EvaLatin 2022

-

Guidelines

Version 1.0

March 16, 2022

Rachele Sprugnoli[1], Flavio M. Cecchini[1], Margherita Fantoli[2], Marco Passarotti[1]
1. CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milan, Italy
rachele.sprugnoli[AT]unicatt.it
marco.passarotti[AT]unicatt.it
flavio.cecchini[AT]unicatt.it
2. KU Leuven
Oude Markt 13, 3000 Leuven, Belgium
margherita.fantoli@kuleuven.be

Contents

1 Introduction **5**

2 Data **7**
2.1 Data Format 8
2.2 Training Data 9
2.3 Test Data 10

3 Tasks and Sub-tasks **13**
3.1 Tasks 13
3.1.1 Lemmatization 13
3.1.2 Part-of-Speech Tagging 15
3.1.3 Morphological Features 20
3.2 Sub-tasks 25

4 Evaluation **27**

5 How to Participate **29**
5.1 Submitting Runs 29
5.2 Writing the Technical Report 30

A Selection of Resources for Latin **31**

B Types Affected by a Bug **33**

Chapter 1

Introduction

EvaLatin 2022 is the second edition of the evaluation campaign of Natural Language Processing (NLP) tools for the Latin language. The campaign is designed with the aim of answering two main questions:

- How can we promote the development of resources and language technologies for the Latin language?
- How can we foster collaboration among scholars working on Latin and attract researchers from different disciplines?

EvaLatin 2020 [7] was organized around 2 tasks (i. e. Lemmatization and PoS Tagging); for EvaLatin 2020 we instead propose 3 tasks (i. e. Lemmatization and PoS Tagging, and Identification of Morphological Features), each with 3 sub-tasks (i. e. Classical, Cross-Genre, Cross-Time). Shared data and a scorer are provided to the participants. Participants can choose to participate in either one or all tasks and sub-tasks. The organizers rely on the honesty of all participants who might have some prior knowledge of part of the data used for evaluation not to unfairly use such knowledge.

EvaLatin 2022 is organized within the "Second Workshop of Language Technologies for Historical and Ancient Languages" (LT4HALA 2022), colocated at LREC 2021[1]. The workshop will be held in Marseille, France. EvaLatin 2022 is organized by the CIRCSE research centre at the Università Cattolica del Sacro Cuore in Milan, Italy, in the context of the *LiLa: Linking Latin* ERC project[2]. An agreement has been established with the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) of the University of Liège, Belgium, for the use of the homonymous corpus, and a collaboration has been set up with the Katholieke Universiteit Leuven, Belgium.

For any update, please check the LT4HALA 2022 website: `https://circse.github.io/LT4HALA/2022/`.

[1] `https://lrec2022.lrec-conf.org/en/`

[2] `https://lila-erc.eu/`. The LiLa project has received funding from the European Research Council (ERC) under the European Union's *Horizon 2020* research and innovation programme – Grant Agreement No. 769994.

Chapter 2

Data

The data set of EvaLatin 2022 consists of texts mainly taken from the LASLA corpus [4], a resource manually annotated since 1961 by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège[1], Belgium, which are then converted into the annotation formalism of the Universal Dependencies (UD) project[2] [3], which is the one used by this evaluation campaign.

The LASLA corpus contains approximately 1,700,000 words (punctuation is not present in the corpus), corresponding to 133,886 unique tokens and 24,339 unique lemmas. The data are based on the editions available at the time, but the final texts are determined by the annotators. Each token is annotated by a trained classicist, and usually the same annotator consistently takes care of a set of associated texts. The annotation takes place through a web-based interface where the annotator chooses between a set of possible analyses or adds a new analysis when necessary. To minimize human errors, a sentence cannot be validated if not every token has been processed. At the end of such procedure, an index of forms and associated morphological analyses is generated and subsequently corrected by the annotator. Finally, a second philologist verifies and corrects the final version, and the most complicated cases are discussed within the LASLA team. The guidelines for annotation are provided by the manual [6].
To these texts from the LASLA corpus, in the test data we add another text annotated by members of the CIRCSE research center.

The conversion from the original fixed-length format of LASLA to the CoNLL-U format and the UD formalism has also been developed internally to the CIRCSE research center and is based on Python[3] scripts complemented by the access to the LiLa lexical knowledge base [5]. The conversion is then followed by a further step of uniformization to make all annotated texts (including those not taken from the LASLA corpus) as coherent as possible between themselves and with respect to the the UD formalism; during this process, in the framework of EvaLatin 2022, only a subset of morpholexical features are retained.

[1] http://web.philo.ulg.ac.be/lasla/textes-latins-traites/
[2] www.universaldependencies.org
[3] https://www.python.org/

Overall, the performed conversion and uniformization are not only a transcription into a different annotation system, but also an adjustment to the annotation principles that in the last years have been under constant development for Latin treebanks in the framework of the UD project, and which might differ in some point from the those of the LASLA corpus, or extend them. Chapter 3 will sketch these implementations and adjustments for the different annotation layers considered by EvaLatin 2022, i. e. lemmas, parts of speech and morphological features.

2.1 Data Format

Training data are distributed in the CoNLL-U format[4]. Following such format, the annotations are plain text files having the `conllu` extension and encoded in UTF-8 containing:

- 2 comment lines starting with a hashtag (`#`): one line specifies the sentence unique code (`# sent_id`), while the other line reports the sentence string (`# text`).
- Lines composed of 10 tab-separated fields and containing either the UD-style annotation of a single-word token, or denoting a multi-word token. When values for a given field are not determined, an underscore (`_`) is used instead (there are no empty fields). The 10 fields of the CoNLL-U format are as follows:
  1. `ID`: numerical index of the word in the sentence. For single-word tokens, a progressive integer starting from 1 for each new sentence; for multi-word tokens, it is a range in the form of $i-(i+n)$, meaning that the token in the sentence string is split into $n+1$ words, with indices from i to $i+n$, each with its own analysis;
  2. `FORM`: word form as occurring in the sentence string. For multi-word tokens, the form of a single word is normalized, whereas it might be contracted, or otherwise subject to changes, in the univerbation (e. g. *imprimis* would be decomposed as *in* and *primis*);
  3. `LEMMA`: form conventionally representing the lexeme related to the word form;
  4. `UPOS`: universal PoS tag[5] of the word;
  5. `XPOS`: language-specific PoS of the word (as used e. g. in the same treebank prior to conversion to the UD annotation style);
  6. `FEATS`: list of morpholexical features of the word, either "universal" or "language-specific"[6];
  7. `HEAD`: index of the syntactic head of the word (possibly `0` if the word is the sentence root);

[4] `https://universaldependencies.org/format.html`
[5] `https://universaldependencies.org/u/pos/all.html`
[6] `https://universaldependencies.org/ext-feat-index.html`

8. DEPREL: universal syntactic dependency relation with respect to the HEAD[7];
9. DEPS: syntactic relations in the enhanced dependency graph;
10. MISC: any other annotation.

- Blank lines marking boundaries between sentences and the end of the document.

In our data set, ID, FORM, LEMMA, UPOS and FEATS are the only annotated fields: all other fields are filled in with underscores.

