Development of robust NER Models and Named Entity Tagsets for Ancient Greek

Chiara Palladino* , **Tariq Yousef†**
*Furman University, USA, chiara.palladino@furman.edu
†University of Southern Denmark, Denmark, yousef@sdu.dk

Abstract

This contribution presents a novel approach to the development and evaluation of transformer-based models for Named Entity Recognition and Classification in Ancient Greek texts. We trained two models with annotated datasets by consolidating potentially ambiguous entity types under a harmonized tagset. Then, we tested their performance with out-of-domain texts, reproducing a real-world use case. Both models performed very well under these conditions, with the multilingual model *Ancient Greek Alignment* being slightly superior. In the conclusion, we emphasize current limitations due to the scarcity of high-quality annotated corpora and to the lack of cohesive annotation strategies for ancient languages.

Keywords: Token Classifications, Transformer Models, Named Entities Recognition, Ancient Greek, NLP

1. Introduction

Named Entity Recognition (NER) is a key task in text analysis and information extraction, which includes extracting, classifying, and disambiguating Named Entities (NEs) occurring in texts. The resulting outputs, which typically consist of datasets of classified names or annotated texts, provide important contextual information to facilitate interpretation of a source, and to enhance further explorations of it. Despite the current innovations in the application of transformer models to this task in ancient languages, Ancient Greek NER is still relatively unexplored. In this contribution, we illustrate a workflow to train a robust transformer-based NER in Ancient Greek with existing annotated texts. We ensured a state-of-the-art performance by mapping different entity types onto universal types, and performed a new type of evaluation with out-of-domain texts, reproducing a real-world scenario that provides a reliable assessment of the model's performance. In the conclusion, we present quantitative and qualitative results, and emphasize that current limitations are not due to scarce performance in available models, but to the lack of cohesive strategies for annotating and classifying entities in ancient languages.

2. Related Work

The introduction of Neural Networks and Deep Learning models has been a radical innovation in the computational processing of texts. Deep Learning was revolutionized by the introduction of transformers (Vaswani et al., 2017), which can capture contextual information to improve understanding of the data and retrieve that information from large contexts, and have become the state-of-the-art for extraction and classification tasks. In ancient languages, workflows based on popular transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Conneau et al., 2020) have been applied to a wide variety of tasks, such as POS tagging, authorship attribution, text alignment, automatic translation, and paleographic analysis (Sommerschield et al., 2023; Yousef et al., 2023c).

The task of NER, however, has remained relatively unexplored. In the case of Latin, LatinCy (Burns, 2023) and LatinBERT (Bamman and Burns, 2020) have been shown to outperform state-of-the-art Machine Learning methods of the previous generation when trained on the NER task (Beersmans et al., 2023). BERT-based models have also been applied to Medieval Latin corpora (Torres Aguilar, 2022) and Sumerian (Wang et al., 2022). Compared to Latin, NER in Ancient Greek is less well-resourced: Singh et al. 2021 developed a BERT-based monolingual language model trained on Ancient and Byzantine Greek that showed optimal performance on POS tagging for in-domain data, while Brennan Nicholson trained a BERT model to predict missing characters (Nicholson, 2020). Neither model, however, was trained on NER.

To overcome the lack of training data and language models for ancient languages, Yousef et al. 2023a developed an annotation projection pipeline based on the word level alignment to project NER annotation from the English translations to the original ancient Greek texts. Yousef et al. 2023b trained the first transformer-based model for NER in Ancient Greek, *Ancient Greek Alignment*. The model leveraged on an XLM-R-based multilingual model fine-tuned on the word-alignment task for Ancient Greek and other languages (Yousef et al., 2022a,b;

LT4HALA 2024@LREC-COLING-2024, pages 89–97

Yousef, 2023), and it was trained for NER using ad hoc annotated corpora, achieving an F1-score higher than 90% in training and through evaluation with in-domain texts. However, it showed a much lower performance with less represented categories, particularly place-names, and in the detection of multi-token entities. Furthermore, confusion in entity labeling and the use of different tagsets in the training data led to frequent errors of miscategorization in the output. In this contribution, we illustrate how we improved the training with additional annotated corpora and a generalized entity tagset. Moreover, we present a new strategy for model evaluation using an out-of-domain corpus: this provides a much clearer understanding of the actual performance of a model, and it more closely reproduces a real-world scenario.

