Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 183–188
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

Overview of the EvaLatin 2022 Evaluation Campaign

Rachele Sprugnoli[1], Marco Passarotti[2], Flavio M. Cecchini[2], Margherita Fantoli[3], Giovanni Moretti[2]
[1]Università di Parma, [2]CIRCSE Research Centre, Università Cattolica del Sacro Cuore, [3]KU Leuven
rachele.sprugnoli@unipr.it, marco.passarotti@unicatt.it, flavio.cecchini@unicatt.it
margherita.fantoli@kuleuven.be giovanni.moretti@unicatt.it

Abstract

This paper describes the organization and the results of the second edition of EvaLatin, the campaign for the evaluation of Natural Language Processing tools for Latin. The three shared tasks proposed in EvaLatin 2022, i. e. Lemmatization, Part-of-Speech Tagging and Features Identification, are aimed to foster research in the field of language technologies for Classical languages. The shared dataset consists of texts mainly taken from the LASLA corpus. More specifically, the training set includes only prose texts of the Classical period, whereas the test set is organized in three sub-tasks: a *Classical* sub-task on a prose text of an author not included in the training data, a *Cross-genre* sub-task on poetic and scientific texts, and a *Cross-time* sub-task on a text of the 15th century. The results obtained by the participants for each task and sub-task are presented and discussed.

Keywords: Latin, evaluation, NLP

1. Introduction

EvaLatin 2022 is the second edition of the campaign devoted to the evaluation of Natural Language Processing (NLP) tools for the Latin language. Like in 2020, EvaLatin is proposed as part of the *Workshop on Language Technologies for Historical and Ancient Languages* (LT4HALA 2022), co-located with LREC 2022.[1] Similar to what happens in other international evaluation campaigns, participants have been provided with training and test data that are made freely available for research purposes to encourage further improvement of language technologies for Latin. Participants also had the chance to evaluate their systems using a shared script. Data, scorer and detailed guidelines are all available in a dedicated GitHub repository.[2]

EvaLatin is an initiative organized by the CIRCSE research centre[3] at the Università Cattolica del Sacro Cuore in Milan, Italy, with the support of the *LiLa: Linking Latin* ERC project.[4] 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.

2. Tasks and Sub-tasks

EvaLatin 2022 has three tasks:

1. **Lemmatization**, i. e. the process of transforming each word form into a corresponding conventional "base form", according to its part of speech (i. e. morphosyntactic properties) and etymology, which usually coincides with an entry found in the dictionary (i. e. lemma);
2. **Part-of-Speech tagging**, for which systems are required to assign each token a lexical category, i. e. a Part-of-Speech (PoS) tag, according to the Universal Dependencies (UD) PoS tagset (de Marneffe et al., 2021, §2.2.2), originally inspired by that of (Petrov et al., 2011).[5]
3. **Features Identification**, for which systems have both to correctly identify the UD morphological features (de Marneffe et al., 2021, §2.2.3) pertaining to the token's word form among the specific subset used in the EvaLatin 2022 dataset (see §3.), and to select correct values for them.[6]

Each task has three sub-tasks:

1. **Classical**: the test data belong to the same genres and time period of the training data;
2. **Cross-genre**: the test data belong to two different genres, namely mythological poem and scientific treatise, but roughly to the same time period compared to the ones included in the training data;
3. **Cross-time**: the test data belong to a different time period, namely the Renaissance era, 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 time periods by analyzing the impact of genre-specific and diachronic features.

Shared data and a scorer are provided to the participants, who can choose to take part in either a single task, or in all tasks and sub-tasks.

[1] https://lrec2022.lrec-conf.org/en/
[2] https://github.com/CIRCSE/LT4HALA/tree/master/2022/data_and_doc
[3] https://centridiricerca.unicatt.it/circse_index.html
[4] https://lila-erc.eu/
[5] https://universaldependencies.org/u/pos/index.html
[6] An overview is at https://universaldependencies.org/u/feat/index.html.

3. Data

The dataset of EvaLatin 2022 consists of texts mainly taken from the LASLA corpus (Denooz, 2004), a resource manually annotated since 1961 by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège,[7] Belgium. The texts are then converted into the annotation formalism of the UD project[8] (de Marneffe et al., 2021), 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. 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 until any 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 annotation guidelines are provided by the manual (Philippart de Foy, 2014).

