SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek

Konstantin Schulz and **Florian Deichsler**
Humboldt University Berlin
konstantin.schulz@hu-berlin.de

Abstract

Literary scholars of Latin and Ancient Greek increasingly use natural language processing for their work, but many models and datasets are hard to use due to a lack of sustainable research data management. This paper introduces the Systematic Evaluation Framework for natural language processing models and datasets in Latin and Ancient Greek (SEFLAG), which consistently assesses language resources using common criteria, such as specific evaluation metrics, metadata and risk analysis. The framework, a work in progress in its initial phase, currently covers lemmatization and named entity recognition for both languages, with plans for adding dependency parsing and other tasks. For increased transparency and sustainability, a thorough documentation is included as well as an integration into the HuggingFace ecosystem. The combination of these efforts is designed to support researchers in their search for suitable models.

1 Introduction

Recent years have seen a surge of publications employing natural language processing (NLP) for the analysis of ancient texts (Papantoniou and Tzitzikas, 2020; Ehrmann et al., 2021; Sommerschield et al., 2023). However, as with other historical languages (see Zhou et al. (2023) for Classical Chinese), the communities around Latin and Ancient Greek rarely provide standardized and centralized resources specifically for the training and evaluation of NLP models. The corresponding treebanks in Universal Dependencies are one notable exception. Lemmatization, on the other hand, is notorious for the many different approaches to, e.g., character encoding (Tauber, 2019), handling of diacritics (Kostkan et al., 2023), homographs (Mambrini and Passarotti, 2019), and other challenges. Besides, existing NLP models are scattered across many different technical platforms such as spaCy (Burns, 2023), Flair NLP (Yousef et al., 2023) or Google Cloud (Bamman and Burns, 2020). As a consequence, every member of the Classics community has to collect and evaluate the same resources again. This leads us to the central question of how we can support literary scholars of Latin and Ancient Greek in choosing the right NLP models for their research agenda.

To address this challenge, we present SEFLAG[1], the **S**ystematic **E**valuation **F**ramework for NLP models and datasets in **L**atin and **A**ncient **G**reek. Our work is still in progress, so we share only a small proof of concept with lemmatization and named entity recognition (NER). Next up, dependency parsing will follow. Our contributions are as follows:

- We collect and document existing datasets and NLP models, using recently established standards such as datasheets (Gebru et al., 2021) and model cards (Mitchell et al., 2019).
- We create benchmarks from suitable datasets, use consistent metrics for comparing models' performance on them and publish results in the Hugging Face[2] (HF) ecosystem.
- We document and publish conceptual mappings for connecting specific NLP models and datasets that were originally built using different annotation guidelines.

2 Related Work

Building highly specialized frameworks like SEFLAG can suffer from various problems. For example, transferring modern developments (large language models, analytical categories) to ancient contexts is non-trivial (McGillivray, 2013; Singh et al., 2021; Ehrmann et al., 2021; Riemenschneider and Frank, 2023; Yousef et al., 2023). In particular, the loose distinction between different seg-

[1] https://github.com/daidalos-project/seflag
[2] https://huggingface.co/

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 247–258
November 16, 2024.

ments of a text in Vedic Sanskrit (Biagetti et al., 2021) and other historical languages necessitates elaborate interpretative efforts to introduce modern syntactic concepts like punctuation or main and subordinate clause. This relatively relaxed notion of syntax correlates with the rather pronounced linguistic variation of Latin and Ancient Greek due to their diachronic, diatopic and diastratic differences (Kostkan et al., 2023).

Similar issues arise in the treatment of historical newspapers (Ehrmann et al., 2020) and early modern scientific texts (Odebrecht et al., 2017), which indicates a general trend of higher linguistic variation and lesser availability of language resources for historical languages (Etxeberria et al., 2016). For ancient texts, there is the additional burden of manifold textual transmission (including indirect transmission through citations), which prevents us from establishing texts in their original form with certainty (Sommerschield et al., 2023).

