gazetteer models show significant differences. AG_BERT_mask improves upon AG_BERT_rule in cases where an entity appears in TM NamVar but is a MISC entity, such as GRP entities that in singular could be a person, e.g. Νύμφαι, 'Nymphs'. Second, AG_BERT_mask is the only model that correctly predicts MISC for the majority of the mathematical entities described above. In the few cases where AG_BERT_rule was an improvement on AG_BERT_simple and AG_BERT_mask not, issues stem again from the inconsistencies in the training data. Sometimes forms of the same word appear annotated as different entity categories, e.g. ῞Αιδου, 'Hades', annotated as O, PERS or LOC. The annotation with PERS and LOC stems from the inherent ambiguity of the word Hades, which can indeed refer both to the god Hades and the underworld. In other cases there are differences in annotation choices between the harmonised training data and our annotation, e.g. epithets annotated as PERS in GLAUx TEST, but as MISC in TRAIN. Lastly, for nested entities, AG_BERT_mask predicts the overarching level where the other two predict the second level, e.g. for Κωνσταντίνου [B-LOC nested B-PERS] ἀγορὰν [I-LOC], 'the Forum of Constantine', AG_BERT_mask predicts MISC for Κωνσταντίνου, 'Constantine'.

Under-representation of certain types of entities is also an issue for AG_BERT_mask. One example is personal names ending in an alpha. A specific case is personal names ending in -ίᾳ (feminine noun, dative ending): tokens with this ending are primarily annotated MISC (total: 441, total PERS: 83) in the training data (e.g. Ἀδρίᾳ, 'the Adriatic', MISC with mask = 1), resulting in the prediction of MISC instead of PERS for tokens ending in -ίᾳ such as for Ἀμεινίᾳ, 'Ameinias', with mask = 1. Only when the exact same form appears in the training data, is the prediction correct. For those tokens with mask = 1, naturally AG_BERT_rule's prediction is always correct. Training on the gazetteer mask had a detrimental effect for AG_BERT_mask in this case as several MISC entities in this category, like the examples given above, did receive a mask = 1, allowing the model the possibility that forms like this can be MISC even though they have mask = 1.

5.4.3 Syntax models

Last, AG_BERT_syntax shows significant improvement in predicting I-labels as compared to the AG_BERT model, as described in Section 5.3. This approach improved multi-token entity recognition for entities consisting of up to three separate tokens or with up to three non-entity tokens present between the B- and I- tokens. However, for multi-token entities that have both more than two tokens and gaps between the B- and I- tokens, performance is not increased. The majority of these errors are not caused by any error in the method but either by incorrect syntactic information encoded in GLAUx as the result of automatic analysis or because our rule-based method of using the syntactic trees could not retrieve all I-entities, e.g. we did not add special rules for coordination, which is complicatedly annotated in the syntactic annotation of GLAUx (see Section 3.3).

6 Conclusion

The goal of our study is to consistently and fully automatically annotate attestations of people using transformer-based NER. We trained several transformer models on available data for Ancient Greek NER and evaluated performance both on a Held out TEST set and on randomly annotated data representative for Greek literary data. While all models performed adequately, we conclude that inconsistency in annotation remains an obstacle in achieving high performance —which is in line with the findings by Palladino and Yousef (2024) and Beersmans et al. (2023), especially concerning multi-token entities. The approaches introduced in Sections 5.1-5.3 increase the performance for detecting persons specifically, but we recognise that there is still room for improvement (see Section 7). In future work, we will consider the integration of other available gazetteers,[10] and incorporate attestation counts as weights. The syntactically informed annotation of multi-token entities could equally benefit from an improvement of the rule-based extraction through a more careful analysis of the structure of I-entities in the dependency tree.

7 Limitations

One of the main limitations is our dependency on the capitalisation choices of the compilers of the (digital) editions we rely on. This also makes this approach difficult for truly transferring to even more low-resource languages. Secondly, gazetteers cannot ensure complete coverage of the attestations. In addition, we aimed at finding an exact match between the lemma in the text and the form resulting in the gazetteer. For this reason, small language variations resulted in a mismatch between the text and the gazetteer form. This could be addressed by allowing a certain degree of variation. For the use of syntactic relations, we largely relied on automatic parsing, a notably hard task, which resulted in some missed retrievals due to erroneous syntactic annotation. This aspect is hard to address because large-scale manual syntactical annotation is not achievable.

8 Acknowledgements

We thank the creators of data, and in particular Brady Kiesling, Elton Barker and Rainer Simon (Periegesis project and Recogito) and Chiara Palladino and Josh Kemp (Annotated Odyssey) for providing us with theirs as per our request.

References

2023. Stepbible data repository cc by 4.0.

Nasser Alshammari and Saad Alanazi. 2021. The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, 22(3):295–302.

Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training and evaluation of named entity recognition models for classical Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Monica Berti. 2021. *Digital Editions of Historical Fragmentary Texts*. Propylaeum.

Monica Berti, K. Simov, and M. Eskevich. 2019. Named Entity Annotation for Ancient Greek with INCEpTION. In *Proceedings of CLARIN Annual Conference 201*, pages 1–4, Leipzig. CLARIN 2019.

