Leveraging Part-of-Speech Tagging for Enhanced Stylometry of Latin Literature

Sarah Li Chen,[1,2,3] **Patrick J. Burns,**[4] **Thomas J. Bolt,**[5] **Pramit Chaudhuri,**[6]
Joseph P. Dexter[7,8†]
[1] Phillips Academy Andover
[2] Research Science Institute, Center for Excellence in Education
[3] Department of Computer Science, Stanford University
[4] Institute for the Study of the Ancient World, New York University
[5] Department of Languages & Literary Studies, Lafayette College
[6] Department of Classics, University of Texas at Austin
[7] Institute of Collaborative Innovation and Department of Computer and Information Science, University of Macau
[8] Data Science Initiative and Department of Human Evolutionary Biology, Harvard University
[†] Corresponding author: jdexter@um.edu.mo

Abstract

In literary critical applications, stylometry can benefit from hand-curated feature sets capturing various syntactic and rhetorical functions. For premodern languages, calculation of such features is hampered by a lack of computational resources for accurate part-of-speech tagging and semantic disambiguation. This paper reports an evaluation of POS taggers for Latin and their use in augmenting a hand-curated stylometric feature set. Our analyses show that POS-augmented features not only provide more accurate counts but also perform well on tasks such as genre classification. In the course of this work, we introduce POS n-grams as a feature for Latin stylometry.

1 Introduction

Although most associated with studies of authorship attribution and chronology (Stamatatos, 2009; Jockers and Witten, 2010; Stover and Kestemont, 2016), computational stylometric methods have increasingly been deployed to address broader literary questions and to augment more traditional approaches to criticism (Jockers, 2013; Moretti, 2013; Long and So, 2016; Piper, 2019; Underwood, 2019). For premodern literary traditions, such work has encompassed applications ranging from profiling the evolution of Latin prose style to computational restoration of Greek inscriptions and manuscripts (Dexter et al., 2017; Assael et al., 2022; Graziosi et al., 2023), as well as genre classification across multiple languages (Chaudhuri et al., 2019; Gianitsos et al., 2019; Storey and Mimno, 2020). For instance, for their genre classification work Chaudhuri et al. (2019) developed a set of 26 Latin stylometric features, including curated function word lists (e.g., prepositions, conjunctions, and pronouns), subordinate clauses, and sentence and clause length. Although calculated using only word or character n-gram counts and a small number of language-specific heuristics, these features proved highly effective for genre classification of classical texts, with the best-performing models achieving F1 > 97%.

It is the strength of a model, however, that it can withstand the frailties of individual features, at least up to a point. Hand-curated lists of words, such as those employed by Chaudhuri et al. (2019), may be insensitive to homonyms, semantic ambiguity, and other potentially challenging facets of natural language. While such issues may not impede success on certain tasks, increasing the accuracy of feature counts may be essential for others, especially those involving fine distinctions. Recent developments in NLP for Latin have led to the creation of tools that can plausibly improve on existing stylometric methods. Notably, the EvaLatin 2020 campaign (Sprugnoli et al., 2020) proposed shared tasks in lemmatization and part-of-speech (POS) tagging for classical Latin. Submissions introduced POS tagger models based on gradient boosters (Celano, 2020), ensemble methods (Stoeckel et al., 2020), and LSTMs (Straka and Straková, 2020) that achieved accuracies of up to 96%.

Here, we evaluate several POS taggers and assess how they improve and expand the feature set published by Chaudhuri et al. (2019). We perform error analysis on our POS-augmented features to quantify these improvements. We also train a classifier to distinguish Latin verse from

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 251–259

prose using either a POS-augmented or the original feature set, and we compare the accuracy and feature importances for each set. In doing so, we demonstrate the stylometric and literary relevance of POS-augmented features and showcase a transition from general tool development to specific literary applications in a lower-resource language.

2 Methods

2.1 POS taggers and test corpora

We evaluate 4 POS taggers to identify an optimal model for feature augmentation. Two models are pre-trained: a gradient boosting model developed as the Leipzig team's submission (Celano, 2020) for the EvaLatin 2020 task using LightGBM (Ke et al., 2017), and a FLAIR model developed by Stoeckel et al. (2020) for the EvaLatin 2020 task. We also consider Lapos (Tsuruoka et al., 2011) and MarMoT (Mueller et al., 2013), 2 well-established POS taggers that are not specific to Latin and were not pre-trained.