Besides, existing NLP implementations for textual annotation often do not fully adhere to the FAIR (Wilkinson et al., 2016) guiding principles of research data management. Earlier evaluations of available resources, on the other hand, were often performed indirectly, e.g., by carrying out surveys in the user community (Monachini et al., 2018) rather than directly testing the resources. Finally, large-scale data processing necessitates automation due to its efficiency, but automation can lead to a loss of data quality in highly heterogeneous datasets (Passarotti and Mambrini, 2022) and is often not sufficient for unifying multiple conceptually different resources, e.g., for valency patterns in ancient languages (Luraghi et al., 2024).

Fortunately, there has also been some progress in NLP for Latin and Ancient Greek. Most researchers involved in NLP evaluation choose one of the two languages and a single NLP task, like word embeddings (Stopponi et al., 2023) or topic modeling (Martinelli et al., 2024). A few of them even work on both languages, usually for single tasks that can be addressed through inherently cross-lingual methods (Perrone et al., 2021). Some engineering work has been done by the CLTK team (Johnson et al., 2021), offering a solid basis for data processing in historical languages but still suffering from a lack of evaluation and benchmarking. Nevertheless, all of these approaches are an important foundation for our mission of collecting and disseminating such resources in a centralized manner.

A more coordinated and comprehensive initiative was the LiLa project[3], which managed to successfully collect, harmonize and disseminate multiple existing language resources for Latin (Mambrini et al., 2020). In particular, they launched the EvaLatin evaluation campaign (Sprugnoli et al., 2020). Unfortunately, their project has officially ended and their platform (which is still running) does not cover Ancient Greek at all. Even for Latin, it addresses many NLP tasks, but not all: NER, topic modeling and some others are missing.

Other platforms have partly solved the problem of long-term availability and funding, such as the Perseus Digital Library. Like LiLa, they use Linked Data (Almas et al., 2014) to make their content findable and interoperable, but do not support evaluation reports. Generic language platforms like HF offer such reporting, but cater to a different audience (namely computer scientists and computational linguists), thus neglecting our target group of literary scholars.

3 Methodology

Our intermediate goal is to find existing NLP models and apply them to existing datasets (both having been created by others). Then, we perform one measurement for each of the two languages: Predictions of the NLP model are assessed using the ground truth annotations from the dataset and consistent metrics. For enhanced transparency and reusability, we document conceptual mappings, as most datasets and NLP models were created using rather different annotation schemas.[4]

For example, names of fictional characters in ancient literature may count as *PERSON* names in one NER dataset, but not in others (see section 4). This results in many datasets and NLP models that belong to the same task (namely, NER), but are not easily interoperable. Conceptual bridging (e.g., through mappings) is needed to close this gap and enable combinations of those resources. Even

[3] https://lila-erc.eu

[4] The problem is well-known especially in the treebanking community, where the de-facto standard of Universal Dependencies has been the most prominent effort to harmonize various other existing traditions such as the Index Thomisticus Treebank (Cecchini et al., 2018) or the Latin Dependency Treebank (Bamman and Crane, 2011). Such issues are particularly pressing for low-resource languages like Latin and Ancient Greek, where data sparsity hampers the development and application of various NLP technologies (McGillivray, 2013). In these languages, pushing the boundaries of existing resources by making them interoperable is especially important.

language resources that do not share the same annotation schema may still profit from the unification of certain annotation labels (such as *PERSON* and *PRS*, see section 4), depending on their conceptual overlap.

For enhanced sustainability, in applicable cases like lemmatization, we merge multiple datasets (i.e., various treebanks from Universal Dependencies) into a larger benchmark and publish it as a HF Dataset[5] for the corresponding task. In doing so, we adopt the approach of Sprugnoli et al. (2020) by integrating diverse language material with regard to time and genre.