Yanne Broux and Mark Depauw. 2015. Developing onomastic gazetteers and prosopographies for the ancient world through named entity recognition and graph visualization: Some examples from trismegistos people. In *Social Informatics*, Lecture Notes in Computer Science, page 304–313, Cham. Springer International Publishing.

Kate Byrne. 2007. Nested Named Entity Recognition in Historical Archive Text. In *International Conference on Semantic Computing (ICSC 2007)*, pages 589–596, Irvine, CA, USA. IEEE.

Gregory R. Crane. 2011. Scalable named entity identification in classical studies.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2022. Dynamic Gazetteer Integration in Multilingual Models for Cross-Lingual and Cross-Domain Named Entity Recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2777–2790, Seattle, United States. Association for Computational Linguistics.

Anna Foka, David A. McMeekin, Kyriaki Konstantinidou, Nasrin Mostofian, Elton Barker, O. Cenk Demiroglu, Ethan Chiew, Brady Kiesling, and Linda Talatas. 2021. *Mapping Ancient Heritage Narratives with Digital Tools*, page 55–65. Ubiquity Press.

Alek Keersmaekers. 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*, page 39–50, Online. Association for Computational Linguistics.

Wouter Mercelis and Alek Keersmaekers. 2022. electra-grc.

[10] Another example of an Ancient Greek and Latin gazetteer is Pleiades,`https://pleiades.stoa.org/`

Chiara Palladino and Tariq Yousef. 2024. Development of robust ner models and named entity tagsets for ancient greek.

Pelagios. 2021. Beyond translation: Building better greek scholars.

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

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021)*, pages 128–137. Association for Computational Linguistics.

Chan Hee Song, Dawn Lawrie, Tim Finin, and James Mayfield. 2020. Improving Neural Named Entity Recognition with Gazetteers. *arXiv preprint*. ArXiv:2003.03072 [cs].

Sergio Torres Aguilar. 2022. 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*, page 119–128, Marseille, France. European Language Resources Association.

Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. 2019. Multi-grained Named Entity Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1430–1440, Florence, Italy. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2023. Transformer-based Named Entity Recognition for Ancient Greek. In *Digital Humanities 2023. Book of Abstracts*, pages 420–422, Graz. Centre for Information Modelling - Austrian Centre for Digital Humanities.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022a. 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, David J. Wright, and Monica Berti. 2022b. 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.

Atefeh Zafarian and Habibollah Asghari. 2019. Improving NER Models by exploiting Named Entity Gazetteer as External Knowledge. In *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 107–111, Trento, Italy. Association for Computational Linguistics.

A Hyperparameters

parameter	values
learning rate	uniform distribution: $[1 \times 10^{-6}, 1 \times 10^{-4}]$
weight decay	$\{0.1, 0.01, 0.001\}$
number of training epochs	$\{3, 4, 5, 6\}$

Table 7: Hyperparameter search space

	AG_BERT	ELECTRA	GrεBerta	UGARIT
learning rate	6.041e-05	9.889e-05	2.715e-05	5.784e-05
weight decay	0.01	0.1	0.01	0.01
epochs	3	5	4	5

Table 8: overview final hyperparameters on the regular NER task

	AG_BERT/AG_BERT_mask	ELECTRA	GrεBerta	UGARIT
learning rate	1.263e-05	8.703e-05	2.961e-05	2.490e-05
weight decay	0.01	0.1	0.1	0.001
epochs	6	5	4	6

Table 9: overview final hyperparameters on the PERS/MISC task

B Detailed results

	AG_BERT	**Electra**	**GrεBerta**	**UGARIT**	**support**
B-PERS	0.87	0.87	0.77	0.87	3,090
I-PERS	0.58	0.50	0.05	0.56	83
B-LOC	0.75	0.73	0.58	0.75	1,105
I-LOC	0.17	0.08	0.00	0.13	196
B-GRP	0.82	0.81	0.68	0.84	1,384
I-GRP	0.00	0.00	0.00	0.00	49
macro_f1	0.53	0.50	0.35	0.53	

Table 10: overview detailed results test set

	AG_BERT	**Electra**	**GrεBerta**	**UGARIT**	**support**
B-PERS	0.84	0.82	0.78	0.85	578
I-PERS	0.04	0.00	0.08	0.07	51
B-LOC	0.77	0.73	0.62	0.68	233
I-LOC	0.22	0.14	0.00	0.31	11
B-GRP	0.78	0.78	0.73	0.76	201
macro_f1	0.53	0.50	0.44	0.53	

Table 11: overview detailed results GLAUx_test

C Entity conversion

PH		DEIPN		OD		SB	
original	converted	original	converted	original	converted	original	converted
person	PERS/GRP	person	PERS/GRP	person	PERS/GRP	PERSON	PERS/GRP
place	LOC	ethnic	GRP	place	LOC	LOC	LOC
place.proxy	GRP	place	LOC			PERS-G	GRP
artwork	O	group	GRP			LOC-G	GRP
event	O	title	O			TITLE	O/PERS/GRP
work	O	festival	O				
epithet	O	month	O				
tx	O	language	O				
material	O	constellation	O				
attribute	O						
movement	O						
measure	O						
animal	O						
object	O						
focalisation	O						
intervention	O						
transformation	O						

Table 12: entity conversion table