Besides these texts from the LASLA corpus, the test data also include a text by Sabellicus, a Renaissance historian of the 15th century, annotated by members of the CIRCSE research center.

The conversion from the original fixed-length format of LASLA to the CoNLL-U format[9] and the UD formalism has also been developed at the CIRCSE research center and is based on Python[10] scripts complemented by the access to the LiLa lexical knowledge base (Passarotti et al., 2020). 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 and our specific choices concerning the morphological annotation. In particular, for this campaign just a subset of UD morpholexical features is retained, thus considering only the following features: `Abbr`, `Aspect`, `Case`, `Degree`, `InflClass`, `InflClass[nominal]`, `Mood`, `Number`, `Person`, `Tense`, `VerbForm`, `Voice`. The guiding principle here is to stick only to purely morphological features which can be tracked down in the word form, and at the same time to avoid features which are annotated inconsistently among texts. The former criterion leaves aside more lexically oriented features like `PronType` (the "pronominal type"), which hinge more on semantic arguments rather than on inflectional and syntactic behaviour; on a similar note, we also discard the `Gender` feature[11] (which is lexically determined) in favor of `InflClass` (which is readable from the word form).[12] The consistency criterion excludes a feature like `Polarity` which, though morphologic, is not systematically annotated in the texts at our disposal.

Overall, the accomplished 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. One fundamental example is the `AUX`/`VERB` split of UD, whereby the functional verb *sum* 'to be' is annotated as `AUX` (and not `VERB`, or `B` in LASLA) also in its occurrences as a copula, and not only as part of a periphrastic form. On the morphological level, another example is the separation of the notions represented in UD by the features `Mood` and `VerbForm`, which in LASLA, following the most common grammatical tradition, are conflated under the label of *mode* 'mood': so, in our dataset the *mode indicatif* corresponds to `Mood=Ind` (with `VerbForm=Fin`), while the *mode infinitif* to `VerbForm=Inf` (with no value for `Mood`). At the same time, *temps* 'tenses' are represented by different combinations of values for `Tense`, but also for `Aspect`, which is not directly indicated in LASLA.

For more details about morphological features, we point to the EvaLatin 2022 guidelines on the official website.[13]

3.1. 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 (Straka et al., 2016) based on the model trained on the Perseus UD Latin Treebank[14] (Bamman and Crane, 2011).

Texts are by five Classical authors for a total of more than 300,000 tokens: Caesar, Cicero, Seneca, Pliny the Younger and Tacitus. All texts are in prose but different genres are included: treatises by Caesar, Seneca and Tacitus, public speeches by Cicero, and letters by Pliny the Younger. Table 1 presents details about the training dataset of EvaLatin 2022, while Figure 1 shows an example of the format.

3.2. Test Data

Test data contain only the tokenized words but not the correct tags, which have to be added by the participant systems

[7] `http://web.philo.ulg.ac.be/lasla/textes-latins-traites/`

[8] `www.universaldependencies.org`

[9] `https://universaldependencies.org/format.html`

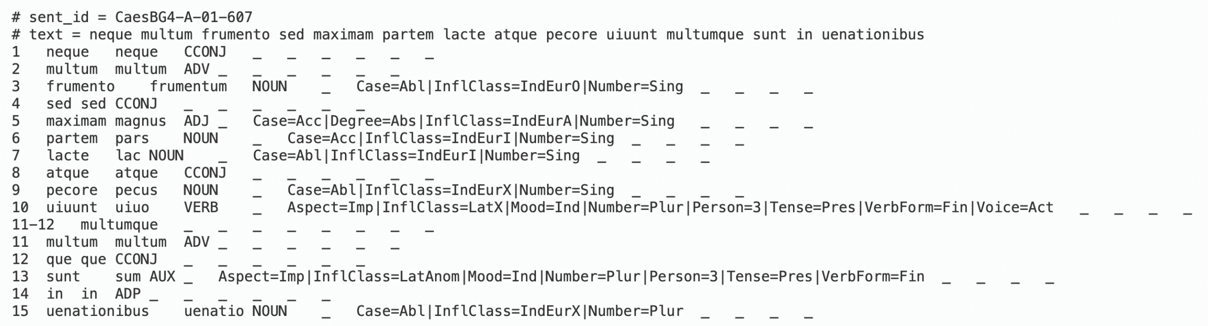
[10] `https://www.python.org/`

[11] We further note that the annotation of grammatical gender in LASLA drastically deviates in its logic from that of UD, making automated conversion problematic.