Additionally, we use datasheets and model cards (see Appendix A and B) to describe language resources systematically. Ideally, such datasheets and model cards should be provided by the creators themselves. However, the adoption of those standards is still insufficient in the Classics community. As the next best option, we create such materials ourselves and try to infer their content from publicly available information about the resources (in scientific publications, source code repositories, etc.). They will be uploaded to the HF Space[6] of an NLP infrastructure (see section 4) and integrated into their website. This infrastructure allows our users to directly apply the evaluated NLP models to their own datasets and learn more about the various tasks through open educational resources.

Model cards are provided separately for each NLP model in our evaluation. They include general metadata like license, version or architecture, but also more complex considerations like ethical implications, ecological factors and possible risks of certain use cases. Apart from literary scholars of Latin and Ancient Greek, we also take neighboring disciplines into account, such as historians, theologians or archeologists dealing with ancient textual materials. From our point of view, their shared characteristics are limited technical background knowledge (Caraher, 2020) and a high interest in practical applications as well as methodological innovation, though all of these aspects are somewhat disputed in the scientific literature (Buchanan, 2015; Mahony, 2016; Damer, 2023).

[5] https://huggingface.co/datasets/daidalos-project/latin_treebanks_ud_test

[6] https://huggingface.co/daidalos-project

4 Implementation

For long-term sustainability, we aim to integrate our work into the Daidalos research infrastructure[7] with institutionalized governance as provided by the datacenter[8] of Humboldt University Berlin[9], which offers a dedicated cloud computing service (as recommended by Almas (2017)). Funding for such an infrastructure has already been secured for an initial period of 3 years, which can be extended to about 10 years depending on periodic evaluation. As a consequence of our integration into that infrastructure, we also build on their community work: Their already existing biannual workshops, national research partnerships with classical scholars and open educational resources on Historical Language Processing[10] are the backbone of our strategy to interact with our target audience and disseminate the evaluation results as widely as possible.

We provide explicit mappings for two cases[11]: NLP models that are evaluated internally (on the test split of their original training data), and externally (i.e., a completely new dataset). The internal case is covered by the *flair_grc_multi_ner* tagger being tested on the data[12] that was curated by Yousef et al., i.e., a mixture of Herodotus, Homer and Athenaeus of Naucratis. Under those circumstances, no mapping is needed at all because the model was conceptualized directly with that dataset in mind.

Mapping the external case is more challenging: We took the LatinCy model and applied it to the Herodotos Project[13] dataset. Each of the two resources uses 4 different entity tags that roughly correspond to the original ones introduced in Grishman and Sundheim (1996). The *PERSON* and *PRS* classes are arguably most compatible. However, the annotation guidelines for neither of the two

[7] https://daidalos-projekt.de

[8] https://www.cms.hu-berlin.de/en/

[9] https://www.hu-berlin.de/en

[10] https://daidalos-projekt.de/dokumentation

[11] See https://github.com/daidalos-project/seflag/blob/main/mappings.yaml. To the best of our knowledge, there is no existing best practice for documenting linguistic annotation mappings. In particular, different conversion software like Pepper (Zipser and Romary, 2010) or Grew (Guillaume, 2021) uses different data formats for serializing the respective conversion instructions.

[12] https://github.com/daidalos-project/seflag/blob/main/documentation/datasheets/yousef_et_al_dataset.md

[13] https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation

language resources have been published anywhere. Thus, we cannot say for sure if their rules for assigning labels to named persons match each other, even considering the vague statement in Burns (2023) that the label applies to "people, including fictional". As a last resort, since the Herodotos dataset was included as training data for the LatinCy model, we may conclude that all 3 entity classes roughly correspond to each other, which allows us to apply mappings and pair the two for evaluation.

5 Results

We report the evaluation results in Table 1. We choose macro F1 and accuracy because they are applicable to many use cases and are widely adopted in the scientific community[14]. Moreover, macro F1 can be indicative of certain characteristics of language resources (Bone et al., 2015) such as the balance of the data distribution. We believe that it is part of our mission to inform potential users about weaknesses in a dataset, such as the strong class imbalance in both NER datasets (see Appendix B): Since most words in a text are non-entities, it is easy to achieve high accuracy by always guessing 'non-entity' as a baseline. This also explains the comparatively low scores in our NER evaluation, where non-entities are treated as rather unimportant. Furthermore, we will publicly upload our results to the HF Hub, so many others can benefit from the insights and do not have to run the evaluations themselves. This saves time and resources for the research community, while also providing easier access to necessary information about language resources.