3. Training Datasets and Tagset Harmonization

We trained the models on available annotated corpora in Ancient Greek. Out of 17 historical corpora surveyed by Ehrmann et al. 2024, only two are in ancient languages (Latin and Coptic, none in Ancient Greek). New Latin corpora have become available through the Corpus Burgundiae (Torres Aguilar et al., 2016) and the LASLA project (Beersmans et al., 2023), and in Sumerian (Bansal et al., 2021). There are only two sizeable annotated datasets in Ancient Greek, which are currently under release: the first one by Berti 2023, consists of a fully annotated text of Athenaeus' *Deipnosophists*, developed in the context of the Digital Athenaeus project [1]. The second one by Foka et al. 2020, is a fully annotated text of Pausanias' *Periegesis Hellados*, developed in the context of the Digital Periegesis project [2]. In addition, we used smaller corpora annotated by students and scholars on Recogito[3]: the *Odyssey* annotated by Kemp 2021; a mixed corpus including excerpts from the *Library* attributed to Apollodorus and from Strabo's *Geography*, annotated by Chiara Palladino; Book 1 of Xenophon's *Anabasis*, created by Thomas Visser; and Demosthenes' *Against Neaira*, created by Rachel Milio. Table 1 provides an overview of the datasets used in the training.

The main issue with annotated corpora is the lack of a cohesive tagset for the classification of named entities. There are no generalized guidelines to annotate Named Entities in ancient texts (Beersmans et al., 2023). Therefore, projects focusing on ancient names use custom tagsets and guidelines that are very specific to the corpus being annotated.

	Person	Location	NORP
Training Dataset			
Odyssey	2.469	698	0
Deipnosophists	14.921	2.699	5.110
Pausanias	10.205	8.670	4.972
Other Datasets	3.283	2.040	1.089
Total	**30.878**	**14.107**	**11.171**
Validation Dataset			
Xenophon	1.190	796	857

Table 1: An overview of the training and validation datasets. For convenience, we have grouped the smallest datasets together.

This problem is particularly crucial because the size of annotated corpora currently available is very small, and ambiguous entities tend to be treated in very different ways: models cannot be trained to optimal results if similar entities are tagged in completely different ways, especially if they belong to underrepresented categories. One of the biggest issues is the often arbitrary use of names of socio-ethnic groups, which are subject to metonymic readings (Poibeau, 2006) or used as proxies for physical locations: these cases are sometimes labeled as places, sometimes as "proxies", sometimes as groups or ethnics. Furthermore, there is no agreement on the classification of groups ("the Athenians") and indications of ethnicity ("Athenian"). Because these cases are strongly dependent on context and interpretation, they are one of the biggest sources of disagreement among annotators (Álvarez Mellado et al., 2021).

In this contribution, we are not proposing a new tagset for the annotation of Named Entities in Ancient Greek. Rather, we suggest a strategy to harmonize already available corpora through tag mapping. We mapped the tagsets used in each corpus onto a general set of entity types, following the model outlined by Burns 2023 for LatinCy, which is based on a simplified version of the OntoNotes v.5.0 release (Weischedel et al., 2013) [4]. The tagset includes the same tags used in LatinCy: PERson (people, including fictional), LOCation (which combined countries, cities and states with non-GPE locations, such as water bodies), and NORP (nationalities, religious, or political groups).

There are several reasons behind the choice of this general tagset. First of all, it ensures consistency with another model for an ancient language that has already been tested successfully for NER. Moreover, it allows more consistency by harmonizing project-specific labels, particularly in complicated cases such as ethnonyms and groups of people. Even though the OntoNotes release is

[1] https://www.digitalathenaeus.org/
[2] https://periegesis.org/
[3] https://recogito.pelagios.org/
[4] http://www.bbn.com/NLP/OntoNotes

based on English, the NORP tag is general enough to include both located groups of people and ethnonyms, but also political and religious organizations in the ancient world. Therefore, it can also be mapped onto a more traditional GRP tag, as proposed by Beersmans et al. 2023, who expanded upon the guidelines outlined by the Herodotos project (Erdmann et al., 2019). Romanello and Najem-Meyer 2022 do not consider located groups, but use the ORG tag for religious and military groups or modern organizations: while we did not encounter enough of these categories to address them specifically, they can be mapped onto our definition of NORP. Because of their intrinsic ambiguity, we decided to avoid context-dependent labeling for proxies (people-for-place: "the Spartans moved war to the Athenians") and methonymic readings (place-for-people: "Athens voted to expel Themistocles"): the former is treated as NORP being a located group, and the latter is tagged as it appears (LOC), without making inferences on its function. Table 4 provides the full list of concordances.

