DIGITAL CLASSICS ONLINE

Daidalos: NER for Literary Studies on Latin and Ancient Greek Texts

Andrea Beyer

Abstract: Literary texts offer a wealth of unstructured data that can be harnessed for data-driven text analysis through Natural Language Processing (NLP). Named Entity Recognition and Classification (NER) is a crucial initial step in this process, enabling the automatic identification of entities such as persons, organizations, locations, and dates. However, NER faces significant challenges, particularly with historical texts in low-resource languages like Latin and Ancient Greek, due to limited annotated corpora and the dynamic nature of language. This paper explores the evolution of NER from simple extraction to semantics-aware entity disambiguation and linking, highlighting the importance of multi-layer annotation systems to enhance data quality and model accuracy. The interdisciplinary *Daidalos* project aims to bridge the gap between Digital Humanities and Classical Studies by providing an NLP infrastructure that supports various data-driven research methods, among others NER. One of the project's case studies demonstrates the potential of NER in Classical literary studies; this is accompanied by proposals on other NER related literary research questions, e.g. on authorship attribution and stereotyping. Additionally, the paper offers some thoughts about teaching NER, presenting a framework to assess the required level of digital literacies when working on a specific research question. Finally, it discusses the implications of generative AI and Large Language Models (LLM) on NER and NLP in Classics, emphasizing the challenges for independent research posed by the high costs and limited transparency of LLMs.

Introduction[1]

Literary texts provide a lot of unstructured data that needs to be extracted and structured, so that a computer can support a data-driven text analysis (Natural Language Processing, NLP). Therefore, automatic taggers search, retrieve and explore the information in a given text corpus and annotate the text accordingly. One of the first and most crucial processing steps in doing so is the extraction technique Named Entity (NE)[2] Recognition and Classification (NER for short). This method enables the automatic identification of entities in texts, entities being for the most part categorised as follows: person, organisation, location and date. This is an essential part of question answering, media monitoring, and opinion mining, however it also helps with machine translation, text summarisation, and text classification.[3] It might additionally be useful for subsequent Information Extraction (IE) tasks like Rela-

1 Acknowledgements: This work is part of a project funded by the German Research Foundation (project no. 518919950) and led by Andrea Beyer, Malte Dreyer, and Anke Lüdeling.

2 Ehrmann et al. (2021), 5: "[…] named entities correspond to different types of lexical units, mostly proper names and definite descriptions, which, in a given discourse and application context, autonomously refer to a predefined set of entities of interest. There is no strict definition of named entities, but only a set of linguistic and application-related criteria which, eventually, compose a heterogeneous set of units." For further discussion see Ehrmann (2008).

3 Ehrmann et al. (2021), 2.

tion Extraction and Entity Linking[4] as well as for Topic Classification, Event Timelines, Network Analysis and various other analysis techniques.[5] Thus, NER has also undergone an evolution by shifting its focus from the mere extraction of information, i.e. the detection and classification of NE, to a more semantics-aware viewpoint, i.e. entity disambiguation and linking, "which can support the cross-linking of multilingual and heterogeneous collections based on authority files and knowledge bases".[6]

Essentially, NER is a sequence-labelling task that is enriched with features at three levels: the morphological level (words, e.g. *Roma, Zeus, Galli*), contextual level (close context or sentences, e.g. *C. Iulius Caesar*, *consulibus M. Tullio Cicerone et C. Antonio Hybrida*), and text level (document, e.g. *C. Plinius Traiano Imperatori*). Developed NER Systems are evaluated in terms of Precision (P), Recall (R) and F-measure (F-score, the harmonic mean of P and R) as well as more fine-grained evaluation metrics.[7] Additionally, the accuracy of NER results is dependent on which resource methods have been used, as well as the inherent challenges that certain research areas may pose, such in the case of Latin and Ancient Greek; here, difficulties may emerge due to the significant changes in the use of these languages over time, as well as the long transmission history of Latin and Ancient Greek texts.

Resource Types

In order to develop NER systems four types of resources exist:[8]

- Typologies: Typologies define a semantic framework for the entities under consideration. They constitute the source of annotation guidelines.

- Lexicons and knowledge bases: On the one hand, information about NE can be of lexical nature given verbatim in a textual unit. By using look-up procedures in lexicons, the NE might be extracted. On the other hand, information about NE can be encyclopaedic in nature, i.e. non-linguistic information on referred entities. For extracting this kind of information, knowledge bases like *Wikipedia* or *Wikidata* are widely used.

- Word embeddings and language models: Word embeddings are dense, low-dimensional vectors that are acquired from the distribution of words in continuous texts and represent the meaning of words. Their ability to be obtained self-supervised, i.e. from unlabelled data, is a major benefit that makes the shift from feature engineering to feature learning possible.

- Corpora: These can consist of labelled and/or unlabelled textual data. Unlabelled texts are used to train language models and embeddings, while labelled corpora are utilised as a learning base or as a point of reference for evaluation purposes.[9]

Low resources are one major problem for research communities working with historical texts and languages, because these communities are rather small and underfinanced. Therefore, they lack the power to significantly improve the unsatisfying situation.