[12] `https://universaldependencies.org/la/feat/InflClass.html`

[13] `https://circse.github.io/LT4HALA/2022/EvaLatin.html`

[14] `https://github.com/UniversalDependencies/UD_Latin-Perseus/`



```
# 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 format of training data.

AUTHORS	TEXTS	# TOKENS
Caesar	De Bello Gallico	44,818
Caesar	De Bello Civili (I, II)	17,287
Cicero	Philippicae (I–XIV)	52,563
Cicero	In Catilinam	12,564
Pliny the Younger	Epistulae (I-VIII, 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, books in parentheses.

to be submitted for the evaluation. 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 has already been provided in tokenized format, one token per line, and with a blank line separating each sentence. The gold standard test data, i. e. the annotation used for the evaluation, was 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, while an example of the format of test data is given in Figure 2.

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

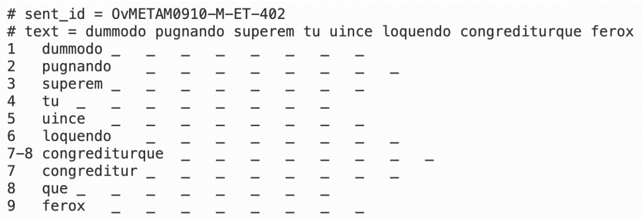
Table 2: Test data for *Classical* sub-task, books in parentheses.

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

Table 3: Test data for *Cross-genre* sub-task, books in parentheses.

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

Table 4: Test data for *Cross-time* sub-task, books in parentheses.



