# Training and Evaluation of Named Entity Recognition Models for Classical Latin

**Marijke Beersmans** and **Evelien de Graaf** and **Tim Van de Cruys** and **Margherita Fantoli**
KU Leuven, Faculty of Arts
Blijde Inkomststraat 21, 3000 Leuven, Belgium
marijke.beersmans@kuleuven.be
evelien.degraaf@kuleuven.be
tim.vandecruys@kuleuven.be
margherita.fantoli@kuleuven.be

## Abstract

We evaluate the performance of various models on the task of named entity recognition (NER) for classical Latin. Using an existing dataset, we train two transformer-based Latin-BERT models and one shallow conditional random field (CRF) model. The performance is assessed using both standard metrics and a detailed manual error analysis, and compared to the results obtained by different already released Latin NER tools. Both analyses demonstrate that the BERT models achieve a better f1-score than the other models. Furthermore, we annotate new, unseen data for further evaluation of the models, and we discuss the impact of annotation choices on the results.

## 1 Introduction

Commonly an important precursor to information extraction, text summarisation and the creation of knowledge bases, Named Entity Recognition (NER) has become a ubiquitous task in Natural Language Processing (NLP). For modern high-resource languages, generic NER off-the-shelf solutions, focusing mainly on identifying locations, organizations and people, can produce highly accurate annotations. For historical languages, even prolific ones like Latin, the task remains a challenge, in part due to a lack of annotated corpora and tools (Ehrmann et al., 2021).

We pursue three main objectives with this paper:

- We compare the performance of three different models for Latin NER using pre-existing, openly available data. The comparison is both quantitative and qualitative.
- Based on the analysis of existing annotations and the results of automatic annotation, we publish a new set of **gold data**, providing documentation of the most critical choices.
- By using the newly annotated data to assess the results of NER, we publish the **automatic annotation** by the best-performing model of a large corpus of literary classical Latin texts and documenting the strengths and weaknesses of the resulting annotation.

The paper contributes to the application of NLP to Latin on a methodological level, since we propose a thorough analysis of the results of NER on Latin and identify the most critical points. In addition, the paper is associated with the publication of NER models and datasets, and documents the choices that have been implemented. The paper is structured as follows: after introducing existing work and datasets related to NER for Classical Languages (Section 2), we describe the data used, and the training of the models and their performance on in-domain and out-of-domain test sets (Section 3). Section 4 provides a qualitative error analysis of the best performing model based on F1 metrics. In section 5, we introduce the annotation of new data from the LASLA corpus, and analyse the results of the automatic annotation by the best-performing model. The data and code related to this paper are made available on a Github repository.[1]

## 2 Related work

Previous work has highlighted the challenges linked to NER for Latin. Ehrmann et al. (2021) identified among others the following relevant challenges concerning NER on historical documents: variable and sparse feature space (generalizing over different genres and domains, cf. Erdmann et al. (2016)), dynamics of language such as spelling variations and change in naming conventions, general lack of resources (e.g. typologies from modern languages not fitting for historical documents). In addition, Burns (2023) underlined another difficulty of the already scarce resources: differences in orthographic conventions and annotation

[1] https://github.com/NER-AncientLanguages/Ner-Latin-RANLP.



*Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023*, pages 1–12,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_001

schemes. Lastly, both Chastang et al. (2021) and Torres Aguilar (2022) consider the frequency of overlapped and nested entities in Latin as a challenge.

When it comes to existing models, Chastang et al. (2021) trained a CRF-based model on Latin medieval charters from Burgundy. Later Torres Aguilar (2022) tested two approaches for creating a multilingual pipeline for medieval charters (French, Spanish and Latin): the first uses contextual and static embeddings coupled to a BiLSTM-CRF (Bidirectional Long-Short Term Memory) classifier, and the second employs a fine-tuning method using the pre-trained multilingual BERT and RoBERTa models. For both of these efforts, custom charter corpora were annotated. In the context of the Herodotos project — which aims to catalogue ancient ethno-political groups and their interactions — Erdmann et al. (2016, 2019) created a neural, BiLSTM-CRF based entity recognizer (Lample et al., 2016) trained on classical Latin texts. In addition, NER is included in text analysis pipelines for Latin, such as the Classical Language Toolkit (CLTK; Johnson et al., 2021) and LatinCy (Burns, 2023).

In recent years, transformer-based models (with the BERT architecture as one of the prime instantiations) have become the norm for various NLP applications (Ehrmann et al., 2022; Sprugnoli et al., 2022; Sommerschield et al., 2023). These models have been leveraged, *inter alia*, for Latin morphosyntactic tagging (Wróbel and Nowak, 2022; Mercelis and Keersmaekers, 2022; Nehrdich, 2022) and translation alignment for ancient languages (Yousef et al., 2022b), which could also be leveraged for named entity projection from modern languages given a parallel corpus (Yousef et al., 2023). For Greek NER, a BERT-based approach equally proved to be effective (Yousef et al., 2022a). There already exists a transformer-based model for Latin (LatinBERT; Bamman and Burns, 2020) but to the best of our knowledge, it has not yet been fine-tuned on the task of named entity recognition.