4 Feng et al. (2018), 4071.

5 Chastang et al. (2021).

6 Ehrmann et al. (2021), 3. Other NE-related specific research directions are temporal information processing and geoparsing: Ehrmann et al. (2021), 5.

7 Ehrmann et al. (2021), 6–7.

8 Ehrmann et al. (2021), 7–8.

9 E.g. Latin NER with literary texts (1st century BC – 2nd century AD), 7.175 NE: Person, Location, Group. See Erdmann et al. (2016).

Methods

The engineering of NER systems is performed by four families of algorithms:[10]

	Description	Example	Constraints
Lexicon-based approaches	NE are detected by comparing a dictionary with the list of words in the selected corpus. These look-up procedures work better for historical data, because the lexicons do not require constant updating and careful maintenance to stay accurate and effective.	*Trismegistos*, Domain: state (papyri, 4C–1C BC, languages: Egyptian, Ancient Greek, Latin, cf. Broux / Depauw [2015]); *Pleiades*; *Lexicon of Greek Personal Names*; *Prosopographia Imperii Romani.*	Resources might not be available (digital, open access) nor well-maintained (updated databases). The complexity of NE is a problem in itself related to the ambiguity of names (e.g. father and son: same name, different person), spelling variation (e.g. *Caius* or *Gaius*), abbreviations (e.g. *C./ G.*, *SPQR*), patronyms (e.g. *Pelopides* – "descendants of Pelops"), and metonyms (e.g. *Tonans* – Jupiter). It is almost impossible to include all potential NEs. Chastang et al. (2021), 9: "Training a dictionary-based recognition model against a list of names can lead to a high ratio of recognition for a particular corpus, but the model is often not robust when applied to unseen texts or different types of data."
Rule-based approaches	NE rules are manually crafted by a developer or linguist on the basis of regularities (patterns) observed in the data, e.g. derivates for ancestry (*-ides*). Rule-based approaches have the advantage of not requiring training data and of being easily interpretable. In contrast, their design needs time and expertise and is thus costly. Chastang et al. (2021), 9: "[…] rule-based models […] show a valid global recall, but a slight tendency to poor precision-ratio on unseen texts."	"*ruLe-based lOCation nAmed-entity recognition method for Latin tExt*" (*LOCALE*); for Ancient Greek no example.	Morphological analysis (stemming, lemmatisation) might be useful for applying the rules, but this comes with a certain percentage of errors.

10 Chastang et al. (2021); Ehrmann et al. (2021), 8–10.

Machine-learning (ML) based approaches	These approaches are also called feature-based, because they use labelled data and learn from it to recognise patterns for applying them on unlabelled data. Due to their capacity to take into account the neighbouring tokens, conditional random fields (CRF) proved particularly well-suited for NER tagging and became the standard for feature-based NER systems.	*Classical Language Tool Kit* (*CLTK*); *CRF*. Domain: news (medieval charters, 10C-13C, cf. Aguilar et al. [2016]); Domain: literature (Classical texts, 1C BC-2C, cf. Erdmann et al. [2016], Beersmans et al. [2023]; Domain: literature (Herodotus, cf. Palladino et al. [2020]).	The ML algorithms require a lot of labelled data which is usually scarce. Besides, in order to resolve ambiguities, a deep understanding of the context (sentences, paragraphs, document) is necessary, something which is a challenge for a rather simple ML algorithm.
Deep learning (DL) approaches	DL systems use artificial neural networks with multiple processing layers, which is why they are called neural-based approaches. On the basis of word embeddings, DL models the semantic and syntactic relationship between various words by learning representations of the given data (corpus) with multiple levels of abstraction. The key benefit of neural networks is their ability to automatically learn input representations instead of relying on manually elaborated features – whether or not the input is topic-specific or rather general.	*Latin BERT* (trained on 640 million tokens spanning 22 centuries, cf. Bamman / Burns [2020]);LatinCy; *Grε(BERT|T)A*, *PHIL(BERT|T)A; AG_BERT_hypopt_reduced_NER; grc-ner-bert; flair_grc_bert_ner*	Deep learning models "frequently operate as opaque 'black boxes' with limited transparency in their decision-making processes" (Sankarapu et al. [2024], 1).

Tab. 1: Methods of Named Entity Recognition (NER).

Challenges of NE Recognition and Classification

Generally, NER on historical documents, particularly those that are written in low-resource languages like Latin and Ancient Greek, faces considerable challenges due to a lack of resources, in particular the documentation on high-quality annotated datasets.[11] The lack of resources for Latin and Ancient Greek summarised by Burns (2019), is still valid:

> "[…] named entity recognition (NER), or the systematic tagging of words in texts by category (so, Roma as a "location" or Σωκράτης as a "person") is not well-supported by standalone tools. With respect to Greek and Latin, a lack of annotated texts and robust language models underlies the problem."[12]

11 Beersmans et al. (2023).

12 Burns (2019), 168.

Even though the AI-driven development in the last few years has brought some advances relating to the models and evaluation procedures, the overall trend has not changed, particularly concerning annotated corpora.