4. Models

We conducted various experiments using different combinations of training datasets and underlying transformer models. We utilized the *Ancient Greek BERT* model developed by Singh et al. 2021, (from now on, the "monolingual" model, or *Model_A*) [5] , and the *Ancient Greek Alignment* model (from now on, the "multilingual" model, or *Model_B*), an XLM-R-based multilingual model[6] fine-tuned on the word alignment task for ancient languages (Yousef et al., 2022a,b; Yousef, 2023). In Ex1 and Ex2, we utilized the Deipnosophists dataset with the monolingual and multilingual models, respectively. In Ex3, we utilized the Pausanias dataset with the monolingual model. In Ex4, we combined both datasets and used the monolingual model. In Ex5 and Ex6, we utilized all available datasets mentioned in Table 1 with the monolingual and multilingual models, respectively. In all experiments, we trained the models for 10 epochs, using 80% of the dataset for training and the remaining 20% for testing. Table 5 provides an overview of the training results.

After training, the models were evaluated with an out-of-domain corpus consisting of the first three books of Xenophon's *Hellenica*, annotated on Recogito by a domain expert. The tagset used in the annotation of Xenophon followed the same internal guidelines adopted by Chiara Palladino, Thomas Visser and Rachel Milio in the training phase. The tagset was subsequently mapped onto the general one, following the same strategy already applied to the rest of the training data. The complete dataset includes a total of 2843 annotated entities, with a larger number of PER entities and a similar quantity of LOC and NORP entitites, as shown in Table1. Table 6 reports the complete overview on the models performance on the validation datasets.

5. Results

Table 2 summarizes the performance of the two models on the test and validation datasets. In the validation stage, both models performed considerably well, showing that a robust training workflow with a tagset harmonization strategy leads to state-of-the-art results with out-of-domain texts, and confirming the reliability of both models on the NER task in a real-world scenario. In particular, the performance achieved with PER and NORP entities was very high in both cases, while for LOC entities it was generally lower. The multilingual model [7] performed better in almost all categories, with an overall F1 score of 93.32% and accuracy of 98.87% in validation and and F1 score of 89.41% and accuracy of 97.5% in training. Place names (LOC) are still the most challenging entity type, with the monolingual model[8] performing at 87.1% and the multilingual model at 88.8%.

The worse performance on LOC can be partly explained by their representation in the training data, as they correspond to about half of the personal names in our datasets. However, this does not explain the much better performance on NORP entities, which are even less represented in the training data, yet led to a high performance in the output. On the one hand, this shows the robustness of the NORP tag chosen for the evaluation, especially considering that ethnonyms and groups are one of the most challenging entity classes for automatic extraction. On the other hand, it suggests that place names need a more careful treatment at the stage of annotation and guidelines design. Both models are now available on HuggingFace

5.1. Qualitative Evaluation

For the qualitative evaluation, we utilized the multilingual model (*Model_B*.) Overall, the multilingual model correctly classified 1118 PER entities, 698 LOC entities, and 809 NORP entities, for a total of 2625 entities (Table 3). It missed 78 entities, and it miscategorized 134 entities in total, with LOC being by far the most frequent. The most frequent errors

[5] https://huggingface.co/pranaydeeps/Ancient-Greek-BERT

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

[7] https://huggingface.co/UGARIT/grc-ner-xlmr

[8] https://huggingface.co/UGARIT/grc-ner-bert

		Test		Validation	
		Model_A (Ex 5)	Model_B (Ex 6)	Model_A (Ex 5)	Model_B (Ex 6)
LOC	precision	82.92%	83.33%	87.10%	**88.66%**
	recall	81.30%	81.27%	87.10%	**88.94%**
	f1	82.11%	82.29%	87.10%	88.80%
NORP	precision	87.10%	88.71%	92.82%	**94.76%**
	recall	90.81%	90.76%	93.42%	**94.50%**
	f1	88.92%	89.73%	93.12%	**94.63%**
PER	precision	92.61%	91.72%	**95.52%**	94.22%
	recall	92.94%	94.42%	95.21%	**96.06%**
	f1	92.77%	93.05%	**95.37%**	95.13%
Overall	precision	88.92%	88.83%	92.63%	**92.91%**
	recall	88.82%	89.99%	92.79%	**93.72%**
	f1	88.87%	89.41%	92.71%	**93.32%**
	accuracy	97.28%	97.50%	98.42%	**98.87%**