We test the models on the Perseus (Bamman and Crane, 2011), PROIEL (Haug and Jøhndal, 2008), and ITTB (Cecchini et al., 2018) Universal Dependencies (UD) Treebanks in addition to EvaLatin's (Sprugnoli et al., 2020) test corpora: a classical dataset consisting of texts from the same genre and time period as the training data, a cross-genre dataset consisting of Latin poetry rather than prose, and a cross-time dataset consisting of medieval rather than classical Latin. These datasets are annotated using the UD POS tag set (Petrov et al., 2012), and training and test sets are pre-split by EvaLatin or the respective UD treebank. We directly evaluate the 2 pre-trained POS taggers on the test data, and we train Lapos and MarMoT on the corresponding training data before evaluating them on each test set.

2.2 Augmenting existing stylometric features

We leverage predicted POS tags in 3 primary ways: to reduce the need for hand-engineered heuristics, to disambiguate polysemous function words, and to calculate additional features based on POS n-grams. Table 1 summarizes our modifications and additions to the published feature set (Chaudhuri et al., 2019).

2.2.1 Minimization of hand-engineered heuristics

Chaudhuri et al. (2019) compute the frequency of conjunctions and frequency of prepositions by identifying the tokens in a text that are in a hand-curated list of words. POS tagging eliminates the need for such lists by enabling direct counts of the corresponding POS tags. POS tagging also allows for frequency calculations with parts of speech that are too numerous to list exhaustively (e.g., all nouns or verbs).

In addition, Chaudhuri et al. (2019) identify superlatives by searching for the infix *-issim-*. We take a first step in improving that feature by only considering words tagged as ADJ or ADV and omitting non-adjective and non-adverb matches. Although an improvement, this count still does not encompass irregular Latin superlatives. We also supplement the hand-engineered feature calculating the frequency of vocatives with a new feature counting the frequency of contiguous blocks of words tagged as INTJ, reflecting the frequency of interjection and exclamation within a text. We exclude lone instances of 'O' to avoid redundancy with the vocative feature and to capture a more specific interjective subset.

2.2.2 Disambiguation of function words

Chaudhuri et al. (2019) rely on n-gram matching to identify keywords and compute corresponding features such as pronoun frequencies. For features that count largely monosemous words (e.g., *ipse*), this approach presents no problems. Some feature computations, however, involve words that can take on multiple meanings in different contexts. In these cases, blunt token matching cannot distinguish between a polysemous word's various usages. This ambiguity limits the value of counting 3 words in particular, *ut* (which can be an adverb or conjunction), *cum* ("when" or "with"), and *quod* ("because" or "which").

As noted above, the frequency of *ut* feature fails to distinguish between adverbial and conjunctive usages. Using POS tagging, we can inspect *ut* at a higher resolution and tabulate separate frequency features for its adverbial (ADV) and conjunctive (SCONJ) meanings. In addition, the feature calculating the frequency of *cum* clauses attempts to isolate conjunctive *cum* from prepositional *cum* by requiring that the word immediately following *cum* not have a standard ablative ending. This rule-based requirement is leaky and prone to false negative calls, in which instances of *cum* are unintentionally excluded from the count. Compared to a gold standard annotation of Livy 22.1-15, Chaudhuri et al. (2019) identify *cum* clauses with a pre-

cision of 1 but a recall of only 0.52. POS tags can directly distinguish between *cum* as "when" (SCONJ) or "with" (ADP) and remove this source of error.

Finally, the features concerning relative clauses (fraction of sentences containing a relative clause and mean length of relative clauses) rely on searching for inflected instances of *qui* (*qui, cuius, cui, quem, quo, quae, quam, qua, quod, quorum, quibus, quos, quarum*, or *quas*). This token matching incorrectly includes *quod* when used as a subordinating conjunction. POS tagging can again distinguish *quod*'s 2 meanings (PRON vs. SCONJ), reducing the error in relative clause features and also enabling the tabulation of a new feature, the frequency of *quod* as a subordinating conjunction.

2.2.3 Frequency of POS tag n-grams

POS tagging enables additional features based on the frequency of POS tag n-grams. These frequency features have been proposed and implemented in English stylometric work (Iyer and Ostendorf, 1999) but, to our knowledge, have never been applied to Latin. The number of possible n-grams, and therefore the number of frequency features, grows exponentially as n increases. We consider up to 2-grams in the current analysis.