An example of the data format is given in Figure 1. This format is used for the training data and participants are expected to produce the same format for the final evaluation.

```
# sent_id = CaesBG4-A-01-607
# text = neque multum frumento sed maximam partem lacte atque pecore uiuunt multumque sunt in uenationibus
1   neque   neque   CCONJ   _   _   _   _   _   _
2   multum  multum  ADV     _   _   _   _   _   _
3   frumento    frumentum   NOUN    _   Case=Abl|InflClass=IndEurO|Number=Sing   _   _   _   _
4   sed sed CCONJ   _   _   _   _   _   _
5   maximam magnus  ADJ     _   Case=Acc|Degree=Abs|InflClass=IndEurA|Number=Sing   _   _   _   _
6   partem  pars    NOUN    _   Case=Acc|InflClass=IndEurI|Number=Sing   _   _   _   _
7   lacte   lac NOUN        _   Case=Abl|InflClass=IndEurI|Number=Sing   _   _   _   _
8   atque   atque   CCONJ   _   _   _   _   _   _
9   pecore  pecus   NOUN    _   Case=Abl|InflClass=IndEurX|Number=Sing   _   _   _   _
10  uiuunt  uiuo    VERB    _   Aspect=Imp|InflClass=LatX|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act   _   _   _   _
11-12   multumque  _   _   _   _   _   _   _   _
11  multum  multum  ADV     _   _   _   _   _   _
12  que que CCONJ   _   _   _   _   _   _
13  sunt    sum AUX     _   Aspect=Imp|InflClass=LatAnom|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   _   _   _   _
14  in  in  ADP     _   _   _   _   _   _
15  uenationibus    uenatio NOUN    _   Case=Abl|InflClass=IndEurX|Number=Plur   _   _   _   _
```

Figure 1: Example of the data format.

2.2 Training Data

Texts provided as training data are the same ones adopted as training and test data for EvaLatin 2020; however, the annotation may slightly differ from that seen in the previous edition of the evalutation campaign. In fact, in 2020 we did not use the LASLA corpus directly, but instead worked with a manually revised version of the automatic annotation performed by UDPipe [8] based on a model trained on the Perseus UD Latin Treebank[8] [1].

Texts are by 5 Classical authors for a total of more than 300,000 tokens: Caesar, Cicero, Seneca, Pliny the Younger and Tacitus. Each author is represented by one specific text genre: treatises in the case of Caesar, Seneca and Tacitus, public speeches for Cicero, and letters for Pliny the Younger. Table 1 presents details about the training data set of EvaLatin 2022.

N.B. On March 17th 2022 we released a new version of the training data: in the previous version some tokens were misspelled due to a bug in the script used for the

[7] https://universaldependencies.org/ext-dep-index.html
[8] https://github.com/UniversalDependencies/UD_Latin-Perseus/

conversion of the LASLA annotation into the CoNLL-U format (72 types - 1,109 tokens). Misspelled types are reported in Appendix B and the corresponding tokens will not be taken into consideration in the evaluation (see Section 4).

AUTHORS	**TEXTS**	**# TOKEN**
Caesar	De Bello Gallico	44,818
Caesar	De Bello Civili (books I and II)	17,287
Cicero	Philippicae (books I-XIV)	52,563
Cicero	In Catilinam	12,564
Pliny the Younger	Epistulae (books I-VIII and X)	60,695
Seneca	De Beneficiis	45,457
Seneca	De Clementia	8,172
Seneca	De Vita Beata	7,270
Seneca	De Providentia	4,077
Tacitus	Historiae	51,420
Tacitus	Agricola	6,737
Tacitus	Germania	5,513
TOTAL	**TEXTS**	**316,573**

Table 1: Training data of EvaLatin 2022

2.3 Test Data

Tokenization is a central issue in evaluation and comparison, because each system could apply different tokenization rules leading to different outputs. In order to avoid this problem, test data will be already provided in tokenized format, one token per line, and with a white line separating each sentence. An example of test data format is given in Figure 2.

Test data contain only the tokenized words but not the correct tags, that have to be added by the participant systems to be submitted for the evaluation. The gold standard test data, i. e. the annotation used for the evaluation, will be provided to the participants after the evaluation. The composition of the test dataset for the *Classical* sub-task is given in Table 2. Details for the data distributed in the Cross-Genre and Cross-Time sub-tasks are reported in Tables 3 and 4 respectively.

AUTHOR	**TEXT**	**# TOKENS**
Livius	Ab Urbe Condita (book VIII)	13,572

Table 2: Test data for *Classical* sub-task.

```
# sent_id = 1
# text = Quaesisti a me Lucili quid ita si prouidentia
mundus regeretur multa bonis uiris mala acciderent
1   Quaesisti   _   _   _   _   _   _   _   _
2   a   _   _   _   _   _   _   _   _
3   me  _   _   _   _   _   _   _   _
4   Lucili  _   _   _   _   _   _   _   _
5   quid    _   _   _   _   _   _   _   _
6   ita _   _   _   _   _   _   _   _
7   si  _   _   _   _   _   _   _   _
8   prouidentia _   _   _   _   _   _   _   _
9   mundus  _   _   _   _   _   _   _   _
10  regeretur   _   _   _   _   _   _   _   _
11  multa   _   _   _   _   _   _   _   _
12  bonis   _   _   _   _   _   _   _   _
13  uiris   _   _   _   _   _   _   _   _
14  mala    _   _   _   _   _   _   _   _
15  acciderent  _   _   _   _   _   _   _   _
```

Figure 2: Example of the test data format.

AUTHORS	TEXTS	# TOKENS
Pliny the Elder	Naturalis Historia (book XXXVII)	11,371
Ovidius	Metamorphoseon (books IX-X)	11,325
TOTAL	TEXTS	22,696

Table 3: Test data for *Cross-genre* sub-task.

AUTHOR	TEXT	# TOKENS
Sabellicus	De Latinae Linguae Reparatione	9,278

Table 4: Test data for *Cross-time* sub-task.

Chapter 3

Tasks and Sub-tasks

Participants can choose to participate in either one or all tasks and sub-tasks described in this Chapter.

3.1 Tasks

This Section provides details on the three tasks included in EvaLatin 2022.

