Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 105–110
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Overview of the EvaLatin 2020 Evaluation Campaign

Rachele Sprugnoli, Marco Passarotti, Flavio M. Cecchini, Matteo Pellegrini
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano
{rachele.sprugnoli,marco.passarotti,flavio.cecchini}@unicatt.it
matteo.pellegrini@unibg.it

Abstract

This paper describes the first edition of EvaLatin, a campaign totally devoted to the evaluation of NLP tools for Latin. The two shared tasks proposed in EvaLatin 2020, i. e. Lemmatization and Part-of-Speech tagging, are aimed at fostering research in the field of language technologies for Classical languages. The shared dataset consists of texts taken from the Perseus Digital Library, processed with UDPipe models and then manually corrected by Latin experts. The training set includes only prose texts by Classical authors. The test set, alongside with prose texts by the same authors represented in the training set, also includes data relative to poetry and to the Medieval period. This also allows us to propose the *Cross-genre* and *Cross-time* subtasks for each task, in order to evaluate the portability of NLP tools for Latin across different genres and time periods. The results obtained by the participants for each task and subtask are presented and discussed.

Keywords: evaluation, lemmatization, PoS tagging

1. Introduction

EvaLatin 2020 is the first campaign being totally devoted to the evaluation of Natural Language Processing (NLP) tools for the Latin language.[1] The campaign is designed following a long tradition in NLP,[2] 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 is proposed as part of the *Workshop on Language Technologies for Historical and Ancient Languages* (LT4HALA), co-located with LREC 2020.[3] EvaLatin is an initiative endorsed by the Italian association of Computational Linguistics[4] (AILC), and is organized by the CIRCSE research centre[5] at the Università Cattolica del Sacro Cuore in Milan, Italy, with the support of the *LiLa: Linking Latin* ERC project.[6]
Data, scorer and detailed guidelines are all available in a dedicated GitHub repository.[7]

[1] https://circse.github.io/LT4HALA/

[2] See for example other campaigns such as MUC (Message Understanding Conference), a competition dedicated to tools and methods for information extraction, SemEval (Semantic Evaluation), which is focused on the evaluation of systems for semantic analysis, CoNLL (Conference on Natural Language Learning), which since 1999 has been including a different NLP shared task in every edition, and EVALITA, a periodic evaluation campaign of NLP tools for the Italian language.

[3] https://lrec2020.lrec-conf.org/en/

[4] http://www.ai-lc.it/

[5] https://centridiricerca.unicatt.it/circse_index.html

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

[7] https://github.com/CIRCSE/LT4HALA/tree/master/data_and_doc

2. Tasks and Subtasks

EvaLatin 2020 has two tasks:

1. **Lemmatization**, i. e. the process of transforming any word form into a corresponding, conventionally defined "base" form, i. e. its lemma, applied to each token;
2. **Part-of-Speech tagging**, in which systems are required to assign a lexical category, i. e. a Part-of-Speech (*PoS*) tag, to each token, according to the Universal Dependencies (UD) PoS tagset (Petrov et al., 2011).[8]

Each task has three subtasks:

1. **Classical**: the test data belong to the same genre and time period of the training data;
2. **Cross-genre**: the test data belong to a different genre, namely lyric poems, but 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 Medieval era, compared to the ones included in the training data.

Through these subtasks, 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 subtasks.