Table 2: Test and validation results of the top two models. Model_A represents the output of Experiment 5, a fine-tuned model based on the ancient Greek monolingual model (Singh et al., 2021), while Model_B represents the output of Experiment 6, a fine-tuned model based on the Ugarit multilingual model (Yousef et al., 2022a).

of classification concerned confusion between the LOC and NORP tags, as it is to be expected. Very rarely confusion occurred between PER and other tags, often being justified by ambiguity in the very lemma of the word or by the presence of foreign names, such as "Mania", which was misclassified as LOC. Entities that were not extracted included some recurring names, such as "Phyle" (8 times) and "Otys" (6 times). The ethnonym "Hellenikon" was not extracted 5 times. There was also a minority of cases where the model correctly identified entities that had been mistakenly omitted by the annotator, which leads us to believe that the results are even better than what the numbers suggest.

Overall, LOC names were most frequently involved in errors of extraction and miscategorization. Interestingly, however, some common nouns were extracted and correctly classified, that could be considered places, such as "doors", "islands", "isthmus", "river", and "acropolis". This presumably reflects the ways in which entity boundaries are established in the training data, where strings like "Phasis river" or "Ionic gulf" are often considered full names, even if the second word is lowercase. However, it is also true that common nouns like "isthmus" are often used in Greek sources to refer to specific places, such as the Isthmus of Corinth: therefore, it is difficult to establish what exactly constitutes an identifiable "place" in these cases. A similar phenomenon occurred with titles, such as "hipparchos" (which can also be a personal name) or "ephoros", and with socio-political organizations, such as "boule" or "demos". It should be noted, however, that these strings were not consistently extracted: this is presumably due to the internal inconsistency of the training data, where analogous instances may or may not have been annotated, whether as names, as part of multi-token entities, or as mentions of specific referents, depending on the project guidelines.

A related issue is represented by multi-token entities, such as "Olympian Zeus" or "Temple of Artemis": these are often not represented in sufficient number to be significant for training and evaluation, and are extremely difficult to annotate, because their boundaries are not always clear. They are also challenging to measure in quantitative evaluation. In our dataset, there were 15 recognizable multi-token entities, of which the model extracted and classified 9 in a coherent way, while 5 were not recognized, and one was dubious. In most cases, even if the entity extracted did not perfectly overlap with the gold standard, it made sense: for example, "Lyceum gymnasium" was counted as an error because the annotator only tagged "Lyceum", but it is a perfectly acceptable alternative name. A remarkable case regarded the "Makra Teiche" (the Long Walls of Athens), which appears lowercase in our text, but was extracted and classified by the model. In other words, there are cases that need to be considered individually and qualitatively in order to be properly assessed, as they often require strict guidelines to establish entity boundaries.

6. Conclusion and Limitations

In this paper, we have shown a workflow to train a robust NER model, whose performance is evaluated on out-of-domain texts, reproducing a realistic scenario of use for a tool of this kind. Our training strategy and tagset harmonization lead to state-of-the-art performance with the two available

		Model Output			
		O	PER	LOC	NORP
Gold St.	O	26,226	33	37	14
	PER	37	1,118	22	8
	LOC	33	29	698	35
	NORP	8	2	38	809

Table 3: Qualitative Evaluation Confusion (Error) Matrix. "O" represents non-entity tokens.

transformer-based models, with a slightly better performance shown in the multilingual model trained on the alignment task.

Despite the encouraging results, the potential of transformers for NER in Ancient Greek is still not fully exploited. It has been shown that even the most refined models, without ad hoc training and fine-tuning, perform poorly on several tasks on ancient and historical corpora (Sprugnoli et al., 2023; González-Gallardo et al., 2023). Transformers are very data-hungry and require a significant amount of annotations for optimal results. This is especially relevant for ancient languages, which are closed systems and, for the most part, significantly smaller corpora than modern languages. This fundamentally limits strategies for upsampling, training and fine-tuning.

