Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 193–197
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

Transformer-based Part-of-Speech Tagging and Lemmatization for Latin

Krzysztof Wróbel[1,2], Krzysztof Nowak[3]
[1]Jagiellonian University, , [2]Enelpol, [3]Institute of Polish Language (Polish Academy of Sciences)
krzysztof@wrobel.pro, krzysztof.nowak@ijp.pan.pl

Abstract

The paper presents a submission to the EvaLatin 2022 shared task. Our system places first for lemmatization, part-of-speech and morphological tagging in both closed and open modalities. The results for cross-genre and cross-time sub-tasks show that the system handles the diachronic and diastratic variation of Latin. The architecture employs state-of-the-art transformer models. For part-of-speech and morphological tagging, we use XLM-RoBERTa large, while for lemmatization a ByT5 small model was employed. The paper features a thorough discussion of part-of-speech and lemmatization errors which shows how the system performance may be improved for Classical, Medieval and Neo-Latin texts.

Keywords: part-of-speech tagging, lemmatization, morphosyntactic tagging, Latin, transformers

1. Introduction

The performance of lemmatization and part-of-speech tagging tools is essential for Latin as it is for all historical languages. Due to relative scarcity of annotated data, newly developed tools may be expected to be effective or at least adaptable to handle Classical, Medieval, and Neo-Latin, despite the fact that their use spans over more than 15 centuries. The recent advancements in NLP technology along with increasing availability of large language models have opened new venues for computational Latin linguistics.

Corpus		Tokens	Sentences	Avg
EvaLatin 2022				
Train		320 355	15 785	20.29
Test	Classical	13 248	385	34.41
Test	Cross-genre	22 086	1 329	16.62
Test	Cross-time	9 174	246	37.29
EvaLatin 2020				
Test	Cross-genre	13 290	597	22.26
	Cross-time	11 556	883	13.09
UD Latin[1]		977 722	58 405	16.74
Lasla[2]		1 728 933	92 170	18.76

Table 1: Corpora used in the study

In this paper, we present our submission to the EvaLatin 2022 shared task (Sprugnoli et al., 2022). First, we briefly characterize the task, focusing on specific challenges the texts included in the test dataset posed. Next, we provide a detailed description of our system and describe its two modalities. Additionally, we show what data were used to enhance the performance of the open variant of the model and provide a thorough analysis of lemmatization and part-of-speech errors. We believe that the present system may be further adapted to address challenges of linguistic annotation of the Medieval and Neo-Latin texts.

2. Training and Test Data

The training dataset of the EvaLatin 2022 shared task contains prosaic texts of five authors composed between the 1st century BC and the beginning of the 2nd century AD. The test dataset includes works which represent various genres and periods of the Latin literature history. The Classical subtask consists of the VIIIth book of Livy's *Ab urbe condita*, a work which is arguably closest to the training data. Two texts in the Cross-Genre sub-task differ from the training data in their literary form and subject domain. The VIIIth and IXth books of the Ovid's epic poem contain narratives of Greek mythology. Pliny the Elder's *Naturalis Historia*, on the other hand, is an encyclopedic work in prose whose XXXVIIth book discusses properties of gemstones. Both texts contain a significant number of words of Greek origin: person and place names in case of *Metamorphoses* and rare terms regarding mineralogy in case of Pliny. The only text included in the Cross-time sub-task dataset is the *De Latinae Linguae Reparatione*, a Renaissance dialogue on history by Marcus Antonius Coccius Sabellicus (†1504). The major challenge seems to be its non-Classical orthography and a number of post-Classical proper names.

3. System Description

Our architecture is based on transformer models, as they are state-of-the-art in part-of-speech tagging and lemmatization. It builds on a morphosyntactic tagger KFTT (Wróbel, 2020) which won the PolEval 2020 task 1 competition (*Morphosyntactic tagging of Middle, New and Modern Polish*) and uses a transformer model contrary to its RNN-based predecessor KRNNT (Wróbel, 2017).

[2]UD corpora include 5 Latin treebanks in the Universal Dependencies format (Zeman, 2022).