2.3 Application to prose vs. verse classification

POS augmentation yields 3 distinct feature sets:

- **Original**: The original set of 26 features published by Chaudhuri et al. (2019).
- **Modified**: Feature set with POS-augmented preposition, conjunction, *ut*, *cum* clause, relative clause, and superlative features replacing the corresponding original features (see the direct modifications in Table 1).
- **Expanded**: All possible features, including the union of the original and modified feature sets and additional features enabled by POS tagging (see the additions in in Table 1).

We extract these 3 feature sets for a selection of 154 prose texts and 180 verse texts drawn from the Tesserae Project (Coffee et al., 2012) and train a random forest model to classify the texts by genre using each individual feature set.

3 Results

3.1 POS tagger evaluation and selection

We first consider the overall accuracy and F1 scores for the 4 taggers' POS tag predictions (Table 2). Among these, the LightGBM and FLAIR models are pre-trained on EvaLatin data, while we train MarMoT and Lapos on EvaLatin training data for the EvaLatin test sets and UD treebank training data for each treebank test set. This retraining accounts for MarMoT and Lapos' higher performance on the UD treebank test sets, compared to to the LightGBM and FLAIR models.

Inconsistencies between dataset annotations provide further explanation for the LightGBM and FLAIR models' worse performance on the UD treebanks. POS annotation guidelines vary between the EvaLatin data and the treebank data (as well as between different UD treebanks). For example, the Perseus Treebank does not use the UD DET tag, whereas EvaLatin does; this difference in annotation accounts for 32% of the FLAIR model's incorrect predictions (6% of its overall error on the Perseus test set). Therefore, treebank datasets impose inherent limits on the performance of the EvaLatin models.

Given these inconsistencies in annotation, we narrow our focus to the 3 EvaLatin test sets and more closely evaluate the 4 taggers trained on the EvaLatin training set: FLAIR, LightGBM, Lapos, and MarMoT. Out of these taggers, FLAIR exhibits the highest accuracies and F1 scores in the classical and cross-genre tasks but the poorest performance in the cross-time task (83% accuracy) (Figure 1). However, the accuracies of all the taggers are generally comparable and have a range of only 2% in the classical test data. We break down these seemingly similar performances by considering subclasses particularly relevant to feature augmentation: the tokens *cum*, *ut*, and *quod*. When considering tokens that fall into these subclasses of interest, the margin between FLAIR and the other taggers on the classical and cross-genre classes widens considerably. For instance, the gap between the F1 scores of the highest and lowest performing classifiers in the classical subtask increases from 0.04 overall to 0.21 in the *ut* class (Figure 1).

Furthermore, performance on these subclasses of interest demonstrates trends that contrast with overall performance. Although FLAIR has the worst overall performance on the cross-time task, it has the highest performance on *quod* and *cum* tokens

Original Feature	**Modification or Addition**
Frequency of prepositions	Count ADP tags (eliminate need for hand-curated list)
Frequency of conjunctions	Count SCONJ and CCONJ tags (eliminate need for hand-curated list)
Frequency of *ut*	Frequency of *ut* tagged as ADV
	Frequency of *ut* tagged as SCONJ
Frequency of *cum* clauses	Only consider *cum* tagged as SCONJ
Fraction of sentences containing relative clause	Only consider forms of *qui* tagged as PRON (exclude instances of *quod* used as SCONJ)
Mean length of relative clauses	
Frequency of superlative adjectives and adverbs	Only consider words tagged as ADJ or ADV
N/A	Frequency of *quod* used as a SCONJ
N/A	Frequency of contiguous instances of INTJ tags
N/A	Frequency of POS tag n-grams and n-skip-grams

Table 1: Table of selected original features from Chaudhuri et al. (2019) (left) and modifications or additions enabled by POS tagging (right). POS augmentation of the feature set includes direct modifications of existing features (indicated by a completed left and right column) as well as additions to the feature set (indicated by "N/A" in the left column).