Apart from the scattered nature of currently available tagsets, some issues remain unresolved. For example, place names are still underrepresented in annotated corpora. However, data availability is insufficient to explain bad model performance, as we have shown above. In general, place names seem to be especially challenging for annotation practices, more than personal and group names. For example, the definition of identifiable "places" sometimes goes beyond capitalized words; furthermore, it may be relevant for a project to tag common nouns that refer to locatable areas. Another issue is represented by multi-token entities, such as "Pythian Apollo" or "Erythraean Sea". In our dataset, we had too few of them to be significant to the evaluation. However, the problem resides once again in annotation practices, as it is often difficult to establish the boundaries of what constitutes a named entity.

In conclusion, we want to emphasize that current challenges in model training and evaluation are not to be attributed to the lack of highly performing models, but to the lack of best practices and documentation in the development of high-quality annotated datasets (Beersmans et al., 2023). This key issue affects the further development of annotation strategies and reliable tagsets: in fact, our mapping strategy was effective in containing potentially ambiguous cases, but it also limited the granularity of entity classification. For example, author names, nicknames and personal names are all grouped under one PER tag, but their different functions could be significant in the context of individual projects. Furthermore, other entity classes were not considered, such as events, objects, and languages. The future necessary steps include the implementation of an extended tagset according to a hierarchical structure, as outlined by Romanello and Najem-Meyer 2022: the hierarchical structure will ensure that existing tagsets can still be harmonized at least at the higher level, but it will also provide a foundation for more accurate annotated corpora in the future.

Acknowledgments

We are deeply grateful to the people who contributed annotations and datasets to make this project possible. In no particular order: Monica Berti (Digital Athenaeus Project, University of Leipzig), Josh Kemp (Odyssey Project, Furman University), Thomas Visser and Rachel Milio (University of Exeter), and the whole team of the Digital Periegesis project: Elton Barker, Anna Foka, Brady Kiesling, Kyriaki Konstantinidou, Linda Talatas.

7. Bibliographical References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. ArXiv:2009.10053 [cs].

Rachit Bansal, Himanshu Choudhary, Ravneet Punia, Niko Schenk, Émilie Pagé-Perron, and Jacob Dahl. 2021. How Low is Too Low? A Computational Perspective on Extremely Low-Resource Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 44–59, Online. Association for Computational Linguistics.

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. 2023. Named Entity Recognition for a Text-Based Catalog of Ancient Greek Authors and Works. *Zenodo (CERN European Organization for Nuclear Research)*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mª Luisa Díez Platas, Salvador Ros Muñoz, Elena González-Blanco, Pablo Ruiz Fabo, and Elena Álvarez Mellado. 2021. Medieval Spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. *Journal of the Association for Information Science and Technology*, 72(2):224–238.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2024. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. 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, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Bèatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation. Original-date: 2017-10-22T06:51:43Z.

Anna Foka, Elton Barker, Kyriaki Konstantinidou, Nasrin Mostofian, O. Cenk Demiroglu, Brady Kiesling, and Linda Talatas. 2020. Semantically geo-annotating an ancient greek "travel guide" itineraries, chronotopes, networks, and linked data. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities '20, page 1–9, New York, NY, USA. Association for Computing Machinery.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G. Moreno, and Antoine Doucet. 2023. Yes but.. Can ChatGPT Identify Entities in Historical Documents? ArXiv:2303.17322 [cs].

Joshua Kemp. 2021. Beyond Translation: Building Better Greek Scholars. *Pelagios Blog*.

Brennan Nicholson. 2020. Ancient-greek-char-bert.

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

Thierry Poibeau. 2006. Dealing with Metonymic Readings of Named Entities. In *The 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*, pages 1962–1968, Vancouver, Canada. Cognitive Science Society.

Matteo Romanello and Sven Najem-Meyer. 2022. Guidelines for the Annotation of Named Entities in the Domain of Classics. Publisher: Zenodo.

Aleksi Sahala. 2021. *Contributions to Computational Assyriology*. Doctoral Thesis, Faculty of Arts, University of Helsinki, Helsinki.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings)*.

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*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, pages 703–747.

Rachele Sprugnoli, Francesco Mambrini, Marco Passarotti, and Giovanni Moretti. 2023. The Sentiment of Latin Poetry. Annotation and Automatic

Analysis of the Odes of Horace. *IJCoL. Italian Journal of Computational Linguistics*, 9(1). Number: 1 Publisher: Accademia University Press.

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*, pages 119–128, Marseille, France. European Language Resources Association.

Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. 2016. Named entity recognition applied on a data base of Medieval Latin charters. The case of chartae burgundiae. In *3rd International Workshop on Computational History (HistoInformatics 2016)*, Krakow, Poland.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Guanghai Wang, Yudong Liu, and James Hearne. 2022. Few-shot Learning for Sumerian Named Entity Recognition. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 136–145, Hybrid. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Ramshaw Lance, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0 LDC2013T19.

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

Tariq Yousef. 2023. *Translation Alignment Applied to Historical Languages: Methods, Evaluation, Applications, and Visualization*. Ph.D. thesis, Leipzig University.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. 2023a. 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. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2023b. 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, and Farnoosh Shamsian. 2023c. Classical Philology in the Time of AI: Exploring the Potential of Parallel Corpora in Ancient Language. In *Proceedings of the Ancient Language Processing Workshop*, pages 179–192, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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 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.

Amir Zeldes and Lance Martin. 2020. Coptic Scriptorium - Entity Annotation Guidelines. Version 1.1.0.

Elena Álvarez Mellado, María Luisa Díez-Platas, Pablo Ruiz-Fabo, Helena Bermúdez, Salvador Ros, and Elena González-Blanco. 2021. TEI-friendly annotation scheme for medieval named entities: a case on a Spanish medieval corpus. *Language Resources and Evaluation*, 55(2):525–549.

8. Appendix

Corpus	Original Tag	OntoNotes Tag
Deipnosophists	Person	PER
	Place	LOC
	Ethnic	NORP
	Group	NORP
	Noclass	MISC
	title	MISC
	festival	MISC
	month	MISC
	language	MISC
	constellation	MISC
Pausanias	Place.proxy	NORP
	Place.regional	LOC
	Place.physical	LOC
	Place.mythical	LOC
	Place.material	LOC
	Person	PER
Other	Place	LOC
	Place.group	NORP
	Ethnonym	NORP
	Person	PER
	Person.group	NORP
	Author	PER
	Patronymic	PER

Table 4: Concordance table used to harmonize the main tagsets used in the training data.

		Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6
LOC	precision	87.10%	86.11%	77.78%	80.54%	82.92%	83.33%
	recall	87.31%	88.59%	78.78%	80.23%	81.30%	81.27%
	f1	87.21%	87.33%	78.28%	80.39%	82.11%	82.29%
NORP	precision	93.35%	93.68%	89.51%	90.76%	87.10%	88.71%
	recall	92.32%	95.55%	92.08%	92.48%	90.81%	90.76%
	f1	92.83%	94.60%	90.78%	91.61%	88.92%	89.73%
PER	precision	94.10%	95.18%	88.78%	92.05%	92.61%	91.72%
	recall	95.67%	97.20%	88.62%	92.34%	92.94%	94.42%
	f1	94.88%	96.18%	88.70%	92.19%	92.77%	93.05%
overall	precision	91.55%	92.86%	85.36%	88.45%	88.92%	88.83%
	recall	91.74%	94.73%	86.17%	88.90%	88.82%	89.99%
	f1	91.64%	93.79%	85.76%	88.67%	88.87%	89.41%
	accuracy	98.21%	98.93%	95.55%	97.22%	97.28%	97.50%

Table 5: Training results of all experiments.

		Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6
LOC	precision	86.15%	**89.69%**	81.91%	86.76%	87.10%	88.66%
	recall	75.35%	75.89%	**91.02%**	85.22%	87.10%	88.94%
	f1	80.39%	82.22%	86.22%	85.99%	87.10%	**88.80**%
NORP	precision	89.53%	90.42%	94.13%	92.26%	92.82%	**94.76**%
	recall	88.00%	**94.61%**	91.01%	91.52%	93.42%	94.50%
	f1	88.76%	92.46%	92.55%	91.89%	93.12%	**94.63%**
PER	precision	93.73%	92.28%	90.84%	**95.83%**	95.52%	94.22%
	recall	86.74%	91.63%	95.79%	93.75%	95.21%	**96.06%**
	f1	90.10%	91.95%	93.25%	94.78%	**95.37%**	95.13%
overall	precision	88.14%	89.44%	89.30%	91.62%	92.63%	**92.91%**
	recall	84.29%	88.57%	93.42%	91.11%	92.79%	**93.72%**
	f1	86.17%	89.00%	91.32%	91.37%	92.71%	**93.32%**
	accuracy	97.09%	98.11%	97.88%	98.18%	98.42%	**98.87%**

Table 6: Performance of different models on the validation dataset.