[2]The LASLA corpus (Denooz, 2007) linked to the LiLa LemmaBank (Fantoli et al., 2022).

Task	Phase	UD Latin	LASLA	EvaLatin '22 Train	EvaLatin '20 X-Genre	EvaLatin '20 X-Time
POS	1	+	+		+	+
	2			+		
Feats	1	+				
	2			+		
Lemmatization	1	+	+		+	+
	2			+	+	+

Table 2: Corpora used in the *open modality* system

Part-of-speech and morphologic tagging are addressed with a transformer encoder model with a token classification head on top. The transformer, first, returns contextual embeddings of each token; next, a linear layer with softmax activation returns normalized scores for each tag seen in training.

In the lemmatization task, the system uses information about predicted parts of speech, but it does not use context of a word. It is solved with sequence to sequence model with input constructed as a word form and predicted part of speech.

In the *open modality* variant of the system, in which external resources can be employed (see Table 2), our models are first trained on a set of corpora that were annotated following different guidelines than the ones adopted in the present competition. In the next phase, the models are re-trained on the EvaLatin 2022 training dataset. Detailed information on each corpus can be consulted in the Table 1. The performance of the system in each task was evaluated using micro-averaged accuracy. 5% of the EvaLatin 2022 training data were used for validation.

For the POS and Feats tasks we used XLM-RoBERTa large (Conneau et al., 2020) – a multilingual encoder. Model training parameters were:

- batch size: 12
- epochs: 10,
- learning rate: 2e-5,
- sequence length: 256.

Lemmatization was performed with ByT5 small model (Xue et al., 2022) whose input are separate bytes of text. Initial experiments with subword models (e.g. mT5 (Xue et al., 2021)) showed worse accuracy. Model training parameters were the following:

- batch size: 128,
- epochs: 5,
- input sequence length: 48,
- output sequence length: 24,
- learning rate: 0.001.

In the *open modality* for the PoS and Feats tasks first training is performed for 2 epochs without early stopping.

All models here described are publicly available.[3]

[3] https://huggingface.co/enelpol/

4. Results

Our system performed best in every task in the competition. In the *closed modality* variant, it was ahead of the second best architecture by 0.9%-4.5% in the PoS task, by 25.5%-31.9% in the Feats task, and by 4.4%-11.0% in the Lemmatization task (Table 3).

Since the system is expected to be employed in Medieval and Neo-Latin corpus projects, it was essential to examine its performance in qualitative terms as well (Nowak et al., 2016). Therefore, we carefully analyzed tagging errors (1) to assess the impact of additional training data on the performance in the *open modality* and (2) to get insight into major challenges that language variation poses to the system. Due to space limitations, however, we only briefly discuss the results of the Lemmatization and PoS task.

4.1. Part-of-Speech Tagging

All texts combined, the PoS tagging errors affect in particular nominal categories, with ADJs misclassified as NOUNs or PROPNs, NOUNs as ADJs, and VERBs as ADJs (see Figure 1). The error distribution varies slightly between sub-tasks and modalities.

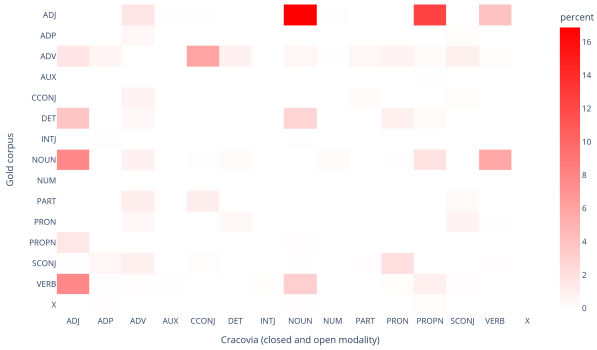

Figure 1: PoS Tagging: Confusion Matrix (*closed* and *open modalities*)

Generally, in the open version of our system, the quality of the PoS tagging improves significantly. The analysis shows (see Figure 2) that the use of annotated resources helps to distinguish NOUNs, PROPNs, VERBs from ADJs. We discuss major improvements below.