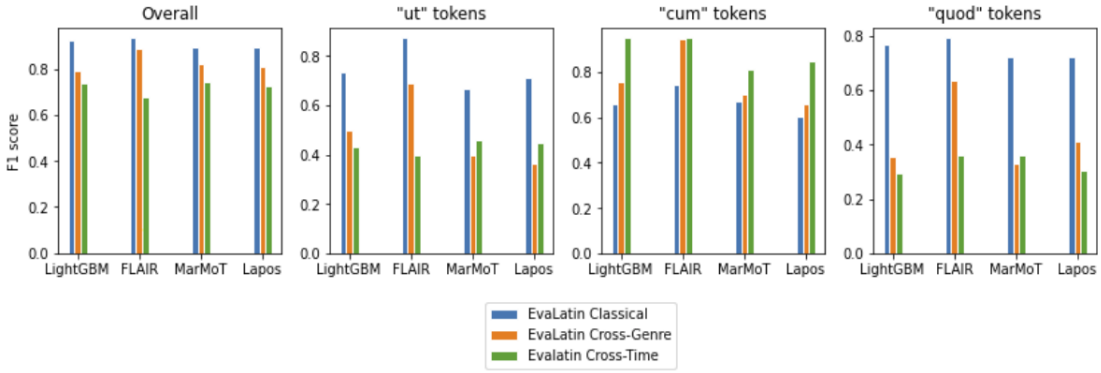

Figure 1: F1 score for LightGBM, FLAIR, MarMoT, and Lapos on EvaLatin test sets overall and on subsets most relevant to feature augmentation (*ut*, *cum*, *quod*).

	Accuracy				F1			
	LightGBM	FLAIR	MarMoT	Lapos	LightGBM	FLAIR	MarMoT	Lapos
EvaLatin Classical	0.96	0.97	0.95	0.95	0.92	0.93	0.89	0.89
EvaLatin Cross-Genre	0.89	0.91	0.87	0.87	0.79	0.89	0.82	0.81
EvaLatin Cross-Time	0.82	0.83	0.85	0.84	0.74	0.68	0.74	0.72
UD Perseus Treebank	0.77	0.80	0.84	0.84	0.50	0.53	0.72	0.73
UD PROIEL Treebank	0.79	0.82	0.95	0.95	0.69	0.73	0.95	0.94
UD ITTB Treebank	0.69	0.73	0.97	0.97	0.46	0.48	0.89	0.90

Table 2: POS tagger accuracy and F1 score across 3 EvaLatin test sets and 3 UD treebank test sets for 2 EvaLatin taggers (LightGBM and FLAIR) and 2 models that are not Latin-specific (MarMoT and Lapos) trained on EvaLatin or treebank data.

in that task. All taggers exhibit their highest performance on the classical subtask and poorer performance on the cross-time and cross-genre subtasks, but F1 scores for *cum* tokens for each tagger show the opposite trend, albeit with smaller margins. Given the pre-trained FLAIR model's strong performance overall as well as on *ut*, *cum*, and *quod* tokens, we use the model to augment the stylometric feature set. We apply the model to texts only within the classical and cross-genre domains, in which it demonstrates high performance.

3.2 Error analysis for POS-augmented features

Calculating stylometric features requires identifying tokens of interest and tabulating their frequency or some other summary metric. The method of token identification underlying a feature determines its accuracy. For example, it is necessary to identify identify conjunctive *cum* accurately to calculate the frequency of *cum* clauses. We perform an error analysis to compare the tokens identified by the original features and by POS-augmented features to the tokens marked by ground truth labels in the EvaLatin classical test dataset.

POS-augmented features overcome some limitations of the original methodology (see Table 3). When identifying conjunctions, counting words tagged as XCONJ (SCONJ or CCONJ) rather than using a hand-curated list increases F1 score from 0.69 to 0.97, an improvement of 0.28. When identifying prepositions, using the ADP tag decreases precision from 1 to 0.99 but increases recall by 0.67, from 0.33 to 1. The identification of *cum* clauses and relative clauses also improves when considering predicted POS tags. In this EvaLatin dataset, Chaudhuri et al. (2019)'s strict, rule-based method identifies *cum* with a precision of 0.92 but a recall of 0.55. Recall increases to 0.91 when counting instances of *cum* marked as SCONJ (Table 3). Chaudhuri et al. (2019)'s relatively loose criteria for identifying relative clauses (retrieve all instances of inflected *qui*) leads to a recall of 1 but a precision of only 0.59. Requiring instances of *qui* to be tagged as PRON increases the recall to 0.67 but still results in 353 false positives, suggesting that the method would benefit from further improvements (Table 3).

