Overview of the EvaLatin 2024 Evaluation Campaign

Rachele Sprugnoli[1],Federica Iurescia[2], Marco Passarotti[2]
Università di Parma, viale D'Azeglio, 85, 43125 Parma, Italy[1],
Università Cattolica del Sacro Cuore, largo A. Gemelli 1, 20123 Milan, Italy[2]
rachele.sprugnoli@unipr.it, {federica.iurescia, marco.passarotti}@unicatt.it

Abstract

This paper describes the organization and the results of the third edition of EvaLatin, the campaign for the evaluation of Natural Language Processing tools for Latin. The two shared tasks proposed in EvaLatin 2024, i. e. Dependency Parsing and Emotion Polarity Detection, are aimed to foster research in the field of language technologies for Classical languages. The shared datasets are described and the results obtained by the participants for each task are presented and discussed.

Keywords: Latin, evaluation, dependency parsing, emotion polarity detection

1. Introduction

EvaLatin 2024 is the third edition of the campaign devoted to the evaluation of Natural Language Processing (NLP) tools for the Latin language. As in 2020 (Sprugnoli et al., 2020a) and 2022 (Sprugnoli et al., 2022), EvaLatin is proposed as part of the *Workshop on Language Technologies for Historical and Ancient Languages* (LT4HALA), co-located with LREC COLING 2024.[1] Similar to what happens in other international evaluation campaigns, participants were provided with shared test data that are made freely available for research purposes to encourage further improvement of language technologies for Latin. Shared scripts were also provided. 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, together with the University of Parma, Italy.

2. Tasks

EvaLatin 2024 is organized around 2 tasks:

- **Dependency Parsing**: the aim of the task is to provide syntactic analysis of Latin texts following the Universal Dependencies (UD) framework (de Marneffe et al., 2021). The output submitted by the participants is a CoNLL-U file with indications of the syntactic head and of the dependency relations in the fields 7 (`HEAD`) and 8 (`DEPREL`) respectively.
- **Emotion Polarity Detection**: the aim of the task is to identify the polarity conveyed by each sentence in the input text, taking into consideration both the vocabulary used by the author and the images that are evoked in the text (Sprugnoli et al., 2023). More specifically, the question to be answered is: which of the following classes best describes how are the emotions conveyed by the poet in the sentence under analysis?
  - `positive`: the only emotions that are conveyed in the text are positive, or positive emotions are clearly prevalent;
  - `negative`: the only emotions that are conveyed in the text are negative, or negative emotions are clearly prevalent;
  - `neutral`: there are no emotions conveyed by the text;
  - `mixed`: lexicon and evoked images produce opposite emotions; it is not possible to find a clearly prevailing emotion polarity.

  Sentences are provided in their original order in the source text.

3. Data

No specific training data are released for the Dependency Parsing task but participants are free to make use of any (kind of) resource they consider useful for the task, including the Latin treebanks already available in the UD collection. In this regard, one of the challenges of this task is to understand which treebank (or combination of treebanks) is the most suitable to deal with new test data.

Also for the Emotion Polarity Detection task, no training data are released but an annotation sample and a manually created polarity lexicon are provided. Also in this task, participants are free to

[1] https://lrec-coling-2024.org/
[2] https://github.com/CIRCSE/LT4HALA/tree/master/2024/data_and_doc
[3] https://centridiricerca.unicatt.it/circse/en.html