Regarding datasets, the Herodotos dataset (at the time of training) is the only available NER dataset for classical Latin (Erdmann et al., 2019, 2023). Additionally, the authors of the LatinCy pipeline are planning to make their custom dataset publicly available (Burns, 2023). Lastly, the multilingual Medieval charter dataset, which includes non-classical Latin (Torres Aguilar, 2022), is available online.[2] We decided to annotate new material to augment the availability of data for classical Latin.

| text | # tokens |
|---|---|
| *BGall.* | 58,621 |
| *NH* | 35,672 |
| *Ep.* | 18,571 |
| *Ars am.* | 17,102 |
| *BCiv.* | 4,819 |

Table 1: Number of tokens per text in the Herodotos dataset

## 3 Data and methods

### 3.1 Data

The Herodotos dataset contains two full texts, Caesar's *Bellum Gallicum* (*BGall.*) and Ovid's *Ars Amatoria* (*Ars am.*), and excerpts from three other texts: a part of the first book of Caesar's *Bellum Civile* (*BCiv.*); book 1, book 2 and a part of book 3 of Pliny the Younger's *Epistulae* (*Ep.*); the preface, first and a part of the second book of Pliny the Elder's *Naturalis Historia* (*NH*). The editions were taken from the Latin Library (Carey, s.d.) and the Perseus Project (Smith et al., 2000). Table 1 contains an overview of the dataset sizes.

The texts are manually annotated for location ('LOC'), person ('PERS') and (socio-ethnic) group ('GRP') entities (Erdmann et al., 2016). The annotations are encoded in BIO-format, where each token is mapped to an 'O' (for 'outside', not an entity) or an entity type with either a B- or an I-prefix. The B-prefix, for 'beginning', indicates the first or only word of an entity whereas the I-prefix, for 'inside', specifies a continuation of a multi-word entity. Nested entities were not considered.

On the whole dataset, minimal preprocessing was performed to iron out formatting mistakes. Afterwards, the five works were divided into two parts: in-domain, used for training and in-domain testing, and out-domain, used exclusively for out-domain testing. The latter should assess the model's generalizing capabilities to texts that are significantly different from the data it was trained on. In this experiment, the in-domain part consisted of the prose texts, (*BGall.*, *Bciv.*, *Ep.* and *NH.*) The out-domain part consisted of the one poetry text, *Ars. Am.*.

[2] https://gitlab.com/magistermilitum/ner_medieval_multilingual/

| type | frequency | |
|---|---|---|
| | Train | Validation |
| O | 82,696 | 13,846 |
| B-PERS | 2,706 | 473 |
| I-PERS | 618 | 125 |
| B-LOC | 839 | 169 |
| I-LOC | 31 | 10 |
| B-GRP | 1,271 | 207 |
| I-GRP | 4 | 2 |

Table 2: Frequency of entity types in train (left) and validation set (right)

The in-domain texts were then split into three sets: a training set (75%), a validation set (12.5%) and an in-domain test set (12.5%). As the BERT-model processes input on the sentence level, the sentence order was randomized. The sentences containing rare multi-word locations and groups were identified and split separately. Each of those splits was later appended to one of the three sets to ensure that each contained entities of every type. The frequencies of the entity types can be found in Table 2 (train and validation split) and in the 'support' column of Table 5 (test split).

To ensure representative testing, the data was augmented with manually annotated test sets from the LASLA corpus in the second part of this paper (see Section 5), both for in-domain prose and out-domain poetry.

### 3.2 Model training and evaluation

We created two models on the Herodotos dataset and compared the results of these models to those obtained using the recently released LatinCy toolkit. The models we trained (finetuned) ourselves are:

- A conditional random field (CRF) model. Erdmann et al. (2016) use a CRF-based baseline in a similar context. This model is fairly simple and will serve as a starting point for comparison.
- LatinBERT (Bamman and Burns, 2020), a specialized BERT model for Latin, trained using the Masked Language Modeling objective on a corpus of 642.7M words, ranging from classical Latin (from 200 BCE onwards) to Neolatin from Wikipedia. We made use of the pre-trained model, and finetuned it on the NER dataset.

The results of these models are compared to LatinCy, a SpaCy pipeline for Latin, and for the LASLA test set (see below) to the Herodotos entity recognizer (Erdmann et al., 2016) as well. In order to train several SpaCy pipelines (Honnibal and Montani, 2017) for Latin (viz. a *small*, *medium* and *large* model), Burns (2023) leveraged the five Latin Universal Dependencies treebanks and several large Latin corpora. LatinCy's named entity recognizers were trained separately from the rest of their respective pipelines, on a custom-made dataset based on the UD treebanks and the dataset of the Herodotos project. For this paper, we tested the *large* ('la_core_web_lg') pipeline, as well as the 'la_core_web_trf' pipeline, which is backed by the multilingual BERT transformer architecture (Devlin et al., 2018).