We also inspect token identification for features that lack definite ground truth labels in our dataset (Table 4). Requiring superlatives to be tagged as ADJ or ADV reduces the superlative count from 330 to 318. Manual inspection reveals that the 12 words omitted are forms of the verb *dissimulo* and are false positive hits. In addition, we count 6 vocatives and 13 INTJ blocks in the test data. There is no overlap between those sets. While the vocative feature identifies instances of direct address following 'O', the INTJ block feature identifies direct address without an 'O' marker and more general interjections such as *age* ("go"), *me hercule* ("by Hercules"), and *ecce* ("behold"). We thus improve the calculated frequency of superlatives feature and complement the calculated frequency of vocatives. Error analyses of remaining POS-augmented features, which include the frequency of conjunctive *quod*, conjunctive *ut*, adverbial *ut*, subordinating conjunctions, and pronouns, yield varying F1 scores with a minimum of 0.74 for conjunctive *quod* (Table 5).

3.3 POS-augmented features in prose vs. verse classification

We evaluate classifier performances with the original, modified, and expanded feature sets described above. There is no significant difference between the accuracy distributions for the different feature sets, although mean accuracy does increase to 98% for the expanded feature set (Table 6).

We also rank features in each set according to

	***Cum* Clauses**		**Relative Clauses**		**Conjunctions**		**Prepositions**	
	SCONJ	Original	PRON	Original	XCONJ	Original	ADP	Original
TP	217	132	725	729	5549	3743	3726	1227
FP	7	11	353	501	151	1425	36	0
FN	21	106	4	0	135	1941	16	2515
Precision	0.97	0.92	0.67	0.59	0.97	0.72	0.99	1.00
Recall	0.91	0.55	0.99	1.00	0.98	0.66	1.00	0.33
F1	0.94	0.69	0.80	0.74	0.97	0.69	0.99	0.49

Table 3: Use of POS tag information improves the identification of *cum* clauses, relative clauses (marked by forms of *qui*), conjunctions, and prepositions. TP denotes true positives, FP denotes false positives, and FN denotes false negatives. Relative clause identification requires punctuation information omitted by EvaLatin, so we evaluate relative clauses on the UD ITTB test data instead.

	Superlatives	**Superlatives (ADJ and ADV)**	**Vocatives**	**INTJ Blocks**
Instance count (predicted POS)	330	318	6	13
Instance count (true POS)	N/A	318	N/A	13

Table 4: Number of tokens counted by the original superlative feature, POS-augmented superlative feature, original vocative feature, and INTJ block feature enabled by POS information. Predicted POS tags match POS ground truth labels with 100% accuracy for all words relevant to the features shown, so the instance counts using predicted POS tags and ground truth POS labels are identical.

Gini importance (Table 7). The original and modified feature sets share 5 out of their 10 most highly ranked features (and 7 out of 10 when considering the POS-augmented versions of the superlatives and prepositions features). Furthermore, 4 of the top 6 features in the modified feature set are POS-augmented (frequencies of prepositions, conjunctions, and conjunctive *ut*). In addition, in the fully expanded set, the top 10 features include frequency of relative clauses (notably not the POS-augmented version), prepositions, *quidam*, and gerunds, all of which are also highly ranked in the original or modified set. However, POS n-gram features have the 2 highest Gini importances and represent 6 of the 10 most important features in the set, demonstrating their relevance to the differentiation of Latin genre.

3.4 POS-augmented features in differentiating epic vs. didactic

Despite the improvements enabled by POS-tagged features, the interpretive payoff can seem modest because of the relative simplicity of the evaluation task: even the original approach of using hard-coded lists achieves F1 > 97% in distinguishing prose and verse. We therefore apply our suite of feature sets to the subtler question of distinguishing works of Latin narrative epic and didactic poetry, which are composed in the same hexameter verse form. These genres differ in topical content and rhetorical structure: epic typically recounts stories of war, while didactic describes technical and scientific matters; epic alternates between narrative and speech, while didactic consists of philosophical argument and explanation. These characteristic qualities are not directly captured in the feature sets, which focus on functional and syntactic elements rather than literary ones. Prior research has demonstrated, however, that these genres can be distinguished on the basis of such features (Chaudhuri et al., 2019), and we find reasonably discrete groupings in our selective hexameter corpus; in particular, certain didactic authors are more clearly separated from their epic peers.