```
# sent_id = OvMETAM0910-M-ET-402
# text = dummodo pugnando superem tu uince loquendo congrediturque ferox
1   dummodo _   _   _   _   _   _   _   _
2   pugnando    _   _   _   _   _   _   _   _
3   superem _   _   _   _   _   _   _   _
4   tu  _   _   _   _   _   _   _   _
5   uince   _   _   _   _   _   _   _   _
6   loquendo    _   _   _   _   _   _   _   _
7-8 congrediturque  _   _   _   _   _   _   _   _
7   congreditur _   _   _   _   _   _   _   _
8   que _   _   _   _   _   _   _   _
9   ferox   _   _   _   _   _   _   _   _
```

Figure 2: Example of the format of test data.

4. Evaluation

The scorer employed for EvaLatin 2022 is a modified version of that developed for the *CoNLL 2018 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies* (Zeman et al., 2018).[15] The evaluation starts by aligning the outputs of the participating systems to the gold standard: given that our test data are already tokenized and split by sentences, the alignment at the token and sentence levels is always perfect (i. e. 100.00%). Then, POS tags, lemmas and features are evaluated and the final ranking is based on accuracy.

Each participant was permitted to submit runs for either one or all tasks and sub-tasks. It was mandatory to produce one run according to the so-called "closed modality", according to which the only annotated resources that could be used to train and tune the system are those distributed by the organizers. Also external non-annotated resources, like word embeddings, were allowed. The second run could be produced according to the "open modality", for which the use of additional annotated external data is allowed.

As for the baseline, we provided the participants with the scores obtained on our test data by UDPipe, using the model trained on the Perseus UD Latin Treebank[16] (Bamman and Crane, 2011), the same available in the tool's web

[15] https://universaldependencies.org/conll18/evaluation.html

[16] https://github.com/UniversalDependencies/UD_Latin-Perseus/

interface.[17]

5. Participants and Results

Two teams took part in EvaLatin 2022 submitting runs for all tasks and sub-tasks. Only one team (namely, Kraków) submitted one run following the open modality for each task and sub-task, whereas the other submitted runs in the closed modality only. Details on the participating teams and their systems are given below:

- Kraków, Jagiellonian University, Institute of Polish Language, Enelpol (Poland) (Wróbel and Nowak, 2022). This team employs transformer models for their runs: in particular, they use XLM-RoBERTa large (Conneau et al., 2020) for both PoS tagging and features identification, and a ByT5 model (Xue et al., 2022) for lemmatization. The runs developed following the open modality are trained adding annotated texts taken from the UD Latin treebanks and the whole LASLA corpus to the official dataset.
- KU-Leuven, KU Leuven, Brepols Publishers (Belgium) (Mercelis and Keersmaekers, 2022). The runs of this team are based on a pre-trained ELECTRA-model (Clark et al., 2020). The *Huggingface Transformers ElectraForTokenClassification* model is used for the PoS tagging task while handcrafted rules are added to handle lemmatization. For the Feature Identification task, a separate classifier is trained for each feature: the predicted labels are then joined at a later time.

Tables 5, 6 and 7 report the final rankings, showing the results in terms of accuracy, including our baseline. For each run, the team name and the modality are specified. Please note that for the *Cross-genre* sub-task the score corresponds to the macro-average accuracy.

6. Discussion

As shown in Tables 5, 6 and 7, all systems largely outperform the baseline: please note that the accuracy rate on Features Identification task is very low because there are several differences between the morphological features 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.

The open-run experiment by the Kraków team yields the best results in each of the tasks and sub-tasks: in particular, an improvement in accuracy is registered in the *Cross-genre* sub-task of the Lemmatization and PoS tasks (respectively +3.46% points and +1.44% points with respect to the run made following the closed modality). This shows that using additional annotated data (e. g. a broader portion of the LASLA corpus and UD treebanks) improves the results, despite the possible inconsistencies in the annotation styles.

Each sub-task contains only one text, with the exception of the *Cross-Genre* sub-task: the standard deviation among the texts of this sub-task (*Metamorphoseon libri* and *Naturalis Historia*) fluctuates between 1.04 and 2.02 (Lemmatization task), 0.22 and 1.75 (PoS task), 0.88 and 3.55 (Features). For the Lemmatization and PoS tagging tasks, the *Metamorphoseon libri* obtain better results than the *Naturalis Historia*, whereas for the Features Identification task, the three systems perform better on the *Naturalis Historia* than on the *Metamorphoseon libri*. This can be explained by the fact that the *Naturalis Historia* starkly differs from the training data because it deals with a very peculiar topic, i. e. precious stones, and thus features a highly specific vocabulary, which impacts the results of the Lemmatization and PoS tasks. For instance, the form *acaustoe* (also a Greek variant) of the ADJ *acaustos* 'incombustible' is wrongly lemmatized by all systems and assigned the PoS NOUN or PROPN. On the contrary, the *Metamorphoseon libri* differ from the training set because they are poetry and not prose, which entails a very different word order and syntax: such variations are likely to strongly impact the Features Identification task.