ADJ ↔ NOUN In both CLASSICAL and CROSS-GENRE sub-tasks, using supplementary

			KU-Leuven	Cracovia	
			closed	closed	open
CLASSICAL	Livy	POS	96.33	97.61	**97.99**
		Lemma	85.44	96.45	**97.26**
		Feats	69.91	95.42	**95.46**
CROSS-GENRE	Ovid	POS	94.66	94.78	**96.78**
		Lemma	87.22	93.05	**96.03**
		Feats	63.06	88.70	**88.81**
	Pliny	POS	89.96	94.47	**95.35**
		Lemma	85.75	90.19	**94.13**
		Feats	58.04	89.95	**90.06**
CROSS-TIME	Sabellicus	POS	92.11	**92.97**	92.70
		Lemma	84.60	91.68	**92.15**
		Feats	60.09	**86.53**	86.50

Table 3: Performance of the Cracovia system for POS, Lemmatization, and Feats tagging task

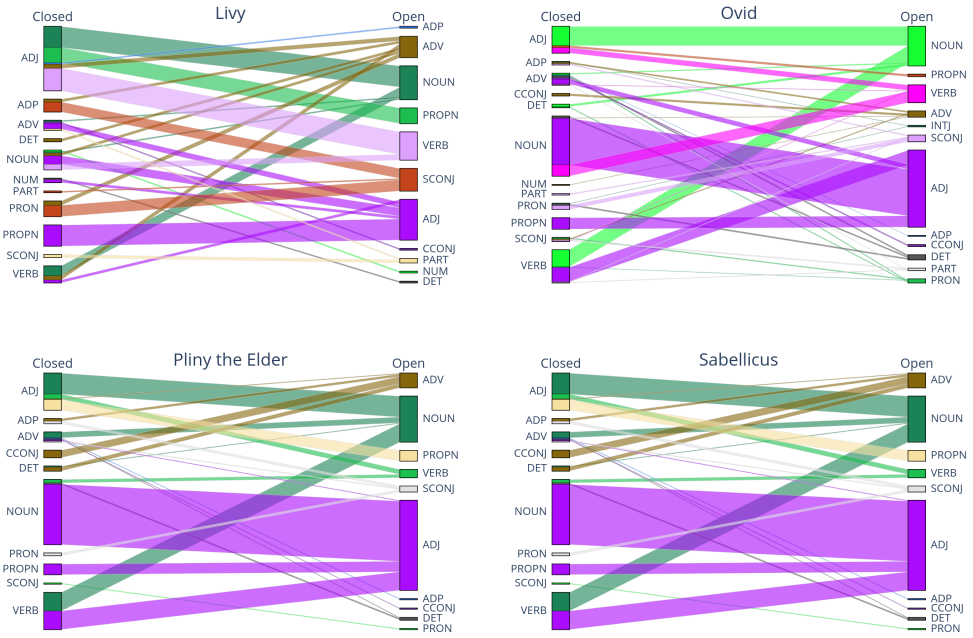

Figure 2: POS Tagging: Closed v. Open Modality

annotated resources leads to better discrimination between homonymous forms of nouns and adjectives, such as *iuuenis* 'young' : 'a young person', *securus*.ADJ 'safe' : *securis*.NOUN 'an axe' or *sacer*.ADJ 'sacred' : *sacrum*.NOUN 'a holy thing'. In the open modality, correct lemmas are assigned, for instance, to Greek-origin terms such as †*synechitus*.ADJ → *synechitis*.NOUN 'a kind of gemstone' or †*iaspidus*.ADJ → *iaspidis*.NOUN 'jasper'.

The improvement is noticeable the other way around, too. Part-of-speech labels are amended for words which were assigned either correct (†*edax*.NOUN → *edax*.ADJ 'edacious') or incorrect lemmas (†*femineum*.NOUN → *femineus*.ADJ 'feminine') in the closed modality.