Apart from the quality of data and methods there are challenges inherent to language corpora which contain texts from different genres, places, and eras. Although the Latin and Ancient Greek texts belong mostly to literature and use standardised, formal language, the dynamics of language might cause some troubles extracting and labelling NE correctly. These errors can be attributed to problems related to normalisation, genre-specifics, ambiguity, and multi-word expressions:

- Normalisation: Without normalising NE to a common standard, circumstances such as spelling variations and slightly different naming conventions may produce multiple results for the same NE, e.g. **C.** Iulius Caesar and **G.** Iulius Caesar, *par**v**um* and *par**u**um*.

- Genre specifics: Poetic texts in particular provide stylistic or rhetorical expressions like paraphrases, metaphors and metonymies, which obfuscate the NE for a non-expert like a standard automatic NER-tagger, e.g. *urbs* instead of Rome, *tonans* instead of Jupiter, *Dis* instead of Pluto.

- Ambiguity: Some of the naming conventions of antiquity like homonyms between father and son are very challenging for automatic NER taggers because they do not 'understand' the concept of one name and two referents, i.e. they were not trained to distinguish multiple homonymous referents. Similarly, this applies to adjectives, e.g. *romanus* meaning Roman (adjective, no NE) or Roman (noun, NE: person). In both examples, a NER tagger is prone to make mistakes in the classification process, e.g. identifying too many or too few entities.

- Multi-word expressions: Occurrences of multi-word expressions ('composed entities') come with different challenges. Firstly, they consist of more than one word with different linguistic construction possibilities like *vallum Hadriani*, *consulibus M. Tullio Cicerone et C. Antonio Hybrida*, *P. Lentulus Sura*, *P. et Ser. Sullae Ser. Filii*. Secondly, they might occur as discontinuous text spans like *rebus Sancti Vincentii Matiscensis*. Thirdly, they might be nested into each other or might overlap like *Guillelmus de Sancti Stephano de Ponte*. In every case they are highly complex and therefore mostly not recognised correctly, i.e. the NE will probably be identified only partially and often not correctly classified.

Approaches to enhance NER results

As mentioned above, to gain better results it is necessary to increase the quantity of data for training and evaluation purposes as well as to improve the quality of this data. By providing more annotated texts of different genres and eras the issue of quantity could be addressed, but without more sophisticated annotations the aforementioned challenges would be almost the same. Thus, the quality of data and consequently of the models needs to be enhanced. Accordingly, it is suggested to enhance the annotations by introducing a so-called multi-layer annotation system; this allows researchers to make explicit annotations in cases of uncertainty and to reveal the distinction in complexity between manual and automatic annotation through one layer of automatic annotation and multiple layers for manual ones.[13] Most notable is the latter functionality, which offers insights into the contrast between a broader automatic approach and a more subtle manual approach. Based on these multi-layer annotations, NER taggers using a DL approach could learn to recognise patterns of fuzzy multi-word expressions, extract them in their entirety as one NE, and classify the NE accurately.

13 Chastang et al. (2021).

Application of NER in Classics

As part of the Humanities and *inter alia* concerned with literary studies, modern Classical Studies has intersections with the Digital Humanities (DH) and with the Computational Literary Studies (CLS). Nowadays, digital research and teaching methods enrich the established methodology and offer new perspectives on the interaction with Latin and Ancient Greek texts. This encompasses all main areas of German Classical Studies: editions, translations, commentaries, interpretations, didactics. However, there is still a huge knowledge gap between the people who develop new resources and tools and the people who might use them for their research and teaching. This gap is two-sided: On the one hand, 'traditional' researchers and teachers lack a sufficient level of digital literacies[14] for fully exploiting the provided tools and methods. On the other hand, specialists for computational methods are often not sufficiently acquainted with the domain-specific needs and research interests – occasionally, they are not even familiar with the Classical languages or literary studies. Thus, matching the research interests and goals of both perspectives sometimes seems impossible. This marks an exciting starting point for the *Daidalos* project, which aims to develop an NLP research infrastructure for different competency levels.

Daidalos: Key Goals and Features

The interdisciplinary *Daidalos* project[15] intends to bring computational and traditional approaches closer together by providing an NLP infrastructure which offers a software-as-a-service for NLP methods like NER, word embeddings, or sentiment analysis. Additionally, *Daidalos* includes working in research tandems, offering material for further education, and consulting researchers or research groups on third-party funding or more generally on research questions appropriate to digital methods. Therefore, *Daidalos* has the following goals and features:

- Bridging the Gap between DH and CLS on one side and German (!) Classical Studies on the other side by working in so-called research tandems for a better understanding of what literary researchers need and want.[16]
- No-code and low-code access for various data-driven research methods and data visualisations.[17]
- Reuse of existing resources and literature overview on application possibilities.
- Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek: SEFLAG.[18]

14 This term sums up different technology-driven literacies, i.e. digital literacy, data literacy, and AI literacy. A domain-specific and case-sensitive theoretical model and framework has been designed as part of the *Daidalos* project. See Beyer (2026).

15 https://daidalos-projekt.de (last access 11.07.2025). The *Daidalos* Project (2023–2026) is funded by the German Research Foundation. Project number: 518919950. *Daidalos* is on GitLab (https://scm.cms.hu-berlin.de/daidalos/daidalos-platform [last access 11.07.2025]) and Zenodo (https://zenodo.org/communities/daidalos/ [last access 11.07.2025]).