Taking a more in-depth look at the results on the test set as a whole, the easiest text to tackle with regard to Lemmatization for the KU Leuven model are the *Metamorphoseon libri* (*Cross-Genre*, accuracy of 87.22%), whereas the two Kraków models perform better on the *Ab Urbe Condita* (*Classical*, accuracy of 96.45% and 97.26%). The hardest text to tackle for all the systems appears to be the *De Latinae Linguae Reparatione* (with an accuracy ranging from 84.6% to 92.15%). This result might be surprising if one considers that this text has a significantly lower percentage of out-of-vocabulary lemmata and a lower lemma/token ratio than the *Naturalis Historia* (respectively 21.67% vs. 33.67%, and 20.2% vs. 25.6%). The results might be due to the fact that, whereas the *Naturalis Historia* is annotated following LASLA conventions, the *De Latinae Linguae Reparatione* is annotated in the frame of a different project; but probably the decisive factor is that, the *De Latinae Linguae Reparatione* being a significantly later text, orthographic variations (such as systematic *e* instead of *ae*, or spellings like *ocium* for *otium* 'leisure', or *phama* for *fama* 'reputation') have a stronger impact than expected on any task and/or system highly relying on word forms. In fact, Features Identification is more heavily impacted (losses in accuracy of up to -9.92% with respect to the *Classical* sub-task) than Lemmatization or PoS tagging (losses of up to -5.11%), which abstract more towards a lexical or syntactic level.

For the PoS tagging task, all systems perform best on the *Ab Urbe Condita*, with very similar results (accuracy ranges from 96.33% to 97.99%). The most difficult text is once again the *De Latinae Linguae Reparatione* (accuracy from 92.11% to 92.70% points). The most frequent errors occur in the categories ADJ, NOUN and PROPN. This is mostly due to the nominal use (i. e. as heads of noun phrases) of adjectival forms, like the adjective (ADJ) *Romanus* 'Roman', that can appear annotated as PROPN in the sense of 'the Roman citizen', or the presumed NOUN *malum* '(an) evil', which is nothing else than the neuter form of the ADJ *malus* 'bad'. Especially the first case is due to a general inconsistency in the annotation of the

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

Classical	
Kraków-open	97.26
Kraków-closed	96.45
KU-Leuven	85.44
Baseline	80.36

Cross-Genre	
Kraków-open	95.08 (1.34)
Kraków-closed	91.62 (2.02)
KU-Leuven	86.48 (1.04)
Baseline	79.03 (1.52)

Cross-time	
Kraków-open	92.15
Kraków-closed	91.68
KU-Leuven	84.60
Baseline	81.92

Table 5: Results of the Lemmatization task for the three sub-tasks in terms of accuracy. The number in brackets indicates standard deviation calculated among the two documents of the test set for the *Cross-Genre* sub-task.

Classical	
Kraków-open	97.99
Kraków-closed	97.61
KU-Leuven	96.33
Baseline	78.23

Cross-Genre	
Kraków-open	96.06 (1.01)
Kraków-closed	94.62 (0.22)
KU-Leuven	92.31 (3.32)
Baseline	76.58 (1.75)

Cross-time	
Kraków-closed	92.97
Kraków-open	92.70
KU-Leuven	92.11
Baseline	74.26

Table 6: Results of the PoS task for the three sub-tasks in terms of accuracy. The number in brackets indicates standard deviation calculated among the two documents of the test set for the *Cross-Genre* sub-task.

Classical	
Kraków-open	95.46
Kraków-closed	95.42
KU-Leuven	69.91
Baseline	24.98

Cross-Genre	
Kraków-open	89.43 (0.88)
Kraków-closed	89.32 (0.88)
KU-Leuven	60.55 (3.55)
Baseline	23.34 (1.16)

Cross-time	
Kraków-closed	86.50
Kraków-open	86.50
KU-Leuven	60.09
Baseline	27.84

Table 7: Results of the Feature Identification task for the three sub-tasks in terms of accuracy. The number in brackets indicates standard deviation calculated among the two documents of the test set for the *Cross-Genre* sub-task.

datasets not solved with the conversion and uniformization process described in §3.. Moreover, in Latin, adjectives and (proper) nouns almost completely overlap on their inflectional paradigms, so that a distinction based on formal criteria can incur in difficulties. Also, the difference between NOUN and PROPN is of a purely semantic rather than morphosyntactic or functional-vs.-lexically grounded nature; this makes PROPN anomalous in the UD PoS scheme, and explains why a system like KU Leuven can drop as low as 59.8% in accuracy for this PoS, and Kraków's reach some of its lowest scores.