PROPN ↔ ADJ Additional training data in the *open* variant of our system improves considerably the distinction between homonymous `PROPN` and `ADJ` in all but the CROSS-TIME sub-tasks. The improvement concerns both frequent lexical units, such as *Romanus*.PROPN : *Romanus*.ADJ 'Roman', and less frequent words, such as *Phlaegreus*.PROPN → *Phlaegreus*.ADJ 'of Phlegra'. Likewise, ethnonyms are usually better distinguished from homonymous adjectives: *Persus*.ADJ 'Persian' → *Persae*.PROPN 'Persians' or *Campanus*.ADJ 'of Campania' → *Campani*.PROPN 'Campanians'.

VERB ↔ NOUN, ADJ The open variant of the system reduces considerably the number of incorrect idiosyncratic annotations, such as *supero*.VERB 'surmount' instead of *superi*.NOUN for *superi* 'the gods', †*uitro*.VERB instead of *uitrum* 'glass'.NOUN for *uitri*,

or †*sideo*.VERB instead of *siderita*.NOUN 'a kind of gemstone' for *sideritis*. It also leads to improved annotation of deverbal nouns, such as *sectura* 'a cut', *partus* 'a birth', which in the closed version were misclassified as VERB forms of *resp. seco* 'to cut' and *pario* 'to bring forth,'.

For Livy's and Ovid's works, the open variant performs better in labelling participles as VERBs rather than NOUNs. It also improves recognition of verb forms in the *Metamorphoses*: *sileo*.VERB 'to keep silence' for *sileam* or *auguror*.VERB 'to augur' for *auguror*. In the closed modality, these first-person forms, untypical of prosaic discourse, are misclassified as †*auguror*.NOUN and †*silea*.NOUN.

4.2. Lemmatization

It comes of no surprise that the open variant of our system improves lemmatization results, as both lemmatization and part-of-speech tagging are closely related tasks and depend one on another.

In the CLASSICAL sub-task, for example, a number of proper nouns unseen in the training dataset are correctly lemmatized, such as *Samnites*, *Samnium*, *Samnis*, *Priuernum*, *Latium*, *Antium*, *Antiati etc.* In the CROSS-GENRE sub-task, on the other hand, the open variant of the system assigns correct lemmas to words of Greek origin related to mythology (Ovid: *heros*, *nympha*, *thalamus*) and mineralogy (Pliny: *smaragdus*, *crystallus*, *sardonyx*), as well as to proper names (Ovid: *Alcmene*, *Iphis*, *Byblis*, *Dryope*).

Correct lemmas are also reached for a number of words which occur frequently in the test data, but (1) are rare or absent from the training dataset (Ovid: *lilium* or Pliny: *gutta*); (2) present phonetic assimilation unseen in the training dataset (*traluceo* : *transluceo*); or (3) have alternative spellings (*etiam nunc* : *etiamnunc*).

In the CROSS-TIME sub-task, the open variant of our system improves significantly the lemmatization of words which display post-classical or non-standard orthography that is not accounted for in the training dataset. Correct lemmas are assigned to word forms such as:

- qu-/c-: *quum* → *cum*
- -n-/-m-: *tanquam* → *tamquam*
- -ae-/-e-: *pene* → *paene*

Likewise, a number of proper nouns, both attested and not attested in Classical texts, are correctly lemmatized in the open modality (for instance *Laurentius*, *Lactantius*, *Strabo*, *Plato etc.*).

Despite using supplementary annotated data in the open modality, a number of lemmatization errors persist (4). They include among others:

- *sui* 'their *etc.* (*sc.* friends, followers)' is frequently misclassified as *suus*.DET;
- ethnonyms, which are either assigned lemmas in singular rather than plural (e.g. *uolscus* instead of *uolsci*) or are confused with adjectives (e.g. *carchedonii*.PROPN instead of *carchedonius*.ADJ);
- homonymous forms of low-frequency words, such as *pedum*.PROPN 'a town in Latium' (incorrectly lemmatized as *pes*.NOUN 'a foot') or almost full homonym pairs, such as *aer* 'the air' : *aes* '(any) base metal'.