Fig. 2 shows that this central result replicates for POS-augmented features. The inclusion of POS n-gram features, however, reduces generic separation, with the notable exception of Lucretius' *De Rerum Natura*, which remains emphatically distinct. The differences in results across the 3 feature sets therefore illustrate the complex relationship between the 2 genres as a whole and the individual works comprising each genre – on the one hand, broadly similar in their sequences of parts of speech; on the other hand, crucially different in sentence length and sentence subordination, and above all different from one author to another.

	quod (SCONJ)	*ut* (SCONJ)	*ut* (ADV)	SCONJ	PRON
TP	125	365	112	1553	4172
FP	45	26	27	137	105
FN	43	27	26	130	136
Precision	0.74	0.93	0.81	0.92	0.98
Recall	0.74	0.93	0.81	0.92	0.97
F1	0.74	0.93	0.81	0.92	0.97

Table 5: Performance metrics for POS-augmented features not discussed in the main text. These features identify conjunctive *quod*, conjunctive *ut*, adverbial *ut*, subordinating conjunctions, and pronouns with F1 scores ranging from 0.74 to 0.97. TP denotes true positives, FP denotes false positives, and FN denotes false negatives.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
Original	0.96	0.99	0.96	0.99	0.97	0.97	0.015
Modified	0.99	0.97	1.00	0.97	0.95	0.98	0.017
Expanded	1.00	1.00	1.00	0.95	0.97	0.98	0.021

Table 6: 5-fold classifier accuracies for models using the original feature set, the directly modified feature set, and the fully expanded feature set in the prose vs. verse classification task.

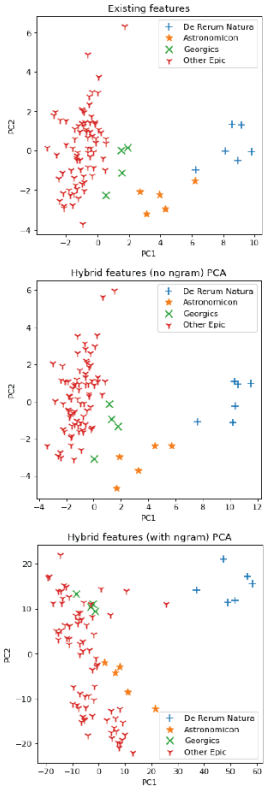

Figure 2: Principal component analyses of Latin narrative and didactic epic with the original (top), POS-augmented (middle), and hybrid stylometric and POS n-gram (bottom) feature sets.

4 Conclusion

We evaluate state-of-the-art POS taggers and select a FLAIR tagger to augment the stylometric feature set published by Chaudhuri et al. (2019). Using predicted POS tags, we first reduce dependency on hand-engineered heuristics in feature calculations to gain more complete POS counts, increasing recall by 0.32 for conjunctions and 0.67 for prepositions when comparing POS-augmented features to their original counterparts. Second, we disambiguate polysemous words such as *cum*, *ut*, and *quod*, increasing F1 score from 0.69 to 0.94 for *cum* clause identification and from 0.74 to 0.80 for relative clause identification. Finally, we calculate newly enabled features including POS n-gram frequencies.

We then train a random forest classifier to distinguish verse from prose, and through feature importance analysis we demonstrate that POS-augmented and POS n-gram features in particular quantify stylometric qualities highly relevant to genre classification. In these ways, we apply advances in Latin NLP to literary critical questions regarding generic style. More generally, we showcase a methodology for Latin that we hope will inform the quantitative criticism of other premodern languages as well.

5 Limitations

The current work uses established models for which performance on benchmark tasks has been well documented, such as the EvaLatin UDPipe

Rank	Original		Modified		Expanded	
1	superlatives	0.31	superlatives*	0.14	AUX*	0.13
2	*quidam*	0.14	*quidam*	0.13	SCONJ ADP 2-gram*	0.09
3	gerunds	0.13	prepositions*	0.10	relative clauses	0.07
4	relative clauses	0.09	conjunctions*	0.09	prepositions*	0.07
5	vocatives	0.08	gerunds	0.07	*quidam*	0.06
6	*idem*	0.07	*ut* (SCONJ)*	0.07	gerunds	0.05
7	personal pronouns	0.04	*antequam*	0.05	ADJ PROPN 2-gram*	0.04
8	*antequam*	0.03	alius	0.05	INTJ blocks*	0.04
9	prepositions	0.02	mean sentence length	0.05	PART ADP 2-gram*	0.04
10	*alius*	0.01	*idem*	0.05	ADP PRON 2-gram*	0.03