The next two subsections describe the training setup for our models; section 3.3 discusses the results of the models we trained, as well as a comparison to LatinCy's performance.

#### 3.2.1 CRF

For the CRF model, we made use of an implementation based on CRFsuite (Okazaki, 2007). We specified the optimization method as *l-bfgs*, set the maximum number of iterations to 100 and considered all possible transitions, The following hand-crafted features are incorporated: whether the word is a digit, capitalised or fully upper-cased; whether the word is the first or last word of a sentence; the last three letters; the last two letters; a context window of two left words and two right words. Following Palladino et al. (2020), the whole word itself was not included, because this might aid generalization to other contexts.

Hyperparameter optimization was performed using a 50-fold random search, to optimize the two regularisation coefficients $c1$ (search space exponentially distributed on scale 0.5) and $c2$ (search space exponentially distributed on scale 0.05). The best hyperparameters were 0.183 and 0.086 for $c1$ and $c2$ respectively.

#### 3.2.2 LatinBERT

Prior to the finetuning of LatinBERT, we incorporated the original subword tokenizer into our own, custom tokenizer to ensure the model was fully compatible with the *transformers* library (Wolf et al., 2020). All words were lowercased during tokenization. We proceeded to utilize the *transformers* trainer API both with and without hyperpa-

| Hyperpar. | Initial | Optimized |
|---|---|---|
| Learning rate | 2.00e-5 | 7.89e-5 |
| Weight decay | 0.01 | 0.10 |
| Number of train epochs | 3 | 3 |

Table 3: initial hyperparameters (LatinBERT1) vs. optimised hyperparameters (LatinBERT2)

rameter optimization (results reported under LatinBERT2 and LatinBERT1 respectively). During the experiments with hyperparameter optimization, we specified the optimization method as *random*. The metric for evaluation is the *validation loss*, and the goal is to minimize it based on a ten-fold search. Table 3 provides a comparison of the hyperparameters used. In both cases the per-device train batch size is 16 and the warmup ratio is 0.1.

### 3.3 Results

In Table 4 we report the micro-averaged f1 (or accuracy) based on the token labeling. The micro-averaged f1 computes the proportion of correctly classified observations out of all observations. In Table 5, for every entity type ('PERS', 'LOC', 'GRP'), we report the f1 score (harmonic mean of precision and recall) on the entity level, where the full entity is only considered correct if the annotations for all its comprising tokens match the gold standard exactly, and the macro f1, where the results for each model are averaged across the various labels without taking class size into account. In Appendix A, more detailed counts per label are provided (Table 10).

The overall results in Table 4 show that there is a drop in performance going from in- to out-of-domain, signaling a difficulty to generalize from prose to poetry. Both LatinBERTs outperform the other models in- and out-of-domain. However, it is important to note that optimizing the hyperparameters causes a slight increase in macro-f1 on the in-domain dataset, but a symmetrical, decrease on the out-of-domain dataset. Looking at the entity level metrics in Table 5, 'PERS' is the class that is the easiest to predict for every model. For the models exclusively trained on the Herodotos data (the CRF and LatinBERTs), single word groups are a relatively well-understood category in-domain, but cause problems out-of-domain. Unfortunately, no multi-token 'GRP' were correctly detected, which can be explained by their rarity. Multi-token 'LOC' are also rarely detected, with only the BERT models being able to recognize some in-domain (See again Table 10).

## 4 Error analysis

### 4.1 Ambiguous annotations in the training data

Although guidelines for named entities in classical scholarship exist (Romanello and Najem-Meyer, 2022), for classical Latin texts, they are still lacking (see Section 5). This is reflected in our dataset. We can hypothesize that this impacts the overall performance of the models. In particular, some tokens are annotated as different entities throughout the dataset. In some cases, this is due to the inherent ambiguity of the token, as in the following examples:

- **Homonyms**: *Galli* (genitive singular of 'Gallus', name of a man) as 'PERS' in *Ars am.* 3.334 or 'GRP' in *BGall.* 1.1 ('the Gauls');
- Tokens that occur both as **entity and non-entity** in the dataset: e.g. *Liber* (a divinity, but also 'book'), forms of *Sol* (divinity 'Sun' and the sun), and *Gratia* ('grace', but also the divinity 'Grace') appear both as entities (personifications, usually capitalized) and non-entities (regular use);
- **Patronyms** such as *Atrides* ('descendant of Atreus'): sometimes forms of these refer to one specific person, sometimes to a group.