LT4HALA 2024@LREC-COLING-2024, pages 190–197

```
# 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   S   Polarity=Neg    _   _   _   LiLaflcat=i
2   multum  multum  ADV M   Degree=Pos  _   _   _   LASLAVariant=2|LiLaflcat=i
3   frumento    frumentum   NOUN    A2  Case=Abl|Gender=Neut|InflClass=IndEurO|Number=Sing  _   _   _   LiLaflcat=n2
4   sed sed CCONJ   S   _   _   _   _   LiLaflcat=i
5   maximam magnus  ADJ C1  Case=Acc|Degree=Abs|Gender=Fem|InflClass=IndEurA|Number=Sing    _   _   _   LiLaflcat=n6
6   partem  pars    NOUN    A3  Case=Acc|Gender=Fem|InflClass=IndEurI|Number=Sing   _   _   _   LiLaflcat=n3
7   lacte   lac NOUN    A3  Case=Abl|Gender=Masc|InflClass=IndEurI|Number=Sing  _   _   _   LiLaflcat=n3
8   atque   atque   CCONJ   S   _   _   _   _   LASLAVariant=1|LiLaflcat=i
9   pecore  pecus   NOUN    A3  Case=Abl|Gender=Neut|InflClass=IndEurX|Number=Sing  _   _   _   LASLAVariant=1|LiLaflcat=n3
10  uiuunt  uiuo    VERB    B3  Aspect=Imp|InflClass=LatX|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act   _   _   _   LiLaflcat=v3
11-12   multumque   _   _   _   _   _   _   _   _
11  multum  multum  ADV M   Degree=Pos  _   _   _   LASLAVariant=2|LiLaflcat=i
12  que que CCONJ   S   _   _   _   _   LiLaflcat=i
13  sunt    sum AUX B6  Aspect=Imp|InflClass=LatAnom|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin  _   _   _   LASLAVariant=1|LiLaflcat=v6
14  in  in  ADP R   AdpType=Prep    _   _   _   LiLaflcat=i
15  uenationibus    uenatio NOUN    A3  Case=Abl|Gender=Fem|InflClass=IndEurX|Number=Plur   _   _   _   LiLaflcat=n3
```

Figure 1: Example of the test data format.

pursue the approach they prefer, including unsupervised and/or cross-language ones.

Both tasks aim to improve a state of the art that is currently not optimal. With regard to Dependency Parsing, UD treebanks currently show different degrees of harmonization, and Latin is not an exception in this respect (Gamba and Zeman, 2023). With regard to Emotion Polarity Detection, there are no available training data for Latin yet, as this is an unexplored territory for this language. It is important to notice that in both tasks, some texts include punctuation, some do not, as this is the actual state of the art for Latin treebanks and corpora; for example, the LASLA corpus (see Section 3.1 for further details) does not include punctuation (Denooz, 2004). The diversity of the data currently available for both tasks is an issue we are aware of, and that needs to be addressed. This evaluation campaign aims at addressing this issue, and among the desired outcomes there are strategies to deal with it successfully.

3.1. Test Data

Texts provided as test data for the Dependency Parsing task are by 2 Classical authors (Seneca and Tacitus) for a total of more than 13,000 tokens. Each author is taken as specimen of one specific text genre: Seneca for poetry, more specifically for tragedy, with *Hercules Furens* (more than 7,000 tokens), composed in 1st century AD; Tacitus for prose, more specifically historical and ethnographic treatise, with *Germania* (nearly 6,000 tokens), written in 1st century AD. Precise numbers are given in Tables 1 and 2, while an example of the format of test data is given in Figure 1. Data are taken from the LASLA corpus, a linguistic resource manually annotated since 1961 by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège, Belgium.[4] Original data were converted into the annotation formalism of the UD project and manually annotated for dependency relations. Data are distributed in the CoNLL-U format.[5] Following such format, the annotations are plain text files having the `.conllu` extension and encoded in UTF-8.

AUTHOR	TEXT	#TOKENS
Seneca	Hercules Furens	7,711

Table 1: Test data for poetry.

AUTHOR	TEXT	#TOKENS
Tacitus	Germania	5,669

Table 2: Test data for prose.

Texts provided as test data for the Emotion Polarity Detection task are by 3 authors for a total of 297 sentences (around 100 sentences for each author):

- Seneca, with the final part (lines 1,175-1,344)[6] of the tragedy *Hercules Furens*, composed in 1st century AD;
- Horace, with 16 odes (4 for each book that makes up *Carmina*), composed in 1st century AD;
- Giovanni Pontano, with 12 poems taken from the work *Neniae*, composed in the 15th century.

Test data for the task of Emotion Polarity Detection are distributed in .tsv format: the first column contains a sentence ID and the second the text to be tagged. Tables 3, 4, 5 report the precise number of sentences for each text, while Figure 2 provide an example of the format. Data by Seneca and Horace are taken from the LASLA corpus, while texts by Pontano are taken from the *Poeti d'Italia in*

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

[5] https://universaldependencies.org/format.html

[6] Line numbers according to the following edition: Fitch, J.G. (2018). *Seneca. Tragedies, Volume I: Hercules. Trojan Women. Phoenician Women. Medea. Phaedra.* Cambridge (MA): Harvard University Press.

lingua latina website.[7] For this reason, Pontano's texts have punctuation while those of Seneca and Horace do not.