Table 7: For the original, modified, and expanded feature sets, the 10 features with highest Gini importance (feature name in left subcolumn, Gini importance in right subcolumn). Features improved or newly enabled by POS augmentation are denoted with *. Unless otherwise noted, each feature name in the table corresponds to the frequency of the indicated class.

model, which won all subtasks of the EvaLatin open division (Straka and Straková, 2020). The use of other models that reflect more recent advances is likely to have an effect on tagger accuracy and downstream performance for specific applications. Furthermore, models trained on a more diverse corpus may improve performance on cross-time tasks in particular. Finally, our use of POS n-grams as a stylometric feature is limited to 2-grams. Given their relatively high ranking among features contributing to successful classification, consideration of longer sequences, as well as of n-skip-grams, may be warranted.

6 Acknowledgments

This work was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org), an interdisciplinary group co-directed by P.C. and J.P.D. and supported by an American Council of Learned Societies Digital Extension Grant and a National Endowment for the Humanities Digital Humanities Advancement Grant (grant number HAA-271822-20). T.J.B. was supported by an Engaged Scholar Initiative Fellowship from the Mellon Foundation., P.C. by a Mellon New Directions Fellowship, and J.P.D. by a Neukom Fellowship and a Harvard Data Science Fellowship. S.L.C. conducted part of this research as a 2020 Research Science Institute Scholar. We are grateful to TTLab for sharing the FLAIR model before public release.

References

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer.

Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in converting the *Index Thomisticus* treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium. Association for Computational Linguistics.

Giuseppe G. A. Celano. 2020. A gradient boosting-Seq2Seq system for Latin POS tagging and lemmatization. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–123, Marseille, France. European Language Resources Association (ELRA).

Pramit Chaudhuri, Tathagata Dasgupta, Joseph P. Dexter, and Krithika Iyer. 2019. A small set of stylometric features differentiates Latin prose and verse. *Digital Scholarship in the Humanities*, 34(4):716–729.

Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson. 2012. The Tesserae Project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28(2):221–228.

Joseph P. Dexter, Theodore Katz, Nilesh Tripuraneni, Tathagata Dasgupta, Ajay Kannan, James A. Brofos, Jorge A. Bonilla Lopez, Lea A. Schroeder, Adriana Casarez, Maxim Rabinovich, Ayelet Haimson Lushkov, and Pramit Chaudhuri. 2017. Quantitative criticism of literary relationships. *Proceedings*

of the National Academy of Sciences, 114(16):E3195–E3204.

Efthimios Gianitsos, Thomas Bolt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. Stylometric classification of Ancient Greek literary texts by genre. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 52–60, Minneapolis, USA. Association for Computational Linguistics.

Barbara Graziosi, Johannes Haubold, Charlie Cowen-Breen, and Creston Brooks. 2023. Machine learning and the future of philology: A case study. *TAPA*, 153(1):253–284.

Dag T.T. Haug and Marius L. Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Rukmini Iyer and Mari Ostendorf. 1999. Relevance weighting for combining multi-domain data for n-gram language modeling. *Computer Speech & Language*, 13(3):267–282.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Matthew L. Jockers and Daniela M. Witten. 2010. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2):215–223.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hoyt Long and Richard Jean So. 2016. Literary pattern recognition: Modernism between close reading and machine learning. *Critical Inquiry*, 42(2):235–267.

Franco Moretti. 2013. *Distant Reading*. Verso Books.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Andrew Piper. 2019. *Enumerations: Data and Literary Study*. University of Chicago Press.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. 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).

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Manuel Stoeckel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. 2020. Voting for POS tagging of Latin texts: Using the flair of FLAIR to better ensemble classifiers by example of Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135, Marseille, France. European Language Resources Association (ELRA).

Grant Storey and David Mimno. 2020. Like two Pis in a pod: Author similarity across time in the Ancient Greek corpus. *Journal of Cultural Analytics*, 5(2).

Justin Stover and Mike Kestemont. 2016. Reassessing the Apulein corpus: A computational approach to authenticity. *The Classical Quarterly*, 66(2):645–672.

Milan Straka and Jana Straková. 2020. UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France. European Language Resources Association (ELRA).

Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Kazama. 2011. Learning with lookahead: Can history-based models rival globally optimized models? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 238–246, Portland, Oregon, USA. Association for Computational Linguistics.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.