16 Beyer / Schulz (2023a) and (2024).

17 Beyer / Schulz (2023b).

18 Schulz / Deichsler (2024).

Feature	Already implemented	Planned
Corpus of Latin and Ancient Greek texts	*Perseus* via *Perseids API*	Integration of other digital corpora like *LASLA*, *MQDQ*
NLP methods and visualisation	Pre-processing, word embeddings, NER, sentiment analysis	Part-of-Speech, topic modelling, fine-tuning of parameters
Workshop material	(Hosted on GitLab:) Jupyter Notebooks on Markdown/JN, data preprocessing, word embeddings, sentiment analysis, NER	Integration via JupyterLite for low-code access and increased flexibility during research
SEFLAG	NER, lemmatisation, dependency parsing	Expanding, web integration
Databases	Small documentation on resources	Research literature, AI tools
User-friendly functionality		Text upload, result download
Identity Access Management		Settings, own corpus

Tab. 2: Overview on features of the *Daidalos* NLP research infrastructure.

The *Daidalos* project will apply for a second funding period, in which, amongst other things, the further development of domain-specific NLP methods shall be addressed.

NER – Underrated for Literary Studies in the Classics?

The methods of information retrieval (IR) and linguistics share a common fate in (traditional) literary studies in (German) Classics. Although they might be called the foundation ('ground truth') of any qualitative analysis – in the form of an interpretation of a passage or a text, they are often not explicitly mentioned or are even despised because of their data-driven approaches. Nevertheless, of course, they are used by philologists to analyse text patterns and extract information from texts about e.g. people, author, or historical context. That being said, the key challenge for establishing data-driven methods in the community of practice of Classical philologists is rather a question of awareness,[19] both of methodology and methodological competence. For this reason, *Daidalos* has introduced the so-called research tandems. Both partners know the domain, but they apply different methods for solving their research questions – *close reading* vs. *distant reading*.[20]

In one of these case studies, NER is part of a workflow to find answers to the underlying research question: How can 'gaps' in historiographical texts be found using digital methods?[21] For modelling reasons this question is reduced to a manually analysed example: Can the absence of the historical event 'Conference of Luca' in Cassius Dio's *Roman History* be detected with digital methods?[22] The

19 De Carvalho-Filho et al. (2020).

20 Schubert (2015).

21 This research question addresses the problem whether an omission is based on a source problem or is intentionally done as part of a conscious decision (literary function).

22 The Conference of Luca was held 56 BC by the triumvirs Caesar, Pompey, and Crassus. It is mentioned by Cicero, Plutarch, Velleius Paterculus, Suetonius, and Appian. In contrast, Cassius Dio does not refer to it.

test corpus consisted of nine references, both in Latin and Ancient Greek texts.[23] Thus, two language specific NER taggers were applied (see tab. 3).

	Latin	Ancient Greek
Model Name	la_core_web_lg	UGARIT/flair_grc_bert_ner
Publication	Burns, 2023	Yousef et al., 2023
NLP Software	spaCy	Flair NLP
Architecture	floret vectors Transition-based Parser	BERT (transformer) vectors long short-term memory network conditional random field
Training Data	Caesar, Ovid, Pliny (Elder & Younger)	Homer, Herodotus, Athenaeus
Tagset	persons, locations	persons, locations, peoples

Tab. 3: Reuse of resources in *Daidalos*: NER taggers for Latin and Ancient Greek.

As it is shown in fig. 1, it was possible to find the passage containing the conference of Luca through visualising the NE.[24] Nevertheless, there was also one false positive (Plut. Caes. 21, 4) that could be explained by the literary scholar, but proved what is generally accepted about NLP tasks: having a human in the loop is essential for evaluating the results. In relation to the research request, the results were good enough for a proof-of-concept state, esp. because the known gap (Cass. Dio 39, 24–36) stayed blank.