3. Dataset

The EvaLatin 2020 dataset consists of texts taken from the Perseus Digital Library (Smith et al., 2000).[9] These texts

[8] https://universaldependencies.org/u/pos/index.html

[9] http://www.perseus.tufts.edu/

are first processed by means of UDPipe models (Straka and Straková, 2017) trained on texts by the same author, and then manually corrected by Latin language experts.
Our author-specific models are trained on *Opera Latina* (Denooz, 2004), a corpus which has been manually annotated at the *Laboratoire d'Analyse Statistique des Langues Anciennes* (LASLA) of the University of Liège since 1961.[10] Based on an agreement with LASLA, the *Opera Latina* corpus cannot be released to the public, but we are allowed to use it to create models for NLP tasks. Thus, we convert the original space-separated format of the *Opera Latina* into the field-based CoNLL-U format,[11] on which we train annotation models using the UDPipe pipeline.[12] These models are then run on the raw texts extracted from the Perseus files,[13] which are originally in XML format, after removing punctuation. Finally, the outputs of our automatic annotation are manually checked and corrected by two annotators; any doubts are resolved by a third Latin language expert.
Figure 1 and Figure 2 show examples of our CoNLL-U-formatted training and test data respectively. Please note that our training and test data lack any tagging of syntactic dependencies or morphological features, since EvaLatin 2020 does not focus on the corresponding tasks; besides, tree-structured syntactic data are not available from the *Opera Latina* corpus.

3.1. Training data

The texts provided as training data are by five Classical authors: Caesar, Cicero, Seneca, Pliny the Younger and Tacitus. For each author we release around 50,000 annotated tokens, for a total of almost 260,000 tokens. Each author is represented by prose texts: 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 dataset of EvaLatin 2020.

AUTHORS	TEXTS	# TOKENS
Caesar	De Bello Gallico	44,818
Caesar	De Bello Civili (book II)	6,389
Cicero	Philippicae (books I-XIV)	52,563
Seneca	De Beneficiis	45,457
Seneca	De Clementia	8,172
Pliny the Younger	Epistulae (books I-VIII)	50,827
Tacitus	Historiae	51,420
TOTAL		259,646

Table 1: Texts distributed as training data.

3.2. Test data

Tokenization is a central issue in the evaluation of Lemmatization and PoS tagging: as each annotation system possibly applies different tokenization rules, these might lead to outputs which are difficult to compare to each other. In order to avoid such problem, we provide our test data in an already tokenized format, one token per line, with a white line separating each sentence.
Our test data consist only of tokenized words, but neither lemmas nor PoS tags, as these have to be added by the participating systems submitted for the evaluation. The composition of the test dataset for the *Classical* subtask is given in Table 2. Details for the data distributed in the *Cross-genre* and *Cross-time* subtasks are reported in Tables 3 and 4 respectively.

AUTHORS	TEXTS	# TOKENS
Caesar	De Bello Civili (book I)	10,898
Cicero	In Catilinam	12,564
Seneca	De Vita Beata	7,270
Seneca	De Providentia	4,077
Pliny the Younger	Epistulae (book X)	9,868
Tacitus	Agricola	6,737
Tacitus	Germania	5,513
TOTAL		56,927

Table 2: Test data for the *Classical* subtask.

AUTHORS	TEXTS	# TOKENS
Horatius	Carmina	13,290

Table 3: Test data for the *Cross-genre* subtask.

AUTHORS	TEXTS	# TOKENS
Thomas Aquinas	Summa Contra Gentiles (part of Book IV)	11,556

Table 4: Test data for the *Cross-time* subtask.

4. Evaluation

The scorer employed for EvaLatin 2020 is a modified version of that developed for the *CoNLL18 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies*.[14] 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 and lemmas are evaluated and the final ranking is based on accuracy.

Each participant was permitted to submit runs for either one or all tasks and subtasks.
It was mandatory to produce one run according to the so-called "closed modality": the only annotated resources that could be used to train and tune the system were 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 annotated external data, like the Latin datasets present in the UD project, was allowed.
As for the baseline, we provided the participants with the scores obtained on our test data by UDPipe, using the