3.1.1 Lemmatization

Lemmatization is the process of transforming each word form into a corresponding conventional "base form", according to its part of speech and etymology, which usually coincides with an entry found in the dictionary (i. e. lemma). The criteria followed in the EvaLatin 2022 corpus are summarized below:

- lemmas are written in lowercase characters only;
- the Latin letters *v* and *j* are never used, *u* and *i* are used instead;
- verbs (`VERB`/`AUX`) are lemmatized under the first person singular form of the indicative imperfective present active (or passive, if the active is not available, as for deponent verbs): e.g., *accingere* → *accingo*; in rare cases, the corresponding perfective form might be used, e. g. *memini*, or the third person for impersonal verbs, e. g. *libet*;
- nominal, normally inflectable classes (`ADJ`, `DET`, `NUM`, `NOUN`, `PRON`) are lemmatized under their nominative singular (or plural, if the singular is not available) form, e. g. *senatui* → *senatus*, under a form corresponding to the masculine gender for variable classes (`ADJ`, `DET`, `NUM`, some `PRON`), and degree-less ("positive") if gradable (most `ADJ`), e. g. *pulcherrimorum* → *pulcher*;
- all other invariable or indeclinable words (including adverbs, `ADV`, and non-analyzed tokens with PoS `X`) take their form as lemma, e. g. *cum* → *cum*, μελετην → μελετην,

possibly a degree-less ("positive") form if gradable (some `ADV`), e. g. *amplius* → *ample*, and can sometimes receive a normalization, as happens for inflectable words, e. g. *ac*, *atque* → *atque*;

- suppletive forms are traced back to the respective form satisfying the above criteria (if it exists) in their paradigms, disregarding different etymologies, e. g. *optimam* → *bonus*;
- abbreviations are expanded, they are considered just possible graphical variants of a lexeme: e. g. *Tib* → *tiberius*, *L* → *lucius*, *hs* → *sestertius*;
- the lemma of Roman numerals, like other invariable words, is identical to their form, but coherently lowercased: e. g. *XVII* → *xvii*;
- *lacunae* and *cruces* (corrupted passages with no sensible interpretation) are assigned no lemma at all: e. g. *p..* → _, *sententis* → _(Seneca, *De Vita Beata*);

Please note that the data use Unicode/UTF-8 encoding, so every character is reproduced in its original, intended form, without using "textual encodings": this means in particular that Greek words appear in the Greek alphabet, complete of all possible diacritics, and not in Beta Code[1], and that we have e. g. $\overline{\text{C}}$ '100,000' instead of *Ca* (as in the LASLA corpus).

A note about tokenization

In the LASLA corpus, some space-separated word sequences are considered as a single token, with a unique (univerbated) lemmatization and morphological interpretation, e. g. |*re publica*|, |*quam ob rem*|, sometimes even in the presence of other interposed words between their components. This is not possible (and not even sensible) in the CoNLL-U format and the UD formalism, especially since Latin (contrary e. g. to Vietnamese) has no codified rule nor interpretation for space-separated words. Consequently, in the EvaLatin 2022 corpus such sequences are simply tokenized as they occur in the text, and each resulting word receives an individual appropriate lemmatization, part-of-speech tagging and morphological analysis. This also means that there will possibly be both occurrences of e. g. *quam* | *ob* | *rem* as three separate tokens with different lemmas (*qui* | *ob* | *res*), parts of speech (`DET` | `ADP` | `NOUN`) and morphological features, alongside occurrences of *quamobrem* as a unique token with lemma *quamobrem* and part of speech `SCONJ` (which entails no morphological features). This ultimately depends on upstream editorial choices. If needed, a unitary treatment of such multiword expressions might be achieved through the annotation of syntactic dependency relations, but these are not present in our corpus.

Conversely, there are tokens which are split into two or more words. There are two kinds: tokens containing clitic elements (e. g. *nequitiaeque* → *nequitiae* | *que*), and tokens with a complex internal structure which are split to achieve annotational coherence in

[1] `http://stephanus.tlg.uci.edu/encoding.php`

the data (e. g. *rempublicam* → *rem* | *publicam*). In both cases, they are represented in the CoNLL-U format by means of so-called multi-word tokens[2]. This preserves the original orthographic form, allowing at the same time for a complete linguistic analysis. We note that in the original LASLA corpus clitics already receive an autonomous annotation, but are not always tokenized separately, e. g. *timorist* (→ *timori* | *est*) vs. *uillas* | *que* (← *uillasque*).

As for tokens that receive a unitary analysis in the LASLA corpus but are split in the EvaLatin 2022 data set, two main criteria are followed. The first is the desire to have an as coherent as possible annotation across all our sources: so, for example, if one source decomposes *etsi* as *et* | *si*, we are bound to replicate this analysis for all other sources, too. The second, more typological reason is that some conventionally univerbated words do warrant an internal analysis if the components are morphologically independent enough and/or their syntactic roles are different. We have the former case for *rempublicam* → *rem* | *publicam*, as the two components follow parallel and distinct inflections, and can be further analyzed with distinct lemmas and parts of speech (which does not happen e. g. for *princeps*, etymologically from *primus* and *capio*). The latter case is seen for *siquidem* → *si* | *quidem*, where *si* is a `SCONJ` introducing the subordinate clause and *quidem* a discoursive `PART`, and both would syntactically depend on an external element (the predicate). On the other hand, we leave a token like *quamobrem* `SCONJ` as it is, even if *quam* | *ob* | *rem* is simultaneously attested in the corpus, since a) its internal analysis would return us a not-so-informative self-contained phrase and b) the whole block acts as a unit where the contributions of the single components are blurred, and further, the components are usually crystallized in their forms (i. e. there is no such pattern as the parallel inflections in *res* | *publica*).

3.1.2 Part-of-Speech Tagging

In the Part-of-Speech (PoS) Tagging task, systems are required to assign a morphosyntactic category (PoS tag) to each token. The universal PoSs tags[3] are detailed in the following. A key point to take into consideration is that, besides the morphosyntactic (and partly semantic) categorization they represent, in UD there is also a fundamental distinction between lexical and functional word classes[4], which motivates the existence of couplets such as `ADJ`/`DET`, `NOUN`/`PRON`, `VERB`/`AUX` (the latter language-specific to Latin).

- `ADJ`: adjectives. Class of lexical relational nominal words that specify properties or attributes of some explicit or implicit referent, e. g. *inopinantes* 'unaware', *pulchra* 'beautiful'. Their relationality in Latin is realized by agreement: adjectives take on case (`Case`) and number (`Number`) of their referents and possibly select an inflectional paradigm (`InflClass`) according to their gender. Many adjectives are also morphologically gradable, in that they can express a comparative (`Cmp`) or absolute superlative (`Abs`) `Degree`. Their functional counterpart are determiners (`DET`) and

[2] `https://universaldependencies.org/format.html#words-tokens-and-empty-nodes`
[3] `https://universaldependencies.org/u/pos/index.html`
[4] `https://universaldependencies.org/u/overview/syntax.html#the-primacy-of-content-words`

their specialized subclass of numerals (NUM). Ordinal numerals, though, e. g. *tertia* 'third', are classified as ADJs. Names of populations (e. g. *romanus*) should also usually be ADJs, not PROPNs or NOUNs, as generically adjectives derived from proper names (e. g. *caesareus* or *caesarianus* from *Caesar*).

- ADP: adpositions. Class of functional words usually introducing noun phrase arguments of the clause, but also connecting subordinate clauses headed by nonfinite verb forms. Adpositions are indeclinable in Latin. Classical Latin has a prevalence of prepositions (i. e. occurring before their nominal phrase), but also postpositions occur. Examples: *ad*, *in*, *cum* (acting as a postposion in fixed expressions like *mecum* 'with me').