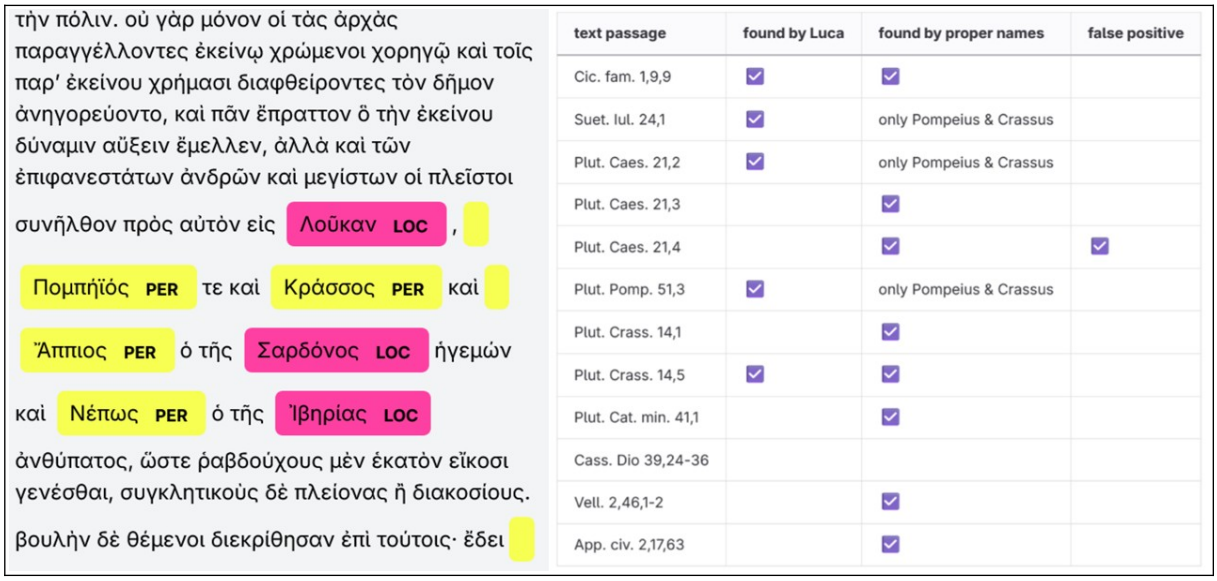

Fig. 1. Left: Example (Plut. Caes. 21) of NER visualisation with *Daidalos*; Right: Overall evaluation of NER results for the test corpus.

23 Cass. Dio 39,24–36; Cic. Fam. 1,9,8–9; Suet. Iul. 24,1; Plut. Caes. 21; Plut. Pomp. 51; Plut. Crass. 14–15; Plut. Cat. min. 41,1–2; App. Civ. 2,17; Vell. 2,46,1–2.

24 If someone already knows all NEs beforehand, he/she could use regular expressions as well. However, using the NER-tagger is a proof-of-concept for finding specific passages without knowing all NEs.

Based on these findings, the research tandem decided to continue further with a combination of NER and sentiment analysis to test a second hypothesis[25] that explains why Cassius Dio might not have mentioned the conference of Luca. The first findings are encouraging, namely that the polarity of the Cassius Dio passage is slightly more negative than any of the others.[26] Thus, the test corpus for this analysis is going to be expanded to the whole work of Cassius Dio, in order to evaluate whether the workflow 'NER and sentiment analysis' paves the way sufficiently to tackle the overall research question, i.e. finding omissions in historiographical Latin and Ancient Greek literature.

Obviously, the presented case study is just one possible application of NER in Classics. Inherently, NER techniques support complex information extraction and classification in corpora too big to read in the provided amount of time (cf. distant reading). This way, they enable literary scholars to visualise frequencies, sequences, recurrent patterns, and hot spots to gain systematic and replicable insights into specifically assembled corpora. Such being the case, some other potential literary use cases are outlined below:

Authorship Attribution

- Corpus: Ps.-Sallust, *Invectiva in M. Tullium Ciceronem*.
- Background: By now, most researchers agree that this speech is an example of a *declamatio* of the late Augustan era, but esp. in Germany some researchers still try to attribute it to Sallust.
- Research question: If the speech is situated in 54 BC, do the timeline and chronology of appearing persons match this date? What evidence is there for a *terminus post quem*?
- Requirements: Digitised Text and a knowledge base of events and persons up to 54 BC.

Relationship of Title Heroes and Protagonists

- Corpus: Sallust, *Bellum Catilinae* & *Bellum Iugurthinum*.
- Background: Both monographs are named after one person (Catilina & Iugurtha) exemplary for the danger they posed to Rome. However, there are other main characters (Caesar & Cato; Marius & Sulla) who drive on the events.
- Research question: How is the relationship between eponymous hero and protagonists weighted? How similar are the two monographs in this respect to each other?
- Requirements: In addition to NER, social network analysis is required.

Time as a Device for Framing

- Corpus: e.g. Cicero, *De Amicitia*, *Tusculanes*; Plato, *Protagoras*, *Politeia*.
- Background: In antiquity, the genre *dialogus* is common to 'teach' philosophy. Some are explicitly set in the past, others span multiple days in a row.

25 Cassius Dio attributes early on a negative relationship between Pompey and Caesar, which might be the reason for Luca's 'gap'. If so, this should be reflected by the (latent) emotions/moods in the text, i.e. more negative polarity in Cassius Dio than in the other passages of the test corpus.

26 The sentiment analysis for both languages is a challenge in it-self, esp. for comparable resources and models. Regardless, this NLP task is introduced in this paper only for the reason of demonstrating a pipeline in which NER is one of the first steps to discover new insights in the classical-philological literary studies.

- Research question: How is time used in the *dialogi*? Do references to time statements fit the fictitious frame of a dialogue? How do time statements support the narration? What can be inferred about the meaning of time in literature of the same era?

- Requirements: <DATE> annotation in a training corpus and a newly trained model.

Stereotypes of Organisations or People over Time

- Corpus: Historiographical texts from 100 BC to 300 AD.

- Background: Different political systems and becoming an empire may have had an influence on Latin literature and the ideas conveyed by the historiographical works.