[10] http://web.philo.ulg.ac.be/lasla/textes-latins-traites/
[11] https://universaldependencies.org/format.html
[12] http://ufal.mff.cuni.cz/udpipe
[13] https://github.com/PerseusDL/canonical-latinLit
[14] https://universaldependencies.org/conll18/evaluation.html

```
# sent_id = 306
# text = Debere se suspicari simulata Caesarem amicitia quod
exercitum in Gallia habeat sui opprimendi causa habere
1   Debere      debeo     VERB  _ _ _ _ _ _
2   se          sui       PRON  _ _ _ _ _ _
3   suspicari   suspicor  VERB  _ _ _ _ _ _
4   simulata    simulo    VERB  _ _ _ _ _ _
5   Caesarem    Caesar    PROPN _ _ _ _ _ _
6   amicitia    amicitia  NOUN  _ _ _ _ _ _
7   quod        quod      SCONJ _ _ _ _ _ _
8   exercitum   exercitus NOUN  _ _ _ _ _ _
9   in          in        ADP   _ _ _ _ _ _
10  Gallia      Gallia    PROPN _ _ _ _ _ _
11  habeat      habeo     VERB  _ _ _ _ _ _
12  sui         sui       PRON  _ _ _ _ _ _
13  opprimendi  opprimo   VERB  _ _ _ _ _ _
14  causa       causa     NOUN  _ _ _ _ _ _
15  habere      habeo     VERB  _ _ _ _ _ _
```