Among verb forms, participial forms in particular are also liable to oscillate, in this case between an annotation as VERB on the one hand, and as ADJ or NOUN on the other hand, depending on the propension for a more morphological or syntactic analysis. Examples from the *Cross-Time* sub-task (i. e. Sabellicus's work) are i) the form *scriptis*, annotated as a NOUN with lemma *scriptum* 'written work' in the test data, but traced back to the VERB *scribo* 'to write' by one of the systems for being originally a participial form; ii) the form *occulto* (occurring in the expression *in occulto* 'secretely'), analyzed as a (participial) form of the VERB *occulo* 'to cover' in the test data, but labeled as a NOUN *occultum* 'secrecy' by one of the systems for being in a nominal context (here, an oblique argument introduced by a preposition). In fact, we see some inconsistencies in this sense between training and test data, and sometimes internally to the training data, too. In particular, the LASLA annotation seems to favor a more "functional" approach whereby e. g. a lexical adjective in a nominal context becomes tagged as a noun, while the tendency in the natively UD-annotated *De Latinae Linguae Reparatione* is to keep it annotated as an ADJ, delegating the representation of its more noun-like behaviour to the layer of syntactic dependency relations.

Similarly, the assignment of the label ADV proves to be particularly difficult with terms such as *uerum* 'certainly', *nunc* 'now' or *quippe* 'of course, by all means', which all lie in the syntactic grey area of sentence connectors and discourse particles, where the border between ADV and CCONJ (and also PART) can be blurred, and sometimes annotation in the data accordingly shows inconsistencies, too.

For the Features Identification task, the easiest text is again the *Ab Urbe Condita* and the hardest *De Latinae Linguae Reparatione*. The gap between the worst and best performing model is significantly larger than in the other tasks: the accuracy ranges from 69.91% to 95.46% on the *Ab Urbe Condita*, and from 60.09% to 86.53% on the *De Latinae Linguae Reparatione*. In general, Case is the most poorly identified feature (followed by InflClass and InflClass[nominal]), with an F1 score ranging from 53% (*Naturalis Historia*) to 95% (*Ab Urbe Condita*). The number of ambiguous forms (e. g. dative and ablative singular of the second declension, plural of second and first declensions; nominative, vocative and accusative of neuter names) and the role of the context for the disambiguation might explain this result.

7. Conclusion

This paper describes the second edition of EvaLatin, the evaluation campaign dedicated to NLP tools for the Latin language. Following the good results in terms of participation and performances obtained in 2020, this edition of EvaLatin has been organized around three tasks: in particular, the Features Identification task has been added to the

Lemmatization and PoS tasks, already proposed in 2020. Although there has been a drop in the number of participants (from 5 to 2), we are satisfied with the achieved results: new annotated data were released and new systems were tested using a common framework. Interestingly, the participating systems are both based on transformer models.

As for the future, we plan to keep organizing a new edition of EvaLatin every two years. Indeed, there are several variables still to address in the campaign, including (a) the authors and genres represented in the texts chosen for the training and test sets, and (b) the shared tasks to perform. With regard to the former, we plan to include Early Medieval documentary texts in the shared data, most likely by relying on the data provided by the Latin Text Archive.[18] For what concerns the latter, a challenge to address in the near future of EvaLatin is syntactic analysis, also in light of the results and the experience of the UD initiative.

8. Acknowledgements

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme via the *LiLa: Linking Latin* project - Grant Agreement No. 769994. The authors want to thank the LASLA, which provided newly lemmatized texts to be included in the test set. They also want to thank Federica Gamba (Charles University, Prague, Czech Republic) for the complete annotation of the *De Latinae Linguae Reparatione*, and Timo Korkiakangas (University of Helsinki, Finland) for having made it possible and permitted its distribution.

9. Bibliographical References

Bamman, D. and Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks. In Caroline Sporleder, et al., editors, *Language technology for cultural heritage*, Theory and Applications of Natural Language Processing, pages 79–98. Springer, Berlin - Heidelberg, Germany.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the International Conference on Learning Representations 2020 (ICLR)*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Denooz, J. (2004). Opera Latina: une base de données sur internet. *Euphrosyne*, 32:79–88.

Mercelis, W. and Keersmaekers, A. (2022). An electra model for latin token tagging tasks. In *Proceedings of LT4HALA 2022-2st Workshop on Language Technologies for Historical and Ancient Languages*.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. (2020). Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, LVIII(1):177–212.

Petrov, S., Das, D., and McDonald, R. (2011). A Universal Part-of-Speech Tagset. *ArXiv e-prints*. arXiv:1104.2086 at `https://arxiv.org/abs/1104.2086`.

Philippart de Foy, C., (2014). LASLA – *Nouveau manuel de lemmatisation du latin*. LASLA, Liège, Belgium.

Straka, M., Hajic, J., and Straková, J. (2016). 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.

Wróbel, K. and Nowak, K. (2022). Transformer-based part-of-speech tagging and lemmatization for latin. In *Proceedings of LT4HALA 2022-2st Workshop on Language Technologies for Historical and Ancient Languages*.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2022). Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

[18] `https://lta.bbaw.de`