AUTHOR	TEXT	#SENT.
Seneca	Hercules Furens (lines 1,175-1,344)	103

Table 3: Test data by Seneca.

AUTHOR	ODE (BOOK_POEM)}	#SENT.
Horace	I_2	7
Horace	I_14	8
Horace	I_28	9
Horace	I_38	2
Horace	II_3	6
Horace	II_11	7
Horace	II_14	3
Horace	II_16	10
Horace	III_2	5
Horace	III_10	4
Horace	III_18	2
Horace	III_24	7
Horace	IV_1	11
Horace	IV_10	1
Horace	IV_12	8
Horace	IV_13	6
TOTAL		**96**

Table 4: Test data by Horace.

AUTHOR	NENIAE	#SENT.
Pontano	I	8
Pontano	II	11
Pontano	III	9
Pontano	IV	14
Pontano	V	6
Pontano	VI	7
Pontano	VII	11
Pontano	VIII	5
Pontano	IX	4
Pontano	X	9
Pontano	XI	8
Pontano	XII	6
TOTAL		**98**

Table 5: Test data by Pontano.

4. Evaluation

Two different scorers are used for the two shared tasks proposed at EvaLatin 2024.

```
100 uelox amoenum saepe Lucretilem mutat Lycaeo Faunus
et igneam defendit aestatem capellis usque meis pluuios
que uentos
101 inpune tutum per nemus arbutos quaerunt latentis et
thyma deuiae olentis uxores mariti nec uiridis metuunt
colubras nec Martialis haediliae lupos utcumque dulci
Tyndari fistula ualles et Vsticae cubantis leuia
personuere saxa
102 di me tuentur dis pietas mea et musa cordi est
103 hic tibi copia manabit ad plenum benigno ruris
honorum opulenta cornu
104 hic in reducta ualle caniculae uitabis aestus et
fide Teia dices laborantis in uno Penelopen uitream que
Circen
105 hic innocentis pocula Lesbii duces sub umbra nec
Semeleius cum Marte confundet Thyoneus proelia nec
metues proterum suspecta Cyrum ne male dispari
incontinentis iniciat manus et scindat haerentem coronam
crinibus inmeritam que uestem
```

Figure 2: Example of the data format for the Emotion Polarity Detection task.

- The scorer employed for the evaluation of the Dependency Parsing task is the one developed for the *CoNLL18 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies* (Zeman et al., 2018).[8] The evaluation starts by aligning the system-produced tokens to the gold standard one; given that we provide test data already sentence-splitted and annotated with morpho-grammatical information, the alignment for tokens, sentences, words, UPOS, UFeats and lemmas should be perfect (i. e. 100.00). Then, CLAS (Content-Word Labeled Attachment Score)[9] and LAS (Labeled Attachment Score)[10] are evaluated in terms of Precision, Recall, F1 and Aligned Accuracy.[11]

- The scorer for the Emotion Polarity Detection task is a Python script that calculates precision, recall and F1 measure for each class assigned at sentence level but also accuracy, macro-average and weighted average. The scorer is available on the EvaLatin web page[12].

As for the baseline, for the Dependency Parsing

[7] https://www.poetiditalia.it/public/

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

[9] CLAS is the labeled F1- score over all relations except those involving function words (aux, case, cc, clf, cop, det, mark) and punctuation (punct). For further details, see (Nivre and Fang, 2017).

[10] LAS is the percentage of tokens assigned both the correct DEPREL and HEAD. For further details, see (Buchholz and Marsi, 2006).

[11] The scorer computes also the Unlabeled Attachment Score (UAS), that is the percentage of tokens assigned the correct HEAD; the Morphology-aware Labeled Attachment Score (MLAS), that is CLAS extended with evaluation of POS tags and morphological features; the Bi-Lexical dependency score (BLEX) that combines content-word relations with lemmatization, but not with POS tags and features. These 3 metrics are not taken into account for this shared task.