Figure 1: Format of training data.

```
# 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: Format of test data.

Classical	
UDPipe-open 1	96.19 (0.89)
UDPipe-closed 1	95.90 (0.83)
JHUCB-closed 2	94.76 (1.04)
Leipzig-closed 1	94.60 (1.11)
JHUCB-closed 1	94.22 (1.38)
Baseline	72.26 (2.88)

Cross-genre	
UDPipe-open 1	87.13
JHUCB-closed 2	85.49
UDPipe-closed 1	85.47
JHUCB-closed 1	82.69
Leipzig-closed 1	81.69
Baseline	62.19

Cross-time	
UDPipe-open 1	91.01
UDPipe-closed 1	87.69
JHUCB-closed 2	85.75
Leipzig-closed 1	83.92
JHUCB-closed 1	83.76
Baseline	76.78

Table 5: Results of the Lemmatization task for the three subtasks in terms of accuracy. The number in brackets indicates standard deviation calculated among the seven documents of the test set for the *Classical* subtask.

model trained on the Perseus UD Latin Treebank[15] (Bamman and Crane, 2011), the same available in the tool's web interface.[16]

5. Participants and Results

A total of five teams are taking part in the PoS tagging task; three of them are also taking part in the Lemmatization task. All the teams have submitted runs for all three subtasks. Only one team (namely, UDPipe) has submitted a run following the open modality for each task and subtask, whereas the others have submitted runs in the closed modality, thus eschewing additional training data. In total, we have received five runs for the Lemmatization task and nine runs for the PoS tagging task. Details on the participating teams and their systems are given below:

- UDPipe, Charles University, Prague, Czech Republic. This team proposes a multi-task model jointly predicting both lemmas and PoS tags. The architecture is a bidirectional long short-term memory (BiLSTM) softmax classifier fed by end-to-end, character-level, pre-trained and contextualized word embeddings. In the run submitted for the open modality, they use all UD Latin treebanks as additional training data (Straka and Straková, 2020).
- Leipzig, Leipzig University, Germany. PoS tags are predicted with a gradient boosting framework fed with word embeddings pre-computed on a corpus of Latin texts of different genres and time periods. Lemmatization is instead based on a character-level translation performed by a long short-term memory (LSTM) sequence-to-sequence model (Celano, 2020).
- JHUBC, Johns Hopkins University and University of British Columbia, Canada. This team tests two systems for both Lemmatization and PoS tagging. The first one is an off-the-shelf neural machine translation toolkit, whereas the second puts together two different learning algorithms in an ensemble classifier: the aforementioned machine translation system and a BiLSTM sequence-to-sequence model (Wu and Nicolai, 2020).
- Berkeley, University of California, Berkeley, USA. The proposed model for the PoS tagging task consists in a grapheme-level LSTM network whose output is the input of a word-level BiLSTM network. This model is fed by a set of grapheme and word embeddings pre-trained on a corpus of over 23 million words (Bacon, 2020).
- TTLab, Goethe University, Frankfurt, Germany. This team tests three approaches to the PoS tagging task (Stoeckel et al., 2020): 1) an ensemble classifier based on a two-stage recurrent neural network combining the taggers MarMoT (Müller et al., 2013) and anaGo;[17] 2) a BiLSTM-CRF (conditional random fields) sequence tagger using pooled contextualized embeddings and a FLAIR character language model (Akbik et al., 2019); 3) another ensemble classifier combining the taggers MarMoT, anaGo, UDify (Kondratyuk and Straka, 2019) and UDPipe.

Tables 5 and 6 report the final rankings, showing the results in terms of accuracy, including our baseline. For each run, the team name, the modality and the run number are specified. Please note that for the *Classical* subtask the score corresponds to the macro-average accuracy obtained on the single text.

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

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

[17] https://github.com/vunb/anago-tagger

Classical	
UDPipe-open 1	96.74 (0.65)
UDPipe-closed 1	96.65 (0.63)
TTLab-closed 2	96.34 (0.60)
Leipzig-closed 1	95.52 (0.65)
TTLab-closed 3	95.35 (0.85)
JHUCB-closed 2	94.15 (0.64)
TTLab-closed 1	93.24 (0.92)
JHUCB-closed 1	92.98 (1.27)
Berkeley-closed 1	90.65 (1.98)
Baseline	70.25 (1.65)

Cross-genre	
UDPipe-open 1	91.11
TTLab-closed 2	90.64
UDPipe-closed 1	90.15
Leipzig-closed 1	88.54
JHUCB-closed 2	88.40
TTLab-closed 3	86.95
TTLab-closed 1	83.88
JHUCB-closed 1	82.93
Berkeley-closed 1	73.47
Baseline	62.96

Cross-time	
UDPipe-open 1	87.69
TTLab-closed 3	87.00
UDPipe-closed 1	84.93
Leipzig-closed 1	83.96
TTLab-closed 2	82.99
JHUCB-closed 1	82.62
TTLab-closed 1	81.38
JHUCB-closed 2	80.78
Berkeley-closed 1	76.62
Baseline	67.58

Table 6: Results of the PoS tagging task for the three subtasks in terms of accuracy. The number in brackets indicates standard deviation calculated among the seven documents of the test set for the *Classical* subtask.

6. Discussion