- ADV: adverbs. Broad class of lexical words modifying (adverbs of manner, degree) or focusing other words in a clause, and also standing for oblique arguments of a predicate (adverbs of cause, place, or time). Adverbs are indeclinable in Latin. They appear as such (e. g. *semper* 'always') or are often formed by means of more or less productive derivational processes from adjectives (e. g. *pulchre* 'beautifully', from *pulcher* 'beautiful'), determiners (e. g. *hic* 'here', from *hic* 'this (one)'), numerals (e. g. *centiens* 'a hundred times', from *centum* 'one hundred'), nouns (e. g. *gregatim* 'in flocks', from *grex* 'herd') and verbs (e. g. *absolute* 'absolutely', from the perfective passive participle of *absoluo* 'I release'). Adverbial forms of adjectival forms retain the possibility to be graded (with Degree either Cmp or Abs); other adverbs might show a particular Degree (e. g. *plus* 'more'). In UD, the lemma of an adverb coincides with its (usually positive, i. e. Degree-less) form, and derived adverbs are not marked for the features of their bases.

- AUX: auxiliaries. Class of functional words that combine with other nominal or verbal forms to form a predicate, often allowing the language to express a set of grammatical categories that are not expressed by those other forms. In Classical Latin, there is only one auxiliary, *sum* 'I am', which has the morphology of a verb and acts as a copula (e. g. *domus pulchra* ***est*** 'the house is beautiful'), an existential/locative (e. g. ***est*** *gladiator in arena* 'there is a gladiator in the arena') or allows for periphrastic conjugations when there is no synthetic form available (e. g. *puer amatus* ***erat*** 'the boy had been loved'; there are no synthetic perfective passive forms in Latin). The verb *eo* 'I go' is also considered an auxiliary in the LASLA corpus; however, it is only used in its imperfective passive infinitive form together with the supine in the peculiar so-called "periphrastic future passive infinitive", as in *[putabam] tibi hanc redditurum* ***iri*** '[I thought] it (i. e. a letter) would have been returned to you'

- CCONJ: co-ordinating conjunctions. Class of functional words that act as connectors between any kind of syntactically equivalent phrases or clauses while keeping them all at the same level of grounding in the sentence. Co-ordinating conjunctions are indeclinable in Latin. Examples: *et*, *atque*, *uel*.

- `DET`: determiners. A class of relational nominal words that are the functional counterparts of adjectives (`ADJ`). They have no full lexical content and are limited to expressing deictic properties. Determiners usually agree with their referents and inflect like adjectives do. Sometimes they can express `Degree`, but are not normally gradable, given they have no lexical content that can be graded. In traditional grammars, they are called pronouns, adjectival pronouns or pronominal adjectives. Examples: *nostros* 'our(s)', *illud* 'that (one)', *eiusmodi* 'of this kind' (indeclinable).

- `INTJ`: interjections. Disparate class of words with expressive and discoursive functions, which are often used as exclamations and are not syntactically bound to the rest of the sentence. Proper interjections tend to be onomatopoeic like *heu* or *ai*, but the LASLA corpus also registers some noun or verbal phrases as such, e. g. *mehercle* 'by Hercules!' (lit. '[may] Hercules [assist] me'), *agedum* 'come on now!' (lit. 'move yet').

- `NOUN`: nouns. Class of lexical, nonrelational words that refer to a person, place, thing, animal, idea or entity in general. Proper nouns are actually a subclass of nouns, but in UD they are treated by means of a different tag, `PROPN`, so that the tag `NOUN` is reserved for what are conventionally labeled as "common nouns". Nouns in Latin always have an inherent grammatical gender (`Gender`, not shown in our data, see §3.1.3) and usually express `Case` and `Number` following an inflectional paradigm (`InflClass`): these two traits are often correlated, but ultimately independent from each other. Nouns are also not gradable by their nature. The functional counterpart of nouns are pronouns (`PRON`). In the LASLA corpus, many nouns of populations are considered to be "proper" (and thus written with a capital letter and marked with the *indice de lemme* `N` or `O`), and consequently annotated as such in our corpus, e. g. *ambiuareti*, even if they could also fit in the `NOUN` category as well. Examples of `NOUN`s: *mater*, *senatus*, *bellum*, *dignitatem*, *auis*.

- `NUM`: numerals. Class of functional relational nominal words, subclass of determiners (`DET`) which express specific numbers. This class substantially identifies with cardinal numbers (in UD annotated with `NumType=Card`, but this feature is not shown in our data; see §3.1.3), whether they are represented in a literal form, e. g. *milia* '1000', or a symbolic one, e. g. *XVIIII* '19'. In the latter case, they are lemmatized under their form, not with an equivalent literal numeral as in the LASLA corpus, and considered (graphically) indeclinable. In Latin, inflection is present only for the first cardinal numbers *unus*/*una*/*unum* '1' (always a numeral) and *duo*/*duae*/*duo* '2' (like first-class adjectives, i. e. `InflClass=IndEurA,IndEurO`, see §3.1.3), and *tres*/*tria* '3' (like a regular second-class adjective, i. e. `InflClass=IndEurI`), compounds of *centum* '100' like *ducenti*/*ducentae*/*ducenta* '200' (again like first-class adjectives), and *mille* '1000' (like a iotic third-declension noun, i. e. `InflClass=IndEurI`) with regard to `Case` (and `Number`). All others are indeclinable and only possess an inherent `Number`. Other numeral elements like ordinals, multiplicatives or distributives are tagged as adjectives (`ADJ`) or adverbs (`ADV`), according to their function.

- PART: particles. Particles are a residual class of functional words that cannot really be assigned to other parts of speech. Their function is often to convey a single grammatical category associated to a word, phrase, or entire clause: an important subclass of particles are discoursive elements like *quidem* or *nam*, which prop up the conversation flow. Other relevant particles are negation elements, imparting negative polarity: *non*, *haud*, *ne*/*ni*. No particle inflects in Latin.

- PRON: pronouns. A class of nonrelational nominal words that are the functional counterparts of nouns (NOUN/PROPN). They refer deictically to some entity without explicitly mentioning it and can act as arguments of a clause the same way nouns do, but cannot modify another word like adjectives (ADJ) or determiners (DET). Our annotation differs from that of traditional grammars in that what are often considered "pronouns" there are tagged as DET if they can be used attributively (similarly to adjectives); we note that in Latin every nominal word can also be head of a phrase, without an explicit noun, so that the ability to occur without a noun does not by itself determine the annotation as PRON or DET. This means that while personal pronouns (e. g. *ego*, *sibi*) are PRONs, possessive elements like *meam*, *suorum* will always be DETs. Pronouns inflect for Case and Number; their InflClass might depend on the Gender of the referred entity, similarly as for adjectives, but often is of pronominal type (LatPron) or even totally idiosyncratic (LatAnom, e. g. *ego*, *me*, *mihi*, *nos*, *nobis*). Other examples: *qui*/*quae*/*quod* (relative), *quis*/*quid*, *is*/*ea*/*id*.

- PROPN: proper nouns. This is actually a semantically determined subclass of NOUN that singles out terms that identify specific, unique entities such as individuals, places, or deities. Examples are *lucilio* (name of a Roman *gens*, formally an adjective), *mosella* (name of a river), *hercules* (name of a divinity). Morphologically, they act the same as NOUNs. We note that the definition of what is a "proper noun" is not unproblematic, so there might be some slight obscillations in the annotation of NOUN/PROPN in our data, e. g. for names of populations like *bellouaci* (involving also ADJs when an adjective is considered to be substantivized, e. g. *romanus*/*romana*/*romanum* vs. its isolated singular masculine *romanus*). On the contrary, adjectives (or nouns) derived from proper names are always just ADJs (or NOUNs): *urbinas* from *urbinum*, *platonicus* from *plato*.