[12] https://github.com/CIRCSE/LT4HALA/blob/master/2024/scorer-emotion.py

task we provide the scores obtained on the test data using UDPipe 2 (Straka et al., 2016) with the model trained on the Perseus Universal Dependencies Latin Treebank[13] (Bamman and Crane, 2011), as it is available from the tool's web interface.[14]

For the Emotion Polarity Detection task, we calculate the baseline by applying a lexicon-based approach to the test data. More specifically, a sentence score is computed by summing the polarity values of all lemmas. Polarity values are taken from LatinAffectus v.4, a prior polarity sentiment lexicon for Latin (Sprugnoli et al., 2020b). The label `positive` is assigned to all the sentences with score above 0 and the label `negative` to sentence for which the score is below 0. For scores equal to 0, we attribute `neutral` to sentences where all words have a score of 0 and `mixed` where positive and negative scores are balancing each other out to a total net sum of 0.

5. Results and Discussion

Three teams took part in the Dependency Parsing task and other three teams took part in the Emotion Polarity Detection task. Regarding the latter, one team did not submit the report and therefore it will not be included in this overview.

5.1. Dependency Parsing

Details on the participating teams and their systems for the Dependency Parsing task are given below:

- Behr. This team submitted one run, leveraging historical sentence embeddings generated via SBERT (Reimers and Gurevych, 2019) as a pivotal strategy to confront the challenge of developing a parser capable of achieving accurate performance irrespective of the chronological period of the Latin texts within the test data (Behr, 2024).

- KU Leuven - Brepols CTLO. The team submitted two runs. The first run adopts a span-span prediction methodology, grounded in Machine Reading Comprehension (MRC), and utilizes LaBERTa (Riemenschneider and Frank, 2023), a RoBERTa model pre-trained specifically on Latin corpora. This run yields meaningful outcomes. Conversely, the second, more exploratory run operates at the token-level, employing a span-extraction approach inspired by the Question Answering (QA) task. This model fine-tunes a DeBERTa model (He et al., 2023) pre-trained on Latin datasets, but the results are extremely low (Mercelis, 2024).

- ÚFAL LatinPipe. Also this team submitted two distinct runs employing a system comprising a fine-tuned concatenation of base and large pre-trained Language Models. Both runs utilize a dot-product attention head for parsing and softmax classification heads for morphology, enabling the joint learning of dependency parsing and morphological analysis. Training data are sampled from seven publicly available Latin treebanks, with additional efforts focused on harmonizing annotations to attain a more cohesive annotation style. The difference between the two runs lies in the treatment of punctuation, that is present in some of the treebanks used for the training set, but is absent in the shared test data (Straka et al., 2024).

Table 6 and 7 show the final ranking. The results are provided in terms of F1, including the baseline. The majority of the submitted runs demonstrate clear improvements over the baseline, with the sole exception being the exploratory KU Leuven - Brepols CTLO run 2. Performances remain consistent across diverse text genres (poetry and prose) and evaluation metrics (LAS and CLAS). The best performing run, ÚFAL LatinPipe_1, exhibits a nearly 25% enhancement over the baseline.

The Dependency Parsing task underscores two primary challenges encountered in the development of models for parsing Latin data: firstly, the variability in the annotation styles across available Latin treebanks, posing a challenge to model training; and secondly, the extensive temporal scope and diverse genres present in Latin texts. The teams addressed these challenges relying on Large Language Models (LLMs) to navigate through them effectively. Behr's approach explicitly targets model performance across different epochs, while KU Leuven - Brepols CTLO adopts a span extraction method, drawing inspiration from QA tasks. However, this experimentation reveals limitations in current QA implementations regarding dependency head prediction, indicating the need for further investigation. The ÚFAL LatinPipe team employs LLMs, conducting data harmonization and fine-tuning on various combinations of treebanks, resulting in superior performance.

Presently, leveraging LLMs, fine-tuning on treebank ensembles, and harmonizing inconsistent annotations emerge as the most encouraging strategies for Dependency Parsing in Latin. This shared task demonstrates promising solutions to parsing challenges: harmonization addresses annotation style diversity, while ensemble approaches mitigate portability issues.