- Research question: How do political organisations and their representatives evolve over time? Are the descriptions rather like stereotypes in the beginning? Do they evolve into stereotypes? Are certain peoples more prone to be stereotyped than others?

- Requirements: <ORG> and <PEOPLE> annotation in a training corpus and a newly trained model. In addition to NER, social network analysis and word embeddings are recommended.

These examples are only thought experiments, of course, and each idea would need more consideration before being applied in literary research. But they hint at the potential value of an IR method like NER in Classical literary studies, if both types of researchers work together on research questions.

NER as Part of Teaching Digital Literacies

As shown above, it is worth knowing how to use NER in the Classical Studies, but the majority of researchers are lacking the necessary skills. Above all, this poses a challenge for the future of Classics in itself, because the researchers are also teachers. So, if the researchers are too unskilled in digital research methods, they will not or cannot teach in a way that enables students to acquire a certain level of digital literacies. Consequently, future researchers and teachers will still have the same problem they face right now. That is why the *Daidalos* project has also developed a framework for describing domain-specific and use-case-sensitive levels of three digital literacies, i.e. digital literacy, data literacy, and AI literacy,[27] which are aggregated as digital literacies. This framework is used to define what is needed for accomplishing digital-based research. Thus, it helps in providing support to literary scholars, and in developing learning material like the curated Jupyter Notebooks.

For example, in the case study presented above, the literary scholar did not know anything about NER (which was expected), but lacked also ‘basic’ knowledge about file formats or even the term ‘annotation’ (which was unexpected, see tab. 4). For one thing, why should a literary scholar be expected to know about annotations? On the other hand, these expectations are reasonable from the view of those researchers who are accustomed to NLP and surrounded by digital devices and data-driven approaches. In fact, both ‘sides’ would be right in their own perspective, illustrating the aforementioned two types of researchers who have difficulties to understand each other. For that reason, although it is hard work to adapt the framework for each case study, it has proven very useful for communication and collaboration in the research tandems.

27 There is an abundance of literature on each of the mentioned literacies and some more that might also be part of the digital literacies framework, e.g. media literacy or information literacy. However, deploying NLP services boils down to the fact that the focus is set on AI and supportive competences related to data and general digital abilities.

Dimension	Get to know	Acquire	Deepen	Create
AI literacy	NER concepts, e.g. taggers, model cards, datasheets	Comparing NER taggers and results	Evaluating NER errors systematically	Refining NER taggers through feedback
Data literacy	Knowledge about annotations in general	Annotating a test corpus	Using a multi-layer annotation scheme	Creating annotation guidelines
Digital literacy	Difference DOCX & TXT	Converting DOCX to TXT	Identifying challenges for pre-processing in TXT	Modifying TXT depending on the use case

Tab 4: Domain-specific and use-case-sensitive framework to describe the needed level of digital literacies for working on an NER task. Not all characteristics need to be acquired for accomplishing a task, but the full overview breaks down the complex requirements into learnable modules.

Obviously, the framework might also be applied to teaching and could facilitate setting the goals of tasks, courses or curricula by structuring explicitly the needed skills. Even if it is mostly not necessary to acquire the full pattern of digital literacies for the selected use case, providing a variety of frameworks for different use cases in Latin and Ancient Greek philology might make it easier for traditionally trained scholars to get an idea of what level of digital literacies is required to apply digital research methods. Consequently, it would be possible to address this need and create learning units with authentic research questions, in order to improve digital literacies in Classics and therefore break the vicious circle of lacking digital competence.

Generative AI and the Future of NER and NLP in Classics

Finally, it is necessary to talk about the implications of the frantic development of generative AI related to NLP tasks and Classics as well. Some people might ask why they should bother to acquire digital literacies, if they just as well could use an AI chatbot, i.e. a Large Language Model (LLM)[28] with a graphical user interface. This opens up a discussion that not only concerns the quality of these models and further technology advancements, but also core values of research like autonomy, replicability, transparency, and openness: In a nutshell, what we can do and what we should do.

Getting back to NER, there is some evidence that LLMs, namely the GPT models, can be deployable for NER tasks (Wang et al. 2023). Although the "intrinsic gap between the two tasks of NER and LLMs"[29] and "the hallucination issue, where LLMs have a strong inclination to over-confidently label NULL inputs as entities"[30] still exist, the researchers showed that using in-context learning – primarily few-shot prompting – combined with a self-verification process of the LLMs is especially beneficial "in the low-resource scenario […] when the amount of training data is extremely scarce"[31]. Consequently, this paper indicates that, by prompting, it is possible to structure unstructured data and process this data in such a way and with such results that are comparable to what specialised taggers

28 As mentioned above, deep learning approaches are being tested and adapted right now for Latin and Ancient Greek. Thus, LLMs are already being applied. In contrast to this meaning, in the following context, the term LLM is related to foundation models that are domain-unspecific and accessed through a chatbot interface by prompting a question or task in the natural language of the user.

29 Wang et al. (2023), 1: "NER is a sequence labeling task in nature, where the model needs to assign an entity-type label to each token within a sentence, while LLMs are formalized under a text generation task."