A qualitative analysis of the lemmatization results empirically reveals some of the problems that were outlined in the research literature (see section 1): Variant spellings of the same lemma exist due to flexible orthography (οὕτως or οὕτω(ς), *parvulus* or *paruulus*), capitalization (*Romanus* or *romanus*), diacritics (τίς, τις or τὶς), and separate entries for specific inflected forms (κύκλος or κύκλῳ, *diuerto* or *diuersus*).

6 Limitations and Risks

Our approach of curating datasheets and model cards for resources that we did not create ourselves leads to information gaps in the documentation. Nevertheless, our effort of inferring information from other sources and disseminating it in a centralized, systematic fashion is highly beneficial for the targeted research community.

Currently, we only report rather simple metrics. To enable deeper insights into model bevavior and dataset structure, we plan to add class-wise confusion matrices, detailed qualitative error analyses and task-specific metrics (like the ones introduced by the Message Understanding Conference (Nadeau and Sekine, 2007)) which allow to distinguish between errors related to entity status and entity type.

Finally, we are very confident that our framework scales well to other planned tasks like part-of-speech tagging, sentiment analysis and dependency parsing.[15] As an abstract representation of the different linguistic annotations, for example, we intend to use a graph model like SALT (Zipser and Romary, 2010). Besides, we have clear interfaces for adding more models, datasets and evaluation metrics. As of now, however, it is unclear to which extent we may need to introduce further metadata (spatial, temporal, stylistic, etc.).

7 Conclusions

Our evaluation framework SEFLAG aims to support literary scholars of Latin and Ancient Greek in selecting the right NLP models for their research. We provide quantitative evaluations of existing models on suitable datasets. Conceptual mappings between tagsets used for the annotation of different language resources are documented explicitly and in a human-readable way. Evaluation results are reported using common metrics (F1, accuracy) and are accompanied by additional documentation for the language resources: datasheets for datasets and model cards for NLP models. Using that additional information, we enable researchers to critically assess the value of such resources for their own research, including questions of dataset characteristics, model architecture, annotation guidelines and contact persons. In short, we provide low-level access to the costly and complex task of NLP evaluation for Latin and Ancient Greek, with a proof of concept that focuses on NER and lemmatization.

In the near future, we will work on fully integrating our framework into the Daidalos research

[14] For examples from the Classics, see Bizzoni et al. (2014); Stoeckel et al. (2020); Köntges (2020).

[15] These tasks have been chosen because they are of general interest to the research community (Berti, 2019; Ehrmann et al., 2021; Beersmans et al., 2023) and are directly relevant to the Daidalos research infrastructure.

Language	Task	Model	Dataset	Metric	Score
Latin	NER	LatinCy	Herodotos	macro F1 ↑	58
	lemmatization	LatinCy	UD Latin	accuracy ↑	88
Ancient Greek	NER	flair_grc_bert_ner	Yousef et al.	macro F1 ↑	64
	lemmatization	greCy	UD Ancient Greek	accuracy ↑	89

Table 1: Evaluation results for NER and lemmatization in Latin and Ancient Greek. The metrics used macro F1 and accuracy. 3 different NLP models have been evaluated on 4 different datasets. Upward arrows for a metric indicate that higher scores are better.

infrastructure. Furthermore, we would like to add more NLP tasks, models and datasets. Finally, we also want to create Open Educational Resources to educate interested researchers about central essentials of the evaluation, such as specific metrics, task concepts and annotation approaches.

Ethics Statement

We address ethical considerations mainly through heavy use of model cards and datasheets. Besides, we respect licensing conditions for datasets by publishing our benchmarks only if all contained sub-datasets allow it from a legal perspective, and only under a license that matches the ones used in the sub-datasets. In general, we refrain from reusing datasets with licenses that are too prohibitive.