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

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

TEAM	F1 POETRY	TEAM	F1 PROSE
ÚFAL LatinPipe_1	74.53	ÚFAL LatinPipe_1	73.19
ÚFAL LatinPipe_2	69.59	ÚFAL LatinPipe_2	68.76
Behr	67.87	Behr	66.53
KU Leuven - Brepols CTLO run 1	57.34	KU Leuven - Brepols CTLO run 1	63.71
BASELINE	48.51	BASELINE	51.81
KU Leuven - Brepols CTLO run 2	5.34	KU Leuven - Brepols CTLO run 2	3.78

Table 6: Dependency Parsing results in terms of CLAS.

TEAM	F1 POETRY	TEAM	F1 PROSE
ÚFAL LatinPipe_1	75.75	ÚFAL LatinPipe_1	77.41
ÚFAL LatinPipe_2	70.68	ÚFAL LatinPipe_2	73.07
Behr	68.33	Behr	69.72
KU Leuven - Brepols CTLO run 1	59.02	KU Leuven - Brepols CTLO run 1	67.32
BASELINE	50.36	BASELINE	56.73
KU Leuven - Brepols CTLO run 2	5.44	KU Leuven - Brepols CTLO run 2	3.70

Table 7: Dependency Parsing results in terms of LAS.

5.2. Emotion Polarity Detection

Details on the participating teams and their systems for the Emotion Polarity Detection task are given below:

- Nostra Domina. This team submitted two runs employing data augmentation algorithms and various Latin LLMs in a neural architecture. Both runs ended up using the same augmentation procedure and LLM, but they differed in their encoder. The first and second runs include a Transformer encoder and BiLSTM encoder, respectively (Bothwell et al., 2024).

- TartuNLP. The team submitted two runs, both based on XLM-RoBERTa, the multilingual version of RoBERTa (Conneau et al., 2020). To deal with the lack of training data, they created two datasets, one by applying LatinAffectus v.4 and the other by using OpenAI's GPT-4. To make the training faster, avoid catastrophic forgetting and capitalize on knowledge transfer, they used parameter efficient fine-tuning methods employing language adapters and multi-stage training. (Dorkin and Sirts, 2024).

Table 8 reports the final ranking, showing the results in terms of F1, including the baseline. Given that Horace and Pontano's test set is made up of various texts, the value reported in the table corresponds to the macro-average F1.

The difficulty of the Emotion Polarity Detection task is evident by looking at the results reported in Table 8. In fact, the baseline is not beaten by every submitted run and it even obtains the best F1 on Pontano's poems. Among the participating systems there is not a single one that performs better than the others on all 3 authors. The TartuNLP_1 run (fine-tuned on a dataset annotated by applying LatinAffectus v.4) is the best performing one on Seneca and Pontano but records the lowest F1 macro-average on Horace for which, on the contrary, the best run is NostraDomina_1 (that uses PhilBERTa-based embeddings (Riemenschneider and Frank, 2023), a Transformer encoder, and a dataset derived from Gaussian clustering). The performances at class level are also different: the NostraDomina team's runs have better results in recognizing positive sentences, while the TartuNLP runs record higher F1 for negative sentences. For all the runs, however, the mixed class is the most difficult to recognize.

In general, there are two important trends that all runs have in common. On the one hand the use of data augmentation methods to make up for the lack of training data, on the other the use of neural models, in particular LLMs.

6. Conclusion

This paper has provided an overview of the NLP tasks addressed in the third edition of the EvaLatin evaluation campaign, namely: Dependency Parsing and Emotion Polarity Detection.

Compared to the tasks of the previous editions of EvaLatin (Lemmatization, PoS tagging, Morphological Feature Identification), the accuracy rates of the tools that participated in the evaluation campaign are lower. This is due both to the higher degree of difficulty of the tasks themselves and to the limited (or nonexistent) availability of training sets to build machine-learning models in a (semi-)supervised manner. To overcome this limitation, the participating systems made extensive use of pre-trained models equipped with knowledge that

TEAM	SENECA	TEAM	HORACE	TEAM	PONTANO
TartuNLP_1	0.26	Baseline	0.40	NostraDomina_1	0.42
Baseline	0.25	TartuNLP_1	0.31	TartuNLP_2	0.32
TartuNLP_2	0.25	TartuNLP_2	0.30	NostraDomina_2	0.31
NostraDomina_2	0.14	NostraDomina_1	0.29	Baseline	0.29
NostraDomina_1	0.12	NostraDomina_2	0.21	TartuNLP_1	0.24