- SCONJ: subordinating conjunctions. Class of functional words that connect a clause to its main clause while backgrounding it. Adpositions can also have this function in conjunction with nominalized verb forms, but subordinating conjunctions are specialized for clauses and cannot introduce noun phrases. Subordinating conjunctions are indeclinable in Latin, but many originate from crystallized forms of determiners, adverbs or verbs. Examples: *postquam*, *dum*, *ut*, *quomodo*.

- VERB: verbs. A class of lexical relational words conveying processes, actions, states or events. In Latin, verbs are observed to take two fundamental types of forms (VerbForm): so-called finite forms (Fin), which follow specific inflectional paradigms (conjugations) and agree with their subjects by means of Person and Number, and

nonfinite forms, some of which (`Part`, `Gdv`) agree with their subjects by means of nominal inflections (declensions) for `Case` and `Number` (and `Gender`, not shown in our data), and others (`Inf`, `Sup`; also `Ger`, not in our data, see §3.1.3) which do not agree, as they do not inflect. The agreeing forms are also gradable (`Degree`). In general, all these forms act as heads of an independent or embedded predicate, of which they express nuances by means of the grammatical categories of `Aspect` and `Voice` (all forms) and `Mood` and `Tense` (only so-called finite forms). A predicate can have different kinds of core (subject, object) or oblique arguments, which inflect for the necessary `Case`s. Examples of verbs are: *rapiebat* (finite), *scire* (nonfinite, infinitive), *potest* (finite). All forms that are observed to potentially have such an argumental structure, be they finite or nonfinite, and that are considered to be regularly part of a verbal paradigm, are annotated as `VERB`s, even when they represent more or less nominalized forms. Conversely, verb forms that do not act as predicates anymore (e. g. they lose their relationality) are annotated according to their new function under which they are lexicalized, and this is represented in the LASLA corpus or UD by a corresponding part of speech, lemmatization and absence of specifically verbal features (`Aspect`, `Voice`, `Mood`, `Tense`, `Person`). For example, we have the `ADJ` *altus* (from a participle of *alo*), the `CCONJ` *scilicet* (from the compounded finite expression *scire licet*), the `NOUN` *natura* (from a participle of *nascor*). We observe that, as far as annotation goes and especially with regard to participles, there is sometimes variation about when exactly a verb form is still considered to act as a `VERB`, or has already become something else.

- `X`: residual class. This tag is used for words that for some reason cannot be assigned a part of speech. In the EvaLatin data set, the tag `X` appears especially for foreign words which are extraneous to an annotation from the perspective of the Latin system, whose lemma is by default identical to their form; and also for *lacunae* in the text and broken forms, and also forms that are impervious to any sound analysis. These latter all also lack a lemma.

Please note that the tags `PUNCT` (punctuation) and `SYM` (symbol), included in the UD PoS tags, are not used in the EvaLatin 2022 data set: both categories are absent, and punctuation has been completely removed from all sources in conformity to what happens in the LASLA corpus.

Also note that the very same word form (~ token), sometimes even with the same lemma and/or etymology, can appear in the data under different morphosyntactic annotations, i. e. PoS and morphological features (and possibly lemmatization, too). This is especially true for indeclinable functional word classes, for which syntactic behaviour and part of speech go in parallel, e. g. *cum* either as a `SCONJ` (***cum*** *uenisset...*) or an `ADP` (***cum*** *amico*), or *post* either as an `ADV` (*nec multo* ***post*** *necessitas abiit*) or an `ADP` (***post*** *victoriam properauerant*). But we have also the form *uentum*, which is either a `NOUN` (*et* ***uentum*** *et aestum uno tempore nactus secundum*) or a `VERB` (*diem quartum quam est in Britanniam* ***uentum***): however, these two forms are also etymologically unrelated and have respective lemmas *uentus* 'wind' and *uenio* 'I come'.

3.1.3 Morphological Features

Features have the form `Name=Value`, with first letters always uppercase; they are sorted in alphabetical order and separated by a pipe, i. e. a vertical bar '|'. Multiple values of the same feature are sorted alphabetically, and they are to be interpreted as a disjunction: only one of those values is the appropriate one, but it cannot be chosen which one because of formal, unresolvable ambiguity (see e. g. the entry for `InflClass` below). We annotate and evaluate only the following features, which have in common their strict morphological nature (in contrast to lexical features like `PronType`[5])[6]:

- `Abbr`: binary feature, denoting with `Yes` the graphical abbreviation of a word. Otherwise, the word is regularly annotated with its part of speech and full lemma, but, if part of an inflecting word class, it receives `InflClass=Ind` (i. e. indeclinable) as a default, and thus it will also lack `Case` and `Number`. Note that punctuation has been completely stripped from the texts, and so (differently from LASLA) abbreviations are never terminated by a period.

- `Aspect`[7]: this feature expresses the way the action, event or state of the predication is considered and presented in its temporal development, and it is distinct both from temporal reference (`Tense`) and Aktionsart (which is a lexical feature and is not annotated). There are values for imperfective (or *infectum*) `Imp`, perfective (or *perfectum*) `Perf` and prospective `Prosp` (used for "future" nonfinite verb forms). The label for inchoative `Inch` is not used in the data.

- `Case`: it can have the classical values for nominative `Nom`, genitive `Gen`, dative `Dat`, accusative `Acc`, ablative `Abl`, vocative `Voc`, and also locative `Loc` appears in the data, since it is distinct from genitive (as it is sometimes traditionally labeled). Case can be seen as the morphological realization of the syntactic/semantic role of a word in the clause; the labelings in UD follow the traditional denominations.

  NB: indeclinable (`InflClass([nominal])=Ind`) forms are not assigned a `Case`, since there is no morphological correspondence with their syntactic/semantic role.

- `Degree`: it can have either values `Cmp` for the comparative or `Abs` for the superlative. which in Latin is basically of absolute nature (i. e. it expresses a very high degree with no specific comparison).

  NB: the label `Pos` for the traditional positive degree does not appear in the data, due to two main reasons: a practical one, because its annotation in LASLA does not easily allow us a meaningful implementation when converting it into the UD formalism; and a typological one, because the notion itself of positive

[5] Cf. the distinction in `https://universaldependencies.org/u/feat/index.html`.

[6] In many cases, Latin makes use of fusional affixes, so that more morphological features are expressed by the same form at once; this is especially the case for verbs.

[7] `https://universaldependencies.org/la/feat/Aspect.html`

degree as a "neutral" degree, opposed to comparative and superlative and thus defined "by negation", is problematic (or at least superfluous in this context).