All the participating teams employ deep learning, and largely overcome the baseline. Systems mainly adopt LTSM networks, often in a bidirectional variant. Two teams also test the efficiency of ensemble classifiers, and one team a neural machine translation approach. Different types of embeddings are adopted: for example, grapheme embeddings, word embeddings, contextualized embeddings. In many cases, these embeddings are trained specifically for EvaLatin 2020 starting from large collections of Latin texts available online.

Not surprisingly, the addition of annotated data to the training set proves to be beneficial: in particular, an increase in accuracy is registered in the *Cross-genre* (+1.64 points of accuracy with respect to the best system in the closed modality) and *Cross-time* (+3.32 points of accuracy with respect to the best system in the closed modality) subtasks of the Lemmatization task.

The standard deviation among the texts of the test set in the *Classical* subtask fluctuates between 0.83 and 1.30 in the Lemmatization task, and between 0.60 and 1.98 in the PoS tagging task. With regard to the Lemmatization task, the easiest text to tackle for all the systems is *In Catilinam* by Cicero (accuracy ranging from 95.94 to 97.61), followed by the first book of the *De Bello Civili* by Caesar (accuracy ranging from 95.66 to 96.94). In the PoS tagging task, the situation is reversed: all the systems obtain better scores on the *De Bello Civili* (accuracy ranging from 93.08 to 97.91) than on *In Catilinam* (accuracy ranging from 93.02 to 97.44).

All the systems suffer from the shift to a different genre or to a different time period with a drop in the performances which, in some cases, exceeds 10 points. Taking a more in-depth look at the results, we can notice that, in general, the participating systems perform better on the Medieval text by Thomas Aquinas than on the Classical poems by Horace in the Lemmatization task, whereas the opposite is true for the PoS tagging task.

As for Lemmatization, Thomas Aquinas presents a less rich and varied vocabulary with respect to Horace: the lemma/token ratio is 0.09 and the percentage of out-of-vocabulary lemmas (i. e. lemmas not present in the training data) is 26%, while in the *Carmina* the lemma/token ratio is 0.26 and the percentage of out-of-vocabulary lemmas is 29%.

As for PoS tagging, Thomas Aquinas proves to be more challenging than Horace. This is probably due to the higher percentage and different distribution of tokens belonging to the categories of prepositions (ADP), conjunctions (CCONJ and SCONJ), auxiliaries (AUX) and numerals (NUM), as a consequence of a different textual and syntactic structure (with respect to the training set) that is more similar to that of modern Romance languages.

In particular, in Thomas Aquinas we observe a more frequent use of prepositional phrases: in Classical Latin, case inflection alone often suffices to convey the syntactic role of a noun phrase, whereas in the same context Medieval Latin might prefer that same phrase to be introduced by a preposition, extending a trend that is already present in Classical Latin (Palmer, 1988). We also find a greater number of subordinate clauses introduced by subordinating conjunctions (for example, the Classical construction of *Accusativus cum infinitivo* tends to be replaced by subordinate clauses introduced by subordinating conjunctions like *quia/quod/ut* 'that' (Bamman et al., 2008)), as well as of coordinated structures with coordinating conjunctions, the latter fact being possibly due to the very infrequent use of the enclitic particle *-que* 'and'. As for auxiliaries, their high number in the text of Thomas Aquinas is due to the fact that its annotation, carried out in the context of the *Index Thomisticus* Treebank (IT-TB) project (Passarotti, 2019), strictly follows the UD guidelines, so that the AUX tag is applied also to verbal copulas. This rule does not apply to the other texts employed in EvaLatin 2020, thus causing a discrepancy in the annotation criteria. Finally, the high occurrence of numerals is caused by the frequent use of biblical quotations (e. g. *Iob 26 14* 'Book of Job, chapter 26, verse 14', from *Summa contra Gentiles*, book 4, chapter 1, number 1).

7. Conclusion

This paper describes the first edition of EvaLatin, an evaluation campaign dedicated to NLP tools and methods for the Lemmatization and PoS tagging of the Latin language.

The call for EvaLatin 2020 has been spurred by the realization that times are mature enough for such an initiative. Indeed, despite the growing amount of linguistically annotated Latin texts which have become available over the last decades, today large collections of Latin texts are still lacking any layer of linguistic annotation, a state of affairs that

prevents users from taking full advantage of digital corpora for Latin.

One aspect that heavily impacts on any NLP task for Latin is the high degree of variability of the texts written in this language, due to its wide diachronic and diatopic diversity, which spans across several literary genres all over Europe in the course of more than two millennia. Just because we need to understand how much this aspect of Latin affects NLP, two subtasks dedicated respectively to the cross-genre and cross-time evaluation of data have been included in EvaLatin 2020.