Classical	Cross-genre		Cross-time
	Ovid	Pliny	
quis	quis	indicus	maior
sui	aer	indi	multus
priuernates	amans	quis	minus
pedum	refero	crystallus	fama
uolsci	quo	sarda	latinus
latini	carus	sestertius	melior
triarius	lotos	margarita	adsum
apuli	ora	uisus	maxime
philo	ausum	carchedonius	epistula
comitia	superus	quod	aliqui

Table 4: 10 most confused lemmas for each task

Some lemmatization choices may also be considered arbitrary and thus should not be expected to be correctly predicted by the tagger. This is the case, for instance, of *hyacinthos* instead of *hyacinthus* or *myrrha* instead of *murra*.

Finally, the last group of tagging errors results from the non-classical orthography employed in Sabellicus' work. However, poor results of the system in the closed modality might have been expected, since the training dataset does not account for spelling variation of Medieval or Neo-Latin texts:

- -o-/-u-: *epistola* → *epistula*
- -ph-/-f-: *phama* → *fama*
- -ci-/-ti-: *ocium* → *otium*
- -oe-/-e-: *foelix* → *felix*

5. Conclusions

The system presented in this paper outperforms competing architecture in lemmatization, part-of-speech and morphological tagging of Latin texts. It handles well the diachronic and diastratic variation of the language whose range of uses and coverage may be compared only to contemporary English. The open variant of the architecture improves significantly the results of both lemmatization and PoS tagging, leaving only small group of specific issues to persist in the resulting data.

Future work can focus on training language models on unlabeled Latin texts instead of using multilingual models, using context for lemmatization, and combining models into one for all tasks. The error analysis shows that careful selection of training data should help in addressing most if not all problems related to

spelling variation, unseen proper names and domain-specific terminology. The use of curated lexical resources should permit to reach preferred lemma labels for the convenience of the linguistic community. The system may be, then, hoped to perform well in a large-scale annotation of Medieval and Neo-Latin texts (Nowak, 2022).

6. Acknowledgments

This work was supported by the PLGrid Infrastructure and by the grant of the Polish Ministry of Science *eFontes. The Electronic Corpus of Polish Medieval Latin* (11H 17 0116 85).

7. Bibliographical References

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., 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, Online, July. Association for Computational Linguistics.

Denooz, J. (2007). Opera latina: le nouveau site internet du LASLA. *Journal of Latin Linguistics*, 9(3), jan.

Nowak, K., Bon, B., and Alexandre, R. (2016). Medialatinitas. Pour une intégration superficielle de ressources textuelles et lexicales en latin. In Damon Mayaffre, et al., editors, *JADT 2016. Journées Internationales d'Analyse Statistique Des Données Textuelles*, Nice, France. Presses de FacImprimeur.

Sprugnoli, R., Passarotti, M., Cecchini, F. M., Fantoli, M., and Moretti, G. (2022). Overview of the EvaLatin 2022 Evaluation Campaign. In Rachele Sprugnoli et al., editors, *Proceedings of the LT4HALA 2022 Workshop - 2nd Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (LREC 2022)*, Paris, France, June. European Language Resources Association (ELRA).

Wróbel, K. (2017). KRNNT : Polish recurrent neural network tagger. In Zygmunt Vetulani et al., editors, *Human language technologies as a challenge for computer science and linguistics : 8th language & technology conference : November 17-19, 2017, Poznań, Poland : proceedings*, pages 386–391. Fundacja Uniwersytetu im. Adama Mickiewicza, Poznań.

Wróbel, K. (2020). Kftt : Polish full neural morphosyntactic tagger. In Maciej Ogrodniczuk et al., editors, *Proceedings of the PolEval 2020 Workshop*, pages 47–53. Institute of Computer Sciences, Polish Academy of Sciences, Warszawa.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

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, March.

8. Language Resource References

Margherita Fantoli and Marco Carlo Passarotti and Eleonora Maria Litta and Paolo Ruffolo and Giovanni Moretti. (2022). *Linking LASLA corpus - LiLa LemmaBank*.

Nowak, Krzysztof. (2022). *eFontes. The Electronic Corpus of Polish Medieval Latin*.

Zeman, Daniel et al. (2022). *Universal Dependencies 2.10*.