- `InflClass`[8]: this feature expresses the class of the inflectional paradigm, both nominal and verbal, of the word. Label denominations are chosen so as to possibly establish links with other Indo-European (hence the prefix `IndEur`) languages where it makes sense, or to point to specific developments of Latin (hence the prefix `Lat`).
  - The possible values and their correspondences with traditional names are: first declension `IndEurA`, second declension `IndEur0`, athematic (gen. pl. *-um*) third declension `IndEurX`, "iotic" (gen. pl. *-ium*) third declension `IndEurI`, fourth declension `IndEurU`, fifth declension `IndEurE`, pronominal declension `LatPron`; first conjugation `LatA`, second conjugation `LatE`, third (athematic) conjugation `LatX`, fourth conjugation `LatI`, fifth or mixed conjugation `LatI2`; anomalous or irregular, i. e. not fitting into any other paradigm, inflections `LatAnom`; indeclinable `Ind`.

  NB[1]: traditional adjective classes are traced back to the nominal paradigms followed by the word forms: so either `IndEurA` (feminine) or `IndEur0` (masculine or neuter) for first-class adjectival forms, or `IndEurI` (usually) or `IndEurX` (a closed class; e. g. *uetus*) for second-class adjectival forms. There exist also a handful of indeclinable (`Ind`) adjectival forms, e. g. *nequam* or *tot*.

  NB[2]: All word forms have been assigned a univocal inflectional class, but there is only one case where disambiguation is systematically not possible when automatically converting from the LASLA to the UD formalism: first-class adjectival forms (including `ADJ`s, `DET`s and `NUM`s) in the dative/ablative plural (all terminating in *-is*) cannot be distinguished whether representing a first-declension (i. e. feminine) or second-declension (i. e. masculine or neuter) ending, so, only for these occurrences, the ambiguous notation `InflClass=IndEurA,IndEur0` appears in the data.

- `InflClass[nominal]`[9]: the same as `InflClass`, but used for nonfinite verbal forms to distinguish their layer of nominal declension from their simultaneous verbal conjugation, which determines e. g. the thematic vowel. Since the basic feature, `InflClass`, refers to the part of speech of the word (in this case `VERB`), and thus to conjugation, the nominal declension linked to `VerbForm` has to be represented by this layered feature.
  - Consequently, possible values are `IndEurA`, `IndEur0`, `IndEurX`, `IndEurI`, `IndEurU`, `Ind`.

- `Mood`: it expresses the attitude or manner with which the speaker presents the action, event or state of the predication, especially in relation to its factuality

[8] `https://universaldependencies.org/la/feat/InflClass.html`
[9] `https://universaldependencies.org/la/feat/InflClass-nominal.html`

(e. g. *realis* ↔ *irrealis*, ut also assertion ↔ command). In Latin, it appears only for so-called finite forms, and can assume the values of indicative `Ind` (*realis*), subjunctive `Sub` (*irrealis*) or imperative `Imp` (command, wish).

NB: "mood" in this (typological) sense has not to be confused with the traditional use of many Latin grammars to also call "moods" nonfinite verb forms like the infinitive or the gerundive. This is language-specifically possible only because such forms do not express a `Mood` (so no ambiguity can arise by the complementary distribution of this grammatical category among verb forms), but we are actually in presence of different phenomena here. In UD, nonfinite verb forms are handled under `VerbForm`.

- `Number`: this feature has two values, either singular `Sing` or plural `Plur`. It is used both for nominal and for verbal (in conjunction to `Person`) inflections. As for all other features, annotation follows morphology, not semantics, so that for example a *pluralia tantum* like *cunae* 'cradle' will always be assigned the value `Plur`, not `Sing`.

  NB: indeclinable (`InflClass([nominal])=Ind`) forms are not assigned a `Number`, since they do not express this category morphologically.

- `Person`: this feature identifies one between the first `1` (the speaker), the second `2` (the interlocutor) or the third `3` person (external actor). In Latin, this category is expressed only by verbs and personal pronouns, and then usually coupled with a `Number`. While for the former this feature is morphological, for the latter it tends towards lexicality, but since personal pronouns are functional words, the complete identification of such forms with a person/number is a motivation for us to keep them annotated with `Person`.

  NB: viceversa, personal determiners like *meus* 'my' are not annotated for `Person`, since the person (and number) of the possessor belongs to another layer (`Person[psor]` and `Number[psor]`) than the `Case`, `InflClass` and `Number` of their nominal declension.

- `Tense`: it specifies the absolute occurrence time of the action, event or state of the predication in relation to when the utterance takes (or is assumed to take) place. It is different from the relative time or expressed time development, which is represented by the `Aspect`. In Latin there are the three tenses past `Past`, present `Pres` and future `Fut`, together with the pluperfect `Pqp` (*plus quam perfectum*), which is retained in the data in conformity to traditional annotations (see below). `Tense` is applied to all moods.

  NB[1]: nonfinite forms do not express a `Tense`, but instead an `Aspect`. The correspondences with traditional terminology are present ↔ imperfective, past/perfect ↔ perfective and future ↔ prospective. Hence, a "present" participle will be marked for `Aspect=Imp` (besides `InflClass[nominal]=IndEurI|`

`VerbForm=Part|Voice=Act` and the appropriate values for verbal inflection, case, number and possibly degree); a gerundive for `Aspect=Prosp`, and so on.

NB[2]: some traditional tenses are actually represented by a combination of features, so for example the "imperfect" has `Aspect=Imp|Tense=Past` (i. e. it is an imperfective past), and so on. A particular case is the perfect: its straightforward, rather uncontroversial analysis would be that of a perfective present, hence `Aspect=Perf|Tense=Pres`; by the same token, the pluperfect would be a perfective past (as opposed to the imperfect), and so `Aspect=Perf|Tense=Past`. However, the perfect is often recognized as having a past reference: in compliance to this interpretation, we keep annotating it as `Aspect=Perf|Tense=Past`. This choice means that the pluperfect needs the label `Pqp` to be differentiated, giving `Aspect=Perf|Tense=Pqp` as a result. On the contrary, there is no such "drift" for future tenses, so that e. g. indicative future and future perfect are regularly marked respectively with `Aspect=Imp|Tense=Fut` and `Aspect=Perf|Tense=Fut`.

- `Variant`[10]: the only value for this feature is `Greek`, meaning that the inflected form of the word is realized by means of a Greek or Graecizing ending rather than a Latin one. It can thus be seen as a subcategorization of `InflClass`. An example is the accusative singular form *iatraliptеn* from *iatralipta*, instead of an expected *iatraliptam* for a first declension. Only forms diverging from Latin paradigms are taken into account: if an ending expressing the same morphological values is identical both in Latin and in Greek, even if the word is of Greek origin, there is no reason to label it with `Variant=Greek`.

- `VerbForm`[11]: this feature is of morphosyntactic nature and accounts for the role of a verbal form in the clause. In Latin, there is a fundamental distinction between traditionally called "finite" forms `Fin`, which can be the head of an independent clause without the support of a copula and always express `Mood`, `Person` and `Tense`, and nonfinite forms, whose syntactic behaviour is similar to that of adjectives, nouns or possibly adverbs. In our data, these latter forms are distinguished between infinitives `Inf`, participles `Part`, gerundivial forms `Gdv` and supines `Sup`. Though this classification can be challenged from a universal perspective (cf. [2]), in particular the distinction betwen participles and gerundives, it is internally coherent and so is kept in our data; in a sense, in Latin the value of `VerbForm` might be considered redundant, as it is already identifiable by the set of values of other morphological features (e. g. `Aspect`, `Voice`, presence/absence of `Mood`, . . .).