In other cases, the differences seem to stem from inconsistent annotation choices:

- Multi-token entities that contain a toponym: e.g. the entity *Amphilocho Athenaeo* ('Amphilochus of Athens') in *NH* is annotated both as 'B-PERS B-GRP' and as 'B-PERS I-PERS'; or a building with a name *aedem Larum* ('the temple of the Lares', *NH* 2.5) is annotated as 'O B-GRP', while *aedem Feroniae* ('the temple of Feronia', *NH* 2.56) is annotated as 'B-LOC I-LOC';
- Persons referred to with only a toponym: e.g. *Cressa* ('the Cretan woman', *Ars am.* 1.327) is annotated as 'B-GRP', while *Cynthius* ('the Cynthian', *Ars am.* 2.239) is annotated as 'LOC';
- Unnamed entities annotated in some cases and not in others: e.g. some of the occurrences of

| micro f1 | | CRF | LB1 | LB2 | LatinCy lg | LatinCy trf | support |
|---|---|---|---|---|---|---|---|
| **Caesar/Pliny's (IN)** | BIO-labels | 0.98 | **0.99** | 0.99 | 0.96 | 0.95 | 14,686 |
| | BI-labels | 0.79 | 0.90 | **0.92** | 0.60 | 0.58 | 1,048 |
| ***Ars am.* (OUT)** | BIO-labels | 0.97 | **0.98** | **0.98** | 0.96 | 0.95 | 17,102 |
| | BI-labels | 0.39 | **0.65** | 0.60 | 0.39 | 0.31 | 570 |

Table 4: micro f1 on the Herodotos selected test-set; **LB** stands for LatinBERT

| | | CRF | LB1 | LB2 | LatinCy lg | LatinCy trf | support |
|---|---|---|---|---|---|---|---|
| **Caesar/Pliny's (IN)** | PERS | 0.80 | 0.91 | **0.92** | 0.64 | 0.64 | 474 |
| | LOC | 0.66 | 0.85 | **0.87** | 0.61 | 0.54 | 218 |
| | GRP | 0.74 | 0.89 | **0.91** | 0.02 | 0.06 | 247 |
| | **macro f1** | 0.74 | 0.88 | **0.90** | 0.43 | 0.44 | 939 |
| ***Ars Am.* (OUT)** | PERS | 0.44 | **0.76** | 0.72 | 0.47 | 0.36 | 375 |
| | LOC | 0.30 | **0.43** | 0.38 | 0.28 | 0.18 | 87 |
| | GRP | 0.25 | **0.45** | 0.40 | 0.00 | 0.05 | 107 |
| | **macro f1** | 0.33 | **0.54** | 0.50 | 0.25 | 0.20 | 569 |

Table 5: f1-score per entity type on the Herodotos selected test-set

*prouincia* ('province') and *terra* ('region') are annotated as 'LOC', and some of the occurrences of *equestri* and *praetori* as 'GRP'.

In addition, entire parts of text are not annotated in *Ars am.* and *NH*. The scarcity of data also appears to be a problem: out of the 180 unique tokens that were not correctly identified by any model, 132 do not occur in the training data.

### 4.2 Qualitative analysis LatinBERT

In this section, we perform a qualitative error analysis of the performance of the two best-performing models, LatinBERT1 and 2, on both the in-domain and out-of-domain sets, in order to better understand the origin of the errors. First, LatinBERT1 and LatinBERT2 share common issues, that are generally not encountered by at least one of the other two models:

- **Boundary detection** proves particularly difficult with lists of names: *Lysiae Demosthenen Aeschinen Hyperiden multosque praeterea, Gracchis et Catoni Pollionem Caesarem Caelium [...]* (*Ep.* 1.20.4). Both models correctly identify 4 separate entities in the first part (*Lysiae .. Hyperiden*) but label '*Pollionem Caesarem Caelium*' as one entity. In addition, we find I-labels predicted for entities not occurring after B-label: for instance, both LatinBERTs predict 'I-LOC' for *Memphitidos* ('of Memphis', *Ars am.* 3.393) ('B-GRP' is the gold data) without assigning 'B-LOC' to a previous token.

- Entities with **foreign names** are often predicted as non-entity: e.g. *Adadu*, *Calymne*, *Therapnaeus*, and *Andromeda*.

- Complete sentences with clear entities predicted as non-entities in out-of-domain data (entities in bold): e.g. *Dextra* ***Lebinthos*** *erat siluisque umbrosa* ***Calymne*** | *Cinctaque piscosis* ***Astypalaea*** *uadis* (*Ars am.* 2.81-2) - non-entity predictions for all entities by LatinBERT2; LatinBERT1 only for *Astypalaea*.

LatinBERT1 and LatinBERT2 differ only in the optimization of the hyperparameters, which seems nonetheless to have a relevant impact on the performance. In a total of 223 cases, the prediction of LatinBERT2 differs from LatinBERT1. Table 8 in Appendix A shows that LatinBERT1 slightly outperforms LatinBERT2 on the label 'B-PERS'. However, in several cases, the prediction of LatinBERT2 classifies the category correctly but with wrong segmentation, predicting 'I-PERS' instead of 'B-PERS', whereas LatinBERT1 also classifies incorrectly. In 46 of the cases where only LatinBERT1 is correct, LatinBERT2 predicts a non-

entity. 42 of these tokens did not appear in the train or validation set and the others were either annotated both as entities and 'O' or appeared only once in the training data. Besides this, many differences can be explained by the difficulties in 'GRP'/'LOC' distinction identified in Section 4.1.