30 Wang et al. (2023), 2.

31 Wang et al. (2023), 11.

would do. Nevertheless, it should be added that, in the light of the evolution of LLMs to Large Action Models (LAM), these findings might soon be outdated. With the arrival of agents[32], a task could be broken up into a lot of individual actions and binary decisions which follow an even better workflow than the prompt engineering so far. Accordingly, the need for training models or curating resources could be reduced, especially for low-resource languages like Latin and Ancient Greek, though, in fact, nobody can really make an accurate estimation for the future.

Despite this, a much more intricate problem is involved when talking about LLMs – in particular LAMs – or AI-driven technology in general. As it is known,[33] the costs for training a LLM, but also for fine-tuning, applying techniques like retrieval-augmented generation (RAG) and low-rank adaptation (LoRA), and for operating the LLMs are extremely high. Thus, the usage of LLMs creates new boundaries, because only global private players are able to finance the 'whole AI ecosystem' so far. This ever-growing dependency on a non-scientific, profit-oriented community comes with crucial drawbacks for an independent research community:

- reduced accessibility due to high prices, costly infrastructure and lack of data privacy,
- restricted replicability due to a lack of openness of models and training data,
- unclear biases due to training the models on the whole internet,
- limited or no transparency at all due to privately owned solutions and the DL approach.

Under these conditions, small domains like Classics have an even harder time working autonomously with digital-based research methods. This means that abstaining from 'buying' seemingly easy access to LLMs, and keeping instead to the old-fashioned NLP approaches might be the only salubrious solution yet. Consequently, researchers still need to learn about NER as an NLP task and become digitally literate.

32 Xi et al. (2023), 1: "AI agents are artificial entities that sense their environment, make decisions, and take actions."

33 Maslej et al. (2024).

Online Sources

https://pleiades.stoa.org/places (last access 11.07.2025).

https://www.lgpn.ox.ac.uk/ (last access 11.07.2025).

https://pir.bbaw.de/ (last access 11.07.2025).

https://trismegistos.org (last access 11.07.2025).

https://github.com/Ivona221/LOCALE (last access 11.07.2025).

http://cltk.org/ (last access 11.07.2025).

http://daidalos-projekt.de (last access 11.07.2025).

References

Bamman / Burns (2020): D. Bamman / P. J. Burns, Latin BERT: A Contextual Language Model for Classical Philology, online 2020, http://arxiv.org/abs/2009.10053 (last access 11.07.2025).

Beersmans et al. (2023): M. Beersmans / E. de Graaf / T. Van de Cruys / M. Fantoli, Training and Evaluation of Named Entity Recognition Models for Classical Latin. Proceedings of the Ancient Language Processing Workshop, online 2023, 1–12, https://aclanthology.org/2023.alp-1.1.pdf (last access 11.07.2025).

Beersmans et al. (2024): M. Beersmans / A. Keersmaekers / E. De Graaf / T. Van De Cruys / M. Depauw / M. Fantoli, "Gotta catch 'em all!": Retrieving people in Ancient Greek texts combining transformer models and domain knowledge. Proceedings of the 1st Workshop on Machine Learning for Ancient Languages, online 2024, 152–164, https://doi.org/10.18653/v1/2024.ml4al-1.16 (last access 11.07.2025).

Beyer (2026, forthcoming): A. Beyer, Textanalyse mit KI: Wie kommt sie in die Klassische Philologie?, in: S. Faller / W. Polleichtner (ed.), Digitalität im Unterricht der Alten Sprachen. (Digitale) Chancen für den Lateinunterricht, Baden-Baden 2026.

Beyer / Schulz (2023a): A. Beyer / K. Schulz, DAIdalos: Forschen und Lernen zugleich?, online 2023, 391–393, https://doi.org/10.18420/inf2023_42 (last access 11.07.2025).

Beyer / Schulz (2023b): A. Beyer / K. Schulz, Data Literacy für die Klassische Philologie: d^{AI}dalos – eine interaktive Infrastruktur als Lernangebot, online 2023, https://doi.org/10.5281/zenodo.8420565 (last access 11.07.2025).

Beyer / Schulz (2024): A. Beyer / K. Schulz, Daidalos: Wie viel Methodenkompetenz braucht ein User? Book of Abstracts – DHd2024, online 2024, 336–338, https://doi.org/10.5281/zenodo.10698299 (last access 11.07.2025).

Broux / Depauw (2015): Y. Broux / M. Depauw, Developing Onomastic Gazetteers and Prosopographies for the Ancient World Through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People, in: L. M. Aiello / D. McFarland (eds.), Social Informatics, Heidelberg 2015, 304–313, https://doi.org/10.1007/978-3-319-15168-7_38 (last access 11.07.2025).

Burns (2019): P. J. Burns, Building a Text Analysis Pipeline for Classical Languages, in: M. Berti (ed.), Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, Berlin / Boston 2019, 159–176, https://doi.org/10.1515/9783110599572-010 (last access 11.07.2025).

Burns (2023): P. J. Burns, LatinCy: Synthetic Trained Pipelines for Latin NLP, online 2023, https://doi.org/10.48550/arXiv.2305.04365 (last access 11.07.2025).