  NB[1]: since the gerundives and gerunds of standard Latin grammars are uncontroversially the same forms from a morphological perspective, they are united under the same label `Gdv` ("gerundival forms") in our data. In fact, their distinction is possible only at a syntactic level (and, even then, it is not always

[10] `https://universaldependencies.org/la/feat/Variant.html`

[11] See documentation at `https://universaldependencies.org/la/feat/VerbForm.html` and [2]

clear-cut), but we are lacking an annotation of syntactic dependency relations. Further, there is also a technical reason in one of our sources lacking this distinction.

NB[2]: in our data, the so-called "passive" supine (i. e. in the ablative case) is subsumed under regular (non-verbal) nouns `NOUN`s. This means that the label `Sup` is found only on "active", i. e. accusative, supines, and it might be seen as a language-specific equivalent to `Conv` (converb, i. e. adverbial verb form).

NB[3]: adverbial forms derived from nonfinite verb forms, e. g. *oboedienter*, from *oboediens*, imperfective active participle of *oboedio*, are labeled directly as adverbs (`ADV`), with lemma identical to their (positive, i. e. `Degree`-less) form, and are consequently not marked for any verbal feature (`Aspect`, `Voice`, ...).

- `Voice`: it is articulated in the two values for active `Act` and passive `Pass` diatheses, mirroring the two parallel, distinct inflectional paradigms seen for imperfective forms of transitive verbs, e. g. *amo*/*amor*, *amas*/*amaris*, *amat*/*amatur*,..., where they indeed correspond to a respectively transitive/active and intransitive/passive syntax and meaning. Since this is a morphological feature, though, any medial, impersonal or similar uses that Latin expresses with the latter conjugation are still labeled as `Pass`, and similarly, deponent verbs (e. g. *sequor* 'I follow') are always assigned a `Pass` value notwithstanding their transitive syntax and active semantic interpretation.

  NB: the forms of the auxiliary verb *sum* 'I am', which has part of speech `AUX`, are not marked for `Voice`. On the one hand there exists no passive conjugation of *sum*, nor is one thinkable or construable, and on the other hand the notion itself of voice is not really compatible with its function as a copula.

Finally, we note that the EvaLatin data set does not annotate the `Gender` (with its possible values feminine `Fem`, masculine `Masc` and neuter `Neut`) feature. This is due to two main reasons. First and more practical, the LASLA formalism does not annotate grammatical gender for (proper) nouns (`NOUN`s and `PROPN`s), and, when it does for other, variable, nominal parts of speech, this happens context-independently: this means that, in relation to number and case, all *a priori* possible values for the given form are assigned, for example `1` "genre commun" (corresponding to `Gender=Fem,Masc,Neut`, so a null annotation) for any genitive plural second-class adjective in *-(i)um*, or `4` "masc. & neutre" (corresponding to `Gender=Masc,Neut`) to any genitive singular first-class adjective in *-i* (opposed to just feminine *-ae*). This *impasse* can be tackled and greatly reduced by recurring to a lexical knowledge base like LiLa [5], but still many ambiguities are left, some of which are unsolvable without manual intervention (like the "common gender" `1`). The magnitude of such a manual inspection is daunting, so we prefer to ignore the `Gender` feature altogether: in fact, and this is the second, more linguistic-theoretical reason, `Gender` is for nouns (mostly) a purely lexical, rather than a morphological, feature, orthogonal to inflectional class, case and number[12] (this means, for example, that a noun of the second

[12]We note that this might be the reason in the first place for the apparently poor choice of

declension, while most probably being masculine, could *a priori* be also feminine or even neuter, given e. g. a Greek origin, and there are indeed attested occurrences for all such cases; at the same time, their number and case will be expressed identically). Inflectional class, on the other hand, is a factual morphological feature and, for first-class adjectives, it is even nearly equivalent to the `Gender` category. For all these reasons, we only retain `InflClass([nominal])` and skip `Gender` in our data set, confining the cases of annotational ambiguity only to dative/ablative plural forms of first-class adjectival forms (`ADJ`s, `DET`s, `NUM`s), cf. **NB[2]** under `InflClass` in the above list. Similarly, in the EvaLatin data set also other more lexical (or syntactic or orthographic) than morphological features, or features with a problematic annotation in our sources that nonetheless appear in other Latin UD treebanks, are omitted: `AdpType`, `AdvType`, `Clitic`, `Compound`, `ConjType`, `Foreign`, `Form`, `NameType`, `Number[psor]`, `NumForm`, `NumType`, `PartType`, `Person[psor]`, `Polarity`, `Poss`, `PronType`, `Proper`, `Reflex`.

3.2 Sub-tasks

Each of the aforementioned tasks has three sub-tasks:

1. **Classical**: test data will be of the same genre (i. e. prose) and period (i. e. Classical) of the training data but of an author not present in the training data;

2. **Cross-genre**: test data will be of two different genres compared to the ones included in the training data;

3. **Cross-time**: test data will be of a different period compared to the ones included in the training data.

Through these sub-tasks, we aim to enhance the study of the portability of NLP tools for Latin across different genres and temporal periods analysing the impact of genre-specific and diachronic features.

not annotating grammatical gender in the LASLA corpus.

Chapter 4

Evaluation

Each participating team will initially have access only to the training data. Later, the unlabeled test data will also be released. After the assessment, the labels for the test data will also be released.

The scorer employed for EvaLatin 2022 is a modified version of the one developed for the *CoNLL18 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies*[1]. An example of the scorer's output is given in Figure 3. The evaluation starts by aligning the system-produced words to the gold standard ones; given that we provide already tokenized and sentence-split test data, the alignment for tokens, sentences and words should be perfect (i. e. 100.00). Then, `UPOS` tags, lemmas and features are evaluated: precision, recall, F1 and accuracy are calculated. The final ranking will be based on accuracy.

NB: On March 17th 2022 we released a new version of the scorer in which tokens affected by a bug in the training set (see Section 2.2) are skipped and not taken into consideration when calculating the accuracy. See Appendix B for the list of types affected by the bug.

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	90.48	90.48	90.48	90.48
UFeats	85.71	85.71	85.71	85.71
Lemmas	95.24	95.24	95.24	95.24

Figure 3: Example of the scorer's output.

As a baseline, we will provide the scores obtained on the test data using UDPipe [8] with a model trained on the Perseus Universal Dependencies Latin Treebank[2] [1], as it is

[1] `https://universaldependencies.org/conll18/evaluation.html`
[2] `https://github.com/UniversalDependencies/UD_Latin-Perseus/`

available from the tool's web interface[3]. Please note that the accuracy rate on `FEATS` is very low because there are several differences between the morphological features in the data set used to train the Perseus model of UDPipe and those in our data. For example, we adopt the feature `InflClass`, not attested in the training data of the Perseus model.