## 5 Annotation of the LASLA corpus

In what follows, we discuss the performance of the same NER models on the LASLA Latin corpus.[3] As the LASLA corpus includes a diverse range of classical Latin texts, it represents an interesting test set to investigate the generalisability of the models. With this procedure, we also establish criteria for the annotation of the most problematic classes. In addition, we augment the test set by including both **prose** and **poetry** works (resp. in-domain and out-of-domain) which do not appear in the training data and that belong to different genres with respect to the training data. Overall, this process allows us to reach conclusions on the urgency of guidelines, of data generation, and the generalisability of existing models across different projects.

The portion of the LASLA corpus used for this experiment is composed of 1,738,435 tokens, belonging to 130 Latin literary texts by 21 authors ranging from the 2nd century BCE to the 2nd century CE. It is linked to the LiLa Knowledge Base, an open-ended Knowledge Base of linguistic Linked Data (Passarotti et al., 2020). The URIs for lemmas and tokens provided by the linking are published to ensure interoperability and reusability of the data.[4]

### 5.1 Texts annotated

To evaluate the performance of the models on the LASLA corpus, we annotated texts from three different authors. As in-domain data, we chose to annotate Tacitus' *Historiae* (*Hist.*) book 1 and the first of Cicero's *Orationes Philippicae* (*Phil.*) and for out-of-domain the first three of Juvenal's *Saturae* (*Juv.*). Tacitus and Cicero were selected as 'in-domain' data since they belong to non-fictional prose. Moreover, the *Phil.* are a different genre (oratory) than the Herodotos training data and Tacitus (Historiography and Epistolography). Juvenal's poetry, with its mentions of historical people, was selected to challenge the model, since the out-of-domain testing of Ovid's *Ars am.*, on the contrary, primarily mentions mythical persons. Good performance on these texts would indicate the models' generalisability.

### 5.2 Annotation process and choices

The texts were annotated by two Latin experts using the BIO-format for the entities location, person, and group (see Section 3.1). The Herodotos project annotation was taken as a reference, and the challenging points were discussed in order to address the shortcomings identified in Section 4.1. Cohesion between the annotations of the two experts was guaranteed by joint annotation of 4,463 tokens of the *Saturae* (Juv. 1-3). The Inter-Annotator Agreement (IAA) was calculated using Cohen's Kappa score (Cohen, 1960). The IAA is calculated both including and excluding the label 'O'. The resulting values are 0.87 (incl. 'O') and 0.74 (excl. 'O'). The confusion matrix (excl. 'O') is shown in Figure 1 of the Appendix A. The biggest disagreement concerns the label 'B-GRP'. The difficulties with the annotation of 'GRP' can be divided into two categories: annotation of adjectives derived from toponyms (*Tuscus* - 'Tuscan', *Aegyptius* - 'Egyptian', *Graecus* - 'Greek') and groups of individuals that do not fit the definition of political/ethnic groups as defined by the Herodotos project. Examples of this last category are names of families (e.g *Gracchos* (2.24) - 'The Gracchii'), names used as a generic category (e.g. *Proculas et Pollittas* (2.68) - 'women like Procula and Pollitta'), gods (*Asianorum ... deorum* (3.218) - 'Asian gods'), and other groups such as *Socraticos ... cinaedos* (2.10 - 'Socratic catamites') and *Manes* (2.149 - 'Shades'). For adjectives derived from toponyms, the annotators agreed to use 'GRP' to align with the Herodotos project. For the other categories, 'GRP' is used following the definition of the subcategory 'PER.Group' from the Automatic Content Extraction Guidelines (Consortium, 2008) for any Person entity referring to more than one person. Finally, we chose **not** to annotate nicknames as 'PERS' entities (e.g. *Uenusina ... lucerna* (1.51) - 'The Venusinian light', Horace, was only annotated as 'B-LOC ... O'). Following the first round of joint annotation, an agreement was reached on problematic points to enhance the consistency of the remaining annotation.

[3] https://www.lasla.uliege.be

[4] https://github.com/NER-AncientLanguages/Ner-Latin-RANLP

| micro f1 | | CRF | LB1 | LB2 | LatinCy lg | Herodotos | support |
|---|---|---|---|---|---|---|---|
| **Tac. and Cic. (IN)** | BIO-labels | 0.96 | 0.97 | **0.98** | 0.96 | 0.97 | 15,737 |
| | BI-labels | 0.61 | 0.78 | **0.79** | 0.66 | 0.72 | 1,320 |
| **Juv. (OUT)** | BIO-labels | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 4,399 |
| | BI-labels | 0.45 | 0.48 | 0.50 | **0.51** | 0.48 | 284 |