Table 8: Emotion Polarity Detection results in terms of F1.

can be fine-tuned for specific NLP tasks by using the data provided by annotated corpora, which, in an ideal virtuous circle, represent one of the outcomes of the application of NLP tools. In such respect, one of the objectives of EvaLatin was (and still remains) providing a venue for developing and evaluating language models for various NLP tasks to support the building of more and larger annotated corpora for Latin.

The task dedicated to Dependency Parsing has shown that the state of the art is good, although still far from optimal. The problem of model portability across different literary genres, albeit roughly distributed on a binary classification (prose and poetry), remains an open challenge, with a substantial impact on the automatic processing of Latin texts, which exhibit a high degree of stylistic variability.

The task of Emotion Polarity Detection was a risky bet, given the scarcity of external resources that could be used, the absence of training sets, and the lack of previously available annotation guidelines. The low accuracy rates of the participating systems highlight the difficulty of the task, which is also due to the high degree of subjectivity intrinsic to the task itself and to the involvement of many different components (lexical, syntactic, encyclopedic, cultural) in determining the emotion evoked by a text.

Emotion Polarity Detection opens the door for EvaLatin to semantic analysis, which includes tasks such as Semantic Role Labeling and Word Sense Disambiguation. It is our intention to consider these types of NLP tasks for the future editions of the evaluation campaign.

7. Acknowledgements

The authors want to thank Lisa Sophie Albertelli, Lorenzo Augello, Roberta Buffolino, Giulia Calvi, Roberta Leotta and Marinella Testori for the annotation of test data for the Emotion Polarity Detection task and Giovanni Moretti for providing the scorer for the Emotion Polarity Detection task.

8. Bibliographical References

David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98, Berlin/Heidelberg, Germany. Springer. Preprint retrievable at http://www.cs.cmu.edu/~dbamman/pubs/pdf/latech2011.pdf.

Rufus Behr. 2024. Behr at EvaLatin 2024: Latin Dependency Parsing Using Historical Sentence Embeddings. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024)*, Torino, Italy. European Language Resources Association.

Stephen Bothwell, Abigail Swenor, and David Chiang. 2024. Nostra Domina at EvaLatin 2024: Improving Latin Polarity Detection through Data Augmentation. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024)*, Torino, Italy. European Language Resources Association.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Flavio Massimiliano Cecchini. 2021. *Formae reformandae*: for a reorganisation of verb form annotation in Universal Dependencies illustrated by the specific case of Latin. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 1–15, Sofia, Bulgaria. The Association for Computational Linguistics (ACL). Retrievable at https://aclanthology.org/2021.udw-1.1/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

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

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

Aleksei Dorkin and Kairit Sirts. 2024. TartuNLP at EvaLatin 2024: Emotion Polarity Detection. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024)*, Torino, Italy. European Language Resources Association.

Federica Gamba and Daniel Zeman. 2023. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. pages 7–16, Washington, D.C. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Wouter Mercelis. 2024. KU Leuven / Brepols-CTLO at EvaLatin 2024: Span extraction approaches for Latin dependency parsing. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024)*, Torino, Italy. European Language Resources Association.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

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

Rachele Sprugnoli, Francesco Mambrini, Marco Passarotti, and Giovanni Moretti. 2023. The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace. *IJCoL. Italian Journal of Computational Linguistics*, 9(9-1):53–71.

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

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020a. 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. European Language Resources Association (ELRA).

Rachele Sprugnoli, Marco Passarotti, Daniela Corbetta, and Andrea Peverelli. 2020b. Odi et Amo. creating, evaluating and extending sentiment lexicons for Latin. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3078–3086, Marseille, France. European Language Resources Association.

Milan Straka, Jan Hajič, and Jana Straková. 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, Portorož, Slovenia. European Language Resources Association (ELRA). Retrievable at https://aclanthology.org/L16-1680.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, BC, Canada. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic Analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024)*, Torino, Italy. European Language Resources Association.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 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. Association for Computational Linguistics.