Acknowledgements

We are grateful to five anonymous reviewers and Andrea Beyer as well as Anke Lüdeling for commenting on earlier drafts of this paper.

This work is part of a project funded by the German Research Foundation (project number 518919950) and led by Andrea Beyer, Malte Dreyer and Anke Lüdeling.

References

Bridget Almas. 2017. Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities. *Data Science Journal*, 16(19):1–17.

Bridget Almas, Alison Babeu, and Anna Krohn. 2014. Linked Data in the Perseus Digital Library. *ISAW Papers*, 7(3).

David Bamman and Patrick J Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. *arXiv preprint arXiv:2009.10053*, pages 1–14.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin Dependency Treebank. In *Language Technology for Cultural Heritage*, pages 79–98. Springer.

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 *Ancient Language Processing Workshop*, pages 1–12.

Monica Berti. 2019. Named entity annotation for ancient greek with inception. In *Proceedings of CLARIN Annual Conference*, pages 1–4.

Erica Biagetti, Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2021. Evaluating Syntactic Annotation of Ancient Languages: Lessons from the Vedic Treebank. *Old World: Journal of Ancient Africa and Eurasia*, 1(1):1–32.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory R Crane. 2014. The Making of Ancient Greek WordNet. In *LREC*, volume 2014, pages 1140–1147.

Daniel Bone, Matthew S Goodwin, Matthew P Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. 2015. Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of autism and developmental disorders*, 45(5):1121–1136.

Sarah A Buchanan. 2015. The Emerging Tradition of Digital Classics. *Annual Review of Cultural Heritage Informatics: 2014*, page 149.

Patrick J Burns. 2023. LatinCy: Synthetic Trained Pipelines for Latin NLP. *arXiv preprint arXiv:2305.04365*.

William Caraher. 2020. Dissecting Digital Divides in Teaching. In Sebastian Heath, editor, *DATAM: Digital Approaches to Teaching the Ancient Mediterranean*, number 16 in Digital Press Books, pages 71–82. The Digital Press at the University of North Dakota.

Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36.

Erika Zimmermann Damer. 2023. What Is a Future for Classics? *American Book Review*, 44(3):47–50.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: The Impresso Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968.

Izaskun Etxeberria, Inaki Alegria, Larraitz Uria, and Mans Hulden. 2016. Evaluating the noisy channel model for the normalization of historical texts: Basque, Spanish and Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1064–1069.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12):86–92.

Ralph Grishman and Beth M Sundheim. 1996. Design of the MUC-6 evaluation. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop Held at Vienna, Virginia, May 6-8, 1996*, pages 413–422.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Thomas Köntges. 2020. Measuring philosophy in the first thousand years of Greek literature. *Digital Classics Online*, pages 1–23.

Jan Kostkan, Márton Kardos, Jacob Palle Bliddal Mortensen, and Kristoffer Laigaard Nielbo. 2023. OdyCy–A general-purpose NLP pipeline for Ancient Greek. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–134.

Silvia Luraghi, Alessio Palmero Aprosio, Chiara Zanchi, and Martina Giuliani. 2024. Introducing PaVeDa–Pavia Verbs Database: Valency Patterns and Pattern Comparison in Ancient Indo-European Languages. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 79–88.

Simon Mahony. 2016. Open Education and Open Educational Resources for the Teaching of Classics in the UK. In Matteo Romanello and Gabriel Bodard, editors, *Digital Classics Outside the Echo-Chamber*, pages 33–50. Ubiquity Press.

Francesco Mambrini, Flavio Massimo Cecchini, Greta Franzini, Eleonora Litta, Marco Carlo Passarotti, and Paolo Ruffolo. 2020. LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web. *Umanistica Digitale*, 4(8):63–78.

Francesco Mambrini and Marco Passarotti. 2019. Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 71–80.