Table 6: micro f1 on the LASLA corpus; **LB** stands for LatinBERT

| | | CRF | LB1 | LB2 | LatinCy lg | Herodotos | support |
|---|---|---|---|---|---|---|---|
| **Tac. and Cic. (IN)** | PERS | 0.65 | 0.83 | **0.85** | 0.66 | 0.74 | 711 |
| | LOC | 0.31 | 0.51 | **0.55** | 0.53 | 0.49 | 222 |
| | GRP | 0.43 | 0.61 | **0.64** | 0.02 | 0.60 | 154 |
| | **macro f1** | 0.46 | 0.65 | **0.68** | 0.40 | 0.61 | 1,087 |
| **Juvenal (OUT)** | PERS | 0.48 | 0.53 | **0.64** | 0.64 | 0.59 | 143 |
| | LOC | 0.32 | **0.46** | 0.36 | 0.44 | 0.27 | 83 |
| | GRP | 0.47 | 0.40 | **0.52** | 0.00 | 0.23 | 36 |
| | **macro f1** | 0.43 | 0.46 | **0.51** | 0.36 | 0.37 | 262 |

Table 7: f1-score per entity type & macro f1 on the LASLA corpus

### 5.3 Results of running the model

Table 6 shows that when labelling single tokens LatinBERT2 outperforms the other models on in-domain data, whereas the models score very close on out-of-domain data, with LatinCY scoring slightly higher than LatinBERT2.[5] Table 7 shows that LatinBERT2 predicts entire entities better than the other models, except for the category 'LOC' on out-of-domain data, where LatinBERT1 performs better. These results confirm LatinBERT2's general good performance, but also its again somewhat unexpected behavior on poetry.

### 5.4 Error Analysis

#### 5.4.1 Challenging aspects of NER prediction

Similarly as to the Herodotos data, many errors can again be related to the inherent ambiguity of Latin and/or the choices made in annotation (cf. Section 4.1). Both on the in- and out-of-domain LASLA data, errors were made that are related to ambiguous tokens that occur both as entity and non-entity, albeit slightly more present in out-of-domain, e.g. *Pax atque Fides, Uictoria, Uirtus* ('The Goddesses Peace, Faith, Victory and Virtue', Juv. 1.115). Also for the LASLA test-set, tokens annotated differently across the Herodotos training data result in multiple errors. For instance, non-capitalized forms of *prouincia* and *urbs* are annotated as 'LOC' in the training data only when they refer to a precise location. Likewise, *princeps* and *imperator* are annotated as 'PERS' only where they refer to specific emperors. Lastly, words like *domus* and *aedes* are sometimes annotated when they indicate a specific location: for example, *aede Apollinis* - 'the temple of Apollo' and *Tiberianam domum* - 'the palace of Tiberius'. Even though the Herodotos training data are not fully consistent in these annotations, the LASLA annotation did strictly follow these guidelines, which highlighted the inconsistent behavior of models with respect to these points.

#### 5.4.2 Qualitative analysis LatinBERT on the LASLA dataset

In Section 4.2 we observed that LatinBERT1 and LatinBERT2 share common issues, that are generally not encountered by at least one of the other two models. On the LASLA corpus, similar and additional observations can be made. **Boundary**

[5]The major increase in performance of LatinCy on the LASLA data can be explained by two reasons: first, 38% of total errors of LatinCy concern the GRP-entities, of which there are relatively less in the LASLA test data (23.5% of the total entities are 'GRP's in Herodotos, whereas in the LASLA 14.1%); second, many other errors are caused by the tendency of LatinCy to predict entities for any and all capitalized words. In the Herodotos data, all sentences start with a capital, creating many errors for LatinCy; in the LASLA, capitalization is absent, hence such errors do not occur.

**detection issues** occur in comparable instances on the LASLA corpus, such as predicting separate entities in lists and predicting I-labels for entities not occurring after B-label. However, an additional boundary complication occurs in poetry in difficult nested cases such as the entity *Cecropiam ... Cotyton* (Juv. 1.7-9) separated by the entity *Baptae* occurring in between (this created the annotation 'B-PERS B-GRP I-PERS'). Both LatinBERTs predict a non-entity for *Baptae* and *Cotyton*. As in the Herodotos test set, **foreign names** again proved particularly difficult, in the LASLA out-of-domain especially those with a Greek accusative ending in 'n' (e.g. *Euphraten* (Juv. 1.104). Of the 10 tokens with this ending only Deucalion (1.81) is predicted correctly as an entity by LatinBERT1.[6] Lastly, in the out-of-domain data we again find **complete sentences** that contain multiple entities for which non-entities are predicted.

A close analysis of the performance on tokens where the manual annotation differed shows some additional challenging categories. Of the 69 tokens where the manual annotation differed, LatinBERT1 got 39 wrong (accounting for 20.5% of its total errors), and LatinBERT2 got 41 wrong (accounting for 22.5% of its total errors). For instance, both LatinBERTs predict 'O' for most **groups of individuals** that did not fit the political/ethnical 'GRP' category, except for some family names (e.g. *Catuli*, *Fabii*). For **Literary works** identified by a personal name, another category where the annotators disagreed but were eventually not annotated, LatinBERT2 predicts an entity but LatinBERT1 'O' (e.g. *Theseide* (1.2); *Heracleas | aut Diomedeas* (1.52-3)). Lastly, for the category of persons referred to with only a toponym, also identified as an issue in Section 4.1, we annotated 'LOC' but the LatinBERTs predicted 'GRP': e.g. *non Maurus erat neque Sarmata nec Thrax* ('it was not a Moroccan nor a Sarmatian nor a Thracian', 3.79).