If it holds true that variation is a challenging issue that affects NLP applications for Latin, one advantage of dealing with Latin data is that Latin is a dead language, thus providing a substantially closed corpus of texts (contemporary additions are just a few, like for instance the documents of the Vatican City or song lyrics (Cecchini et al., forthcoming)). This warrants us to speak of a possible complete linguistic annotation of all known Latin documents in the future.

In the light of such considerations, we have decided to devote the first edition of EvaLatin to Lemmatization and PoS tagging, as we feel the need to understand the state of the art of these two fundamental annotation layers for what concerns Latin.

We hope that the results of our evaluation campaign will help the community move towards the enhancement of an ever-increasing number of Latin texts by means of Lemmatization and PoS tagging as a first step towards a full linguistic annotation that includes also morphological features and syntactic dependencies, and that it will also help foster interest for Latin among the NLP community, confronting the challenge of portability of NLP tools for Latin across time, place and genres.

8. Acknowledgments

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 also wish to thank Giovanni Moretti for his technical assistance.

9. Bibliographical References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In Waleed Ammar, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, MN, USA, June. Association for Computational Linguistics (ACL).

Bacon, G. (2020). Data-driven Choices in Neural Part-of-Speech Tagging for Latin. In Rachele Sprugnoli et al., editors, *Proceedings of the* LT4HALA *2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (*LREC *2020)*, Paris, France, May. European Language Resources Association (ELRA).

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.

Bamman, D., Passarotti, M., and Crane, G. (2008). A Case Study in Treebank Collaboration and Comparison: Accusativus cum Infinitivo and Subordination in Latin. *The Prague Bulletin of Mathematical Linguistics*, 90(1):109–122, December.

Cecchini, F. M., Franzini, G. H., and Passarotti, M. C. (forthcoming). Verba Bestiae: How Latin Conquered Heavy Metal. In Riitta Valijärvi, et al., editors, *Multilingual Metal: Sociocultural, Linguistic and Literary Perspectives on Heavy Metal Lyrics*, Emerald Studies in Metal Music and Culture. Emerald group publishing, Bingley, UK.

Celano, G. (2020). A Gradient Boosting-Seq2Seq System for Latin PoS Tagging and Lemmatization. In Rachele Sprugnoli et al., editors, *Proceedings of the* LT4HALA *2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (*LREC *2020)*, Paris, France, May. European Language Resources Association (ELRA).

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

Kondratyuk, D. and Straka, M. (2019). 75 Languages, 1 Model: Parsing Universal Dependencies Universally. In Kentaro Inui, et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (*EMNLP-IJCNLP*)*, pages 2779–2795, Hong Kong, China, November. Association for Computational Linguistics (ACL).

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In David Yarowsky, et al., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, WA, USA, October. Association for Computational Linguistics (ACL).

Palmer, L. R. (1988). *The Latin language*. Oklahoma University Press, Norman, OK, USA. Reprint.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology*, volume 10 of *Age of Access? Grundfragen der Informationsgesellschaft*, pages 299–320. De Gruyter Saur, Munich, Germany, August.

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`.

Smith, D. A., Rydberg-Cox, J. A., and Crane, G. R. (2000). The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Stoeckel, M., Henlein, A., Hemati, W., and Mehler, A. (2020). Voting for PoS tagging of Latin texts: Using the flair of FLAIR to better Ensemble Classifiers by Exam-

ple of Latin. In Rachele Sprugnoli et al., editors, *Proceedings of the* LT4HALA *2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (*LREC *2020)*, Paris, France, May. European Language Resources Association (ELRA).

Straka, M. and Straková, J. (2017). Tokenizing, PoS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Jan Hajič et al., editors, *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics (ACL). Available at `http://www.aclweb.org/anthology/K/K17/K17-3009.pdf`.

Straka, M. and Straková, J. (2020). UDPipe at EvaLatin 2020: Contextualized Embeddings and Treebank Embeddings. In Rachele Sprugnoli et al., editors, *Proceedings of the* LT4HALA *2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (*LREC *2020)*, Paris, France, May. European Language Resources Association (ELRA).

Wu, W. and Nicolai, G. (2020). JHUBC's Submission to LT4HALA EvaLatin 2020. In Rachele Sprugnoli et al., editors, *Proceedings of the* LT4HALA *2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (*LREC *2020)*, Paris, France, May. European Language Resources Association (ELRA).