Ginevra Martinelli, Paola Impicciché, Elisabetta Fersini, Francesco Mambrini, and Marco Passarotti. 2024. Exploring Neural Topic Modeling on a Classical Latin Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6929–6934.

Barbara McGillivray. 2013. *Methods in Latin Computational Linguistics*. Brill, Leiden; Boston.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Monica Monachini, Anika Nicolosi, and Alberto Stefanini. 2018. Digital Classics: A Survey on the Needs of Ancient Greek Scholars in Italy. In *Proceedings of the CLARIN 2017 Conference*, pages 1–5. Linköping University Electronic Press.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Carolin Odebrecht, Malte Belz, Amir Zeldes, Anke Lüdeling, and Thomas Krause. 2017. RIDGES Herbology: Designing a diachronic multi-layer corpus. *Language Resources and Evaluation*, 51(3):695–725.

Katerina Papantoniou and Yannis Tzitzikas. 2020. NLP for the Greek language: A brief survey. In *11th Hellenic Conference on Artificial Intelligence*, pages 101–109.

Marco Passarotti and Francesco Mambrini. 2022. Issues in Building the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *LLOD Approaches for Language Data Research and Management LLODREAM2022 : International Scientific Interdisciplinary Conference*.

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2021. Lexical semantic change for

Ancient Greek and Latin. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, number 6 in Language Variation, pages 287–310. Language Science Press.

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, pages 15181–15199.

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*, 49(3):703–747.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Manuel Stoeckel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. 2020. Voting for POS tagging of Latin texts: Using the flair of FLAIR to better ensemble classifiers by example of Latin. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135.

Silvia Stopponi, Nilo Pedrazzini, Saskia Peels-Matthey, Barbara McGillivray, and Malvina Nissim. 2023. Evaluation of Distributional Semantic Models of Ancient Greek: Ancient Language Processing. *Proceedings of the Ancient Language Processing Workshop*, pages 49–58.

James K Tauber. 2019. Character encoding of classical languages. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 10:137–158.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(160018):1–9.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2023. Transformer-Based Named Entity Recognition for Ancient Greek. In *Book of Abstracts*, pages 420–422, Graz. Zenodo.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP Evaluation Benchmark For Classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319.

Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*.

A Model Card: Latincy

la_core_web_lg

- Person or organization developing model: Patrick J. Burns; with Nora Bernhardt [ner], Tim Geelhaar [tagger, morphologizer, parser, ner], Vincent Koch [ner]
- Model date: May 2023
- Model version: 3.7.4
- Model type: spaCy
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features: For information on the training workflow see p.4-5 of LatinCy: Synthetic Trained Pipelines for Latin NLP (https://arxiv.org/pdf/2305.04365v1)
- Paper or other resource for more information: **Burns, P.J. 2023. "LatinCy: Synthetic Trained Pipelines for Latin NLP." arXiv:2305.04365 [cs.CL]. http://arxiv.org/abs/2305.04365.
- License: MIT

- Where to send questions or comments about the model: https://diyclassics.github.io/
- Intended Use
  - Primary intended uses: Morphological analysis, POS-Tagging, Lemmatizing, Parsing, NER
  - Primary intended users: Classical Scholars
  - Out-of-scope use cases: unknown
- Data, Limitations, and Recommendations
  - Data selection for training: Training data consists of latin UD-Treebanks, Wikipedia and OSCAR sentence data, the CC-100 Latin dataset and the Herodotos Project NER dataset
  - Data selection for evaluation: Evaluation was done according to the spaCy workflow and is documented in the meta.json file found in the repository (https://huggingface.co/latincy/la_core_web_lg/blob/main/meta.json)
  - Limitations: unknown

B Datasheet: Herodotos Project Dataset

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. created for Herodotos Project to train NER-Tagger (BiLSTM CRF; see: Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeaux-Prunel and Marie-Catherine de Marneffe. 2019. "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities." In Proceedings of North American Association of Computational Linguistics (NAACL 2019). Minneapolis, Minnesota.); Goal of Herodotos Project: catalogue and compendium of ancient ethnic groups; For more info on the corpus see: https://aclanthology.org/W16-4012.pdf