The comparison between the two LatinBERTs shows that on the in-domain LASLA data, LatinBERT2 outperforms LatinBERT1, especially on I-labels (cf. Appendix A, Table 9). When considering I-label errors, both LatinBERTs classify the category correctly for more than half of these errors (40 out of 78 for LatinBERT1; 32 out of 62 for LatinBERT2), but wrongly assign the 'B-' label: the problem thus lies again with the **boundary detection**. On the out-of-domain data, LatinBERT2 outperforms LatinBERT1 in the 'B-PERS' category. As on the Herodotos project test data, in the majority cases where only LatinBERT1 is correct, LatinBERT2 predicts a non-entity: for the in-domain set 22 out of 27 total cases concern words absent from the train/validation set, for out-of-domain 16 out of 18.

This analysis confirmed that the categories identified in Section 4.2 are difficult for NER. It also emphasised the differences between in- and out-of-domain data: models only trained on prose perform worse on poetry due to stylistic and thematic differences.

## 6 Conclusions and future work

The process of training two new models on existing data, comparing their results on previously and newly annotated data, and comparing their performance to existing models allows us to draw several conclusions. First, the good performance of LatinBERT1 and 2 demonstrates the interest of applying **transformer-based models** for the NER task on Latin. Especially for the category 'PERS' the two models yield satisfactory results. However, the analysis of the annotations and the errors has shown that the **development of guidelines** is crucial to ensure the consistent annotation of datasets that can be reused as training- and test-sets across different projects and for different models. In addition, the significantly worse performance of the models on poetry indicates the need for training data for this specific type of texts. Future work should also consider improving the preprocessing and normalization of training data (e.g. harmonizing the use of the 'v/u' 'i/j' pairs), and testing the use of multilingual BERT models that include Latin (mBERT, XLM-Roberta) (Sprugnoli et al., 2022; Nehrdich, 2022). Likewise, additional linguistic information available in the LASLA corpus (e.g. lemmatization and PoS tagging) might improve the results of the NER. Finally, after we establish a system for Named Entity Disambiguation employing information from existing extensive resources, we will explore the potential of mutual reinforcement, i.e. we will consider whether results from one system can improve the other and vice-versa as argued by Kolitsas et al. (2018).

[6] This is particularly surprising since in the Herodotos test-set LatinBERT1 correctly predicted 29 out of 40 of such forms, and LatinBERT2 22.

## References


David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. (arXiv:2009.10053). ArXiv:2009.10053 [cs].

Patrick J. Burns. 2023. Latincy: Synthetic trained pipelines for latin nlp. (arXiv:2305.04365). ArXiv:2305.04365 [cs].

William L. Carey. s.d. The latin library. Accessed: July 1st, 2023.

Pierre Chastang, Sergio Torres Aguilar, and Xavier Tannier. 2021. A Named Entity Recognition Model for Medieval Latin Charters. *Digital Humanities Quarterly*, 015(4).

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Linguistic Data Consortium. 2008. ACE (Automatic Content Extraction) English: Annotation Guidelines for Entities.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. (arXiv:2109.11406). ArXiv:2109.11406 [cs].

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, page 423–446, Cham. Springer International Publishing.

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, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2023. Herodotos-project-latin-ner-tagger-annotation.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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*, page 20–29, Online. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. In *Computational Natural Language Learning*, pages 519–529. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Wouter Mercelis and Alek Keersmaekers. 2022. An ELECTRA model for Latin token tagging tasks. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 189–192, Marseille, France. European Language Resources Association.

Sebastian Nehrdich. 2022. SansTib, a Sanskrit - Tibetan parallel corpus and bilingual sentence embedding model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6728–6734, Marseille, France. European Language Resources Association.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Chiara Palladino, Farimah Karimi, and Brigitte Mathiak. 2020. Ner on ancient greek with minimal annotation. *https://dh2020.adho.org/*.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Matteo Romanello and Sven Najem-Meyer. 2022. Guidelines for the annotation of named entities in the domain of classics.

David A. Smith, Jeffrey A. Rydberg-Cox, and G. Crane. 2000. The perseus project: a digital library for the humanities. *Literary and Linguistic Computing*.

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*, page 1–44.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Sergio Torres Aguilar. 2022. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, page 119–128, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Krzysztof Wróbel and Krzysztof Nowak. 2022. Transformer-based part-of-speech tagging and lemmatization for Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 193–197, Marseille, France. European Language Resources Association.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. 2023. 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*, page 175–182, Dubrovnik, Croatia. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2022a. *Transformer-Based Named Entity Recognition for Ancient Greek*.

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.