Chastang et al. (2021): P. Chastang / S. O. Torres Aguilar / X. Tannier, A Named Entity Recognition Model for Medieval Latin Charters, Digital Humanities Quarterly 15/4 (2021).

de Carvalho-Filho et al. (2020): M. A. de Carvalho-Filho / R. A. Tio / Y. Steinert, Twelve tips for implementing a community of practice for faculty development, Medical Teacher 42/2 (2020), 143–149, https://doi.org/10.1080/0142159X.2018.1552782 (last access 11.07.2025).

Ehrmann (2008): M. Ehrmann, Les Entités Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation. Paris 2008, https://hal.science/tel-01639190v1/document (last access 11.07.2025).

Ehrmann et al. (2021): M. Ehrmann / A. Hamdi / E. L. Pontes / M. Romanello / A. Doucet, Named Entity Recognition and Classification on Historical Documents: A Survey, ACM Computing Surveys 56/2 (2021), 1–47, https://doi.org/10.1145/3604931 (last access 11.07.2025).

Erdmann et al. (2016): A. Erdmann / C. Brown / B. Joseph / M. Janse / P. Ajaka / M. Elsner / M.-C. De Marneffe, Challenges and Solutions for Latin Named Entity Recognition. Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), online 2016, 85–93.

Feng et al. (2018): Feng X. / Feng X. / Qin B. / Feng Z. / Liu T., Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, online 2018, 4071–4077. https://doi.org/10.24963/ijcai.2018/566 (last access 11.07.2025).

Maslej et al. (2024): N. Maslej / L. Fattorini / R. Perrault / V. Parli / A. Reuel / E. Brynjolfsson / J. Etchemendy / K. Ligett / T. Lyons / J. Manyika / J. C. Niebles / Y. Shoham / R. Wald / J. Clark, Artificial Intelligence Index Report 2024, online 2024, https://arxiv.org/abs/2405.19522v1 (last access 11.07.2025).

Palladino et al. (2020): C. Palladino / F. Karimi / B. Mathiak, NER on Ancient Greek with minimal annotation, online 2020, https://doi.org/10.17613/j7jt-b052 (last access 11.07.2025).

Palladino / Yousef (2024): C. Palladino / T. Yousef. Development of Robust NER Models and Named Entity Tagsets for Ancient Greek, LT4HALA, online 2024, https://aclanthology.org/2024.lt4hala-1.11.pdf (last access 11.07.2025).

Riemenschneider / Frank (2023), F. Riemenschneider / A. Frank, Exploring Large Language Models for Classical Philology, online 2023, https://doi.org/10.48550/arXiv.2305.13698 (last access 11.07.2025).

Sankarapu et al. (2024): V. K. Sankarapu / C. Chitroda / Y. Rathore / N. K. Singh / P. Seth, DLBacktrace: A Model Agnostic Explainability for any Deep Learning Models, online 2024, https://doi.org/10.48550/arXiv.2411.12643 (last access 11.07.2025).

Schubert (2015): C. Schubert, Close Reading und Distant Reading. Methoden der Altertumswissenschaften in der Gegenwart, Digital Classics Online 1,1 (2015), 1–6, https://doi.org/10.11588/dco.2015.1.20483 (last access 11.07.2025).

Schulz / Deichsler (2024): K. Schulz / F. Deichsler, SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek, in: M. Hämäläinen / E. Öhman / S. Miyagawa / K. Alnajjar / Y. Bizzoni (eds.), Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, online 2024, 247–258, https://aclanthology.org/2024.nlp4dh-1.24 (last access 11.07.2025).

Wang et al. (2023): Wang S. / Sun X. / Li X. / Ouyang R. / Wu F. / Zhang T. / Li J. / Wang G., GPT-NER: Named Entity Recognition via Large Language Models, online 2023, https://doi.org/10.48550/arXiv.2304.10428 (last access 11.07.2025).

Xi Z. et al. (2023): Xi Z. / Chen W. / Guo X. / He W. / Ding Y. / Hong B. / Zhang M. / Wang J. / Jin S. / Zhou E. / Zheng R. / Fan X. / Wang X. / Xiong L. / Zhou Y. / Wang W. / Jiang C. / Zou Y. / Liu X. / Yin Z. / Dou S. / Weng R. / Cheng W. / Zhang Q. / Qin W. / Zheng Y. / Qiu X. / Huang X. / Gui, T., The Rise and Potential of Large Language Model Based Agents: A Survey, online 2023, https://doi.org/10.48550/arXiv.2309.07864 (last access 11.07.2025).

Yousef et al. (2023): T. Yousef / C. Palladino / S. Janicke, Transformer-Based Named Entity Recognition for Ancient Greek, Digital Humanities 2023: Book of Abstracts, online 2023, 420–422, https://zenodo.org/records/8107629 (last access 11.07.2025).

Figure References

Fig. 1. Left: Example (Plut. Caes. 21) of NER visualisation with *Daidalos*; Right: Overall evaluation of NER results for the test corpus.

Author Contact Information[34]

Dr. Andrea Beyer
Humboldt-Universität zu Berlin
Sprach- und Literaturwissenschaftliche Fakultät
Unter den Linden 6
10099 Berlin
E-mail: beyeranz@hu-berlin.de

34