AUTHOR	**TEXT**	**LEMMA**	**UPOS**	**FEATS**
Livius	Ab Urbe Condita (book VIII)	80.36	78.23	24.98

Table 5: Baseline for the *Classical* sub-task in terms of accuracy

AUTHOR	**TEXT**	**LEMMA**	**UPOS**	**FEATS**
Pliny the Elder	Naturalis Historia (book XXXVII)	77.96	75.35	24.16
Ovidius	Metamorphoseon (books IX-X)	80.11	77.82	22.52
MACRO-AVERAGE		**79.03**	**76.58**	**23.34**

Table 6: Baseline for the *Cross-genre* sub-task in terms of accuracy

AUTHOR	**TEXT**	**LEMMA**	**UPOS**	**FEATS**
Sabellicus	De Latinae Linguae Reparatione	81.92	74.26	27.84

Table 7: Baseline for the *Cross-time* sub-task in terms of accuracy

[3] `http://lindat.mff.cuni.cz/services/udpipe/`

Chapter 5

How to Participate

Participants will be required to submit their runs and to provide a technical report describing their systems.

5.1 Submitting Runs

Each participant can submit runs for each sub-task within each task. A run should be produced according to the "closed modality": the only annotated data to be used to train and tune the system are those distributed by the organizers. Other non-annotated resources, e. g. word embeddings, are instead allowed. The second run will be produced according to the "open modality": annotated external data, such as the Latin data sets of the Universal Dependencies initiative, can be also employed. All external resources are expected to be described in the systems' reports. The closed run is compulsory, while the open run is optional.

Once the system has produced the results for the task over the test set, participants have to follow these instructions to complete their submissions:

- name the runs with the following filename format: `task_sub-task_docName_teamName_systemID_modality.conllu`.
  For example: `pos_classical_BellumCivile_unicatt_1_closed.conllu` would be the first run of a team called *unicatt* using the closed modality for the PoS Tagging task and the Classical sub-task on the `Caesar_BellumCivile_LiberI_TEST.conllu` document.
  `lemma_cross-genre_DeVitaBeata_unicatt_2_open.conllu` would be the second run of a team called *unicatt* using the open modality for the lemmatization tagging task and the Cross-genre sub-task for the `Seneca_DeVitaBeata_TEST.conllu` document.

- send the file to the following email address: `rachele.sprugnoli[AT]unicatt.it`, using the subject "EvaLatin Submission: task - teamName", where the "task" is either *PoS* or *Lemma*.

5.2 Writing the Technical Report

Technical reports will be included in the proceedings of the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) as short papers and will be published along the LREC 2022 proceedings. Reports must be submitted through the START platform (URL available soon). All the reports must meet the following requirements:

- they must be written in English;
- they must be formatted according to the LREC 2022 conference style[1];
- the maximum length is 4 pages (excluding references);
- they should contain (at least) the following sections: description of the system, results, discussion, references.

Reports will receive a light review: we will check for the correctness of the format, the exactness of results and ranking, and overall exposition. If needed, we will contact the authors asking for corrections.

[1] https://lrec2022.lrec-conf.org/en/submission2022/authors-kit/

Appendix A

Selection of Resources for Latin

- Lemma embeddings: https://embeddings.lila-erc.eu/
- Latin texts and embeddings: http://www.cs.cmu.edu/~dbamman/latin.html
- Latin BERT: https://github.com/dbamman/latin-bert
- Word embeddings: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
- Other embeddings: http://embeddings.texttechnologylab.org/
- CLTK: http://cltk.org/
- UD Latin PROIEL: https://github.com/UniversalDependencies/UD_Latin-PROIEL
- UD Latin ITTB: https://github.com/UniversalDependencies/UD_Latin-ITTB
- UD Latin Perseus: https://github.com/UniversalDependencies/UD_Latin-Perseus
- UD Latin LLCT: https://github.com/UniversalDependencies/UD_Latin-LLCT
- UD UDante: https://github.com/UniversalDependencies/UD_Latin-UDante
- Latin texts: https://github.com/PerseusDL
- Collatinus: https://outils.biblissima.fr/en/collatinus/index.php
- LEMLAT v.3: https://github.com/CIRCSE/LEMLAT3
- Treetagger: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- Glossaria: https://glossaria.eu/outils/lemmatisation/#page-content
- Word Formation Latin (WFL) lexicon: http://wfl.marginalia.it/
- Lemmatized corpus with morphological analysis "Corpus Latin antiquité et antiquité tardive lemmatisé": https://doi.org/10.5281/zenodo.4337145

- Latin Lasla Model (Thibault Clérice) for lemmatization and morphological analysis:https://doi.org/10.5281/zenodo.3773327

Appendix B

Types Affected by a Bug

The bug consisted in the accidental removal of all the occurrences of initial and/or final letters "a" and "b" from the following types (all numeral terms of various kinds): *aliquanto, aliquantum, aliquot, ambas, ambo, ambobus, amborum, ambos, bina, binae, binas, binis, binorum, binos, bis, complura, decima, ducenta, duetuicensima, duodecima, duodena, duodena, duodetriginta, duoetuicensima, duplicia, milia, multa, plura, plurima, nonaginta, nona, octaua, octingenta, octoginta, octona, pauca, pauciora, paucissima, prima, priora, quadraginta, quadringenta, quanta, quartadecima, quarta, quaterna, quingenta, quinquaginta, quinta, quina, secunda, sena, septima, septingenta, septuaginta, sescenta, sexcenta, sexaginta, sexta, singula, terna, terna, tertia, trecenta, tria, tricena, tricesima, triginta, trina, undecima, unaetuicensima, una.*

The tokens affected by this bug were 1,109 (out of 316,573).

Bibliography

[1] David Bamman and Gregory Crane. The ancient Greek and Latin dependency treebanks. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98, Berlin/Heidelberg, Germany, 2011. Springer. Preprint retrievable at `http://www.cs.cmu.edu/~dbamman/pubs/pdf/latech2011.pdf`.

[2] Flavio Massimiliano Cecchini. *Formae reformandae*: for a reorganisation of verb form annotation in Universal Dependencies illustrated by the specific case of Latin. In Petya Osenova, Kiril Simov, Myriam de Lhoneux, and Reut Tsarfaty, editors, *Proceedings of the Fifth Workshop on Universal Dependencies (*UDW*, SyntaxFest 2021)*, pages 1–15, Sofia, Bulgaria, dec 2021. The Association for Computational Linguistics (ACL). Retrievable at `https://aclanthology.org/2021.udw-1.1/`.

[3] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, July 2021. Retrievable at `https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies`.

[4] Joseph Denooz. Opera Latina : une base de données sur internet. *Euphrosyne*, 32:79–88, 2004. Retrievable at `https://www.brepolsonline.net/doi/10.1484/J.EUPHR.5.125535`.

[5] Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1):177–212, 2020. Retrievable at `https://www.studiesaggilinguistici.it/ssl/article/view/277`.

[6] Caroline Philippart de Foy. LASLA – *Nouveau manuel de lemmatisation du latin.* LASLA, Liège, Belgium, 2014. Retrievable at `https://orbi.uliege.be/handle/2268/171931`.

[7] Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 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, May 2020. European Language Resources Association (ELRA). Retrievable at https://www.aclweb.org/anthology/2020.lt4hala-1.16.

[8] Milan Straka, Jan Hajič, and Jana Straková. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). Retrievable at https://aclanthology.org/L16-1680.