# A Appendix

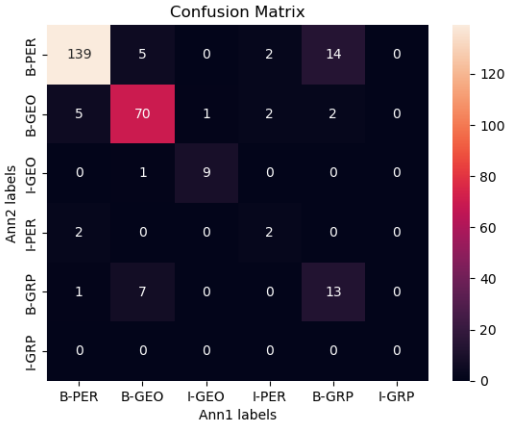


Figure 1: IAA on Juv. *Saturae* 1-3, label 'O' excluded

| **Gold label** | **1 & 2 wrong** | **1 correct** | **2 correct** |
|---|---|---|---|
| O | 0 | 11 | 25 |
| B-PERS | 13 | 47 | 24 |
| I-PERS | 2 | 1 | 1 |
| B-LOC | 14 | 12 | 14 |
| I-LOC | 0 | 0 | 2 |
| B-GRP | 18 | 16 | 12 |
| I-GRP | 1 | 0 | 0 |
| Total | 58 | 87 | 78 |

Table 8: Comparison of differences in prediction between LatinBERT1 (1) and LatinBERT2 (2) on the Herodotos data.

| **Gold label** | **1 & 2 wrong** | | **1 correct** | | **2 correct** | |
|---|---|---|---|---|---|---|
| | IN | OUT | IN | OUT | IN | OUT |
| O | 0 | 0 | 7 | 5 | 7 | 7 |
| B-PERS | 1 | 6 | 14 | 8 | 20 | 22 |
| I-PERS | 2 | 3 | 4 | 0 | 14 | 0 |
| B-LOC | 2 | 7 | 6 | 15 | 11 | 9 |
| I-LOC | 3 | 3 | 0 | 0 | 5 | 0 |
| B-GRP | 14 | 3 | 6 | 4 | 10 | 5 |
| I-GRP | 1 | 0 | 0 | 0 | 0 | |
| Total | 23 | 22 | 37 | 32 | 67 | 43 |

Table 9: Comparison of differences in prediction between LatinBERT1 (1) and LatinBERT2 (2) on in and out-of-domain LASLA data.

| | | CRF | LB1 | LB2 | LatinCy lg | LatinCy trf | support |
|---|---|---|---|---|---|---|---|
| **Caesar/Pliny's (IN)** | B-PERS | 0.83 | 0.93 | 0.94 | 0.75 | 0.73 | 474 |
| | I-PERS | 0.86 | 0.89 | 0.91 | 0.43 | 0.52 | 98 |
| | B-LOC | 0.70 | 0.87 | 0.90 | 0.64 | 0.56 | 218 |
| | I-LOC | 0.00 | 0.33 | 0.50 | 0.00 | 0.00 | 8 |
| | B-GRP | 0.77 | 0.90 | 0.92 | 0.02 | 0.06 | 247 |
| | I-GRP | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3 |
| ***Ars Am.* (OUT)** | B-PERS | 0.48 | 0.76 | 0.72 | 0.47 | 0.36 | 375 |
| | I-PERS | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 1 |
| | B-LOC | 0.30 | 0.43 | 0.38 | 0.28 | 0.18 | 87 |
| | B-GRP | 0.25 | 0.45 | 0.40 | 0.00 | 0.05 | 107 |

Table 10: f1-score per entity type on the Herodotos dataset; **LB** stands for LatinBERT

| | | CRF | LB1 | LB2 | LatinCy lg | Herodotos | support |
|---|---|---|---|---|---|---|---|
| **Tac. and Cic. (IN)** | B-PERS | 0.71 | 0.88 | 0.89 | 0.84 | 0.81 | 711 |
| | I-PERS | 0.78 | 0.81 | 0.85 | 0.24 | 0.79 | 188 |
| | B-LOC | 0.33 | 0.58 | 0.60 | 0.57 | 0.52 | 222 |
| | I-LOC | 0.00 | 0.00 | 0.18 | 0.00 | 0.13 | 42 |
| | B-GRP | 0.43 | 0.61 | 0.62 | 0.03 | 0.60 | 154 |
| | I-GRP | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3 |
| **Juv. (OUT)** | B-PERS | 0.55 | 0.55 | 0.65 | 0.66 | 0.64 | 143 |
| | I-PERS | 0.23 | 0.00 | 0.00 | 0.00 | 0.19 | 7 |
| | B-LOC | 0.35 | 0.50 | 0.39 | 0.44 | 0.27 | 83 |
| | I-LOC | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 14 |
| | B-GRP | 0.47 | 0.40 | 0.52 | 0.00 | 0.23 | 36 |
| | I-GRP | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1 |

Table 11: f1-score per entity type on the LASLA corpus