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? from the documentation: „The data files in the Annotation directory were annotated for named entities by a team of Classics experts at Ohio State University. Texts presently included are excerpts from Caesar's Wars, both Gallic (GW) and Civil (CW), the Plinies' writings, both Elder and Younger, and Ovid's Ars Amatoria. "

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. unknown

Any other comments? No

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. Latin texts "Texts presently included are excerpts from Caesar's Wars, both Gallic (GW) and Civil (CW), the Plinies' writings, both Elder and Younger, and Ovid's Ars Amatoria."

How many instances are there in total (of each type, if appropriate)? 146,066 words

Does the dataset contain all po-ssible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). sample of Latin literature (see previous answers), representative of Classical Latin literature, might not be representative of the entire Latin literature (time, geography)

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. Each instance consists of raw text data

Is there a label or target associated with each instance? If so, please provide a

description. NER Labels: PRS-B, PRS-I, GEO-B, GEO-I, GRP-B, GRP-I or 0; labels follow the BIO scheme; see also: https://aclanthology.org/W16-4012.pdf

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. No

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. Relationships are made explicit according to the BIO scheme

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. Text from Gallic War is split into test and train sets

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Naturally occurring repetitions of names in the texts

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate. The dataset is self-contained and can be downloaded from GitHub (https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation/blob/master/README.md)

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description. No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. If the dataset does not relate to people, you may skip the remaining questions in this section. The dataset contains descriptions of war.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. A number of ethnic groups from antiquity are referred to.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how. Only historical individuals

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description. Only historical individuals

Any other comments? No

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? The data consists of publicly available texts

If the data was reported by subjects or indirectly inferred/derived from other data,

was the data validated/verified? If so, please describe how. unknown

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? from the documentation: „All texts are in Latin taken from the Latin Library Collection (collected by CLTK) or the Perseus Latin Collection. "

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? unknown

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? https://aclanthology.org/W16-4012.pdf S. 87: "an undergraduate, a graduate, and a professor of Classics, each with at least 4 years of experience studying Latin"

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. If the dataset does not relate to people, you may skip the remaining questions in this section. unknown

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. not applicable

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. not applicable

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). not applicable

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. not applicable

Any other comments? No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section. The data was manually annotated for NEs.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. The data can be downloaded from: https://github.com/clmarr/Herodotos-beta/tree/f22fdd92b3318cfb8fc93b004b0947aea14ce9c2/Annotation_1-1-19

Any other comments? No

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

It has been used to train an NER-Tagger for Latin. See: https://aclanthology.org/W16-4012.pdf and https://github.com/alexerdmann/HER/blob/master/HER_NAACL2019_preprint.pdf

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. What (other) tasks could the dataset be used for? See: https://github.com/alexerdmann/HER/blob/master/HER_NAACL2019_preprint.pdf

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms? Strong class imbalance (most tokens are non-entities)

Are there tasks for which the dataset should not be used? If so, please provide a description. No

Any other comments? No

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The data can be downloaded from: https://github.com/clmarr/Herodotos-beta/tree/f22fdd92b3318cfb8fc93b004b0947aea14ce9c2/Annotation_1-1-19

Does the dataset have a digital object identifier (DOI)? No

When will the dataset be distributed? The data can be downloaded from: https://github.com/clmarr/Herodotos-beta/tree/f22fdd92b3318cfb8fc93b004b0947aea14ce9c2/Annotation_1-1-19 https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. AGPL-3.0 license

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. unknown

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. unknown

Any other comments? No

Maintenance

Who will be supporting/hosting/maintaining the dataset? from the documentation: "Contact ae1541@nyu.edu or any of the co-authors with questions regarding this repository."

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? ae1541@nyu.edu

Is there an erratum? If so, please provide a link or other access point. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? new instances for the Ancient Greek language will be added in the future

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. not applicable

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers. unknown

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. unknown

Any other comments? No