A small set of stylometric features differentiates Latin prose and verse

Pramit Chaudhuri
Department of Classics, University of Texas at Austin, TX, USA

Tathagata Dasgupta[1] **and Joseph P. Dexter**
Department of Systems Biology, Harvard Medical School, MA, USA

Krithika Iyer[2]
Plano East Senior High School, TX, USA and Center for Excellence in Education, Research Science Institute, VA, US

Abstract

Identifying the stylistic signatures characteristic of different genres is of central importance to literary theory and criticism. In this article we report a large-scale computational analysis of Latin prose and verse using a combination of quantitative stylistics and supervised machine learning. We train a set of classifiers to differentiate prose and poetry with high accuracy (>97%) based on a set of twenty-six text-based, primarily syntactic features and rank the relative importance of these features to identify a low-dimensional set still sufficient to achieve excellent classifier performance. This analysis demonstrates that Latin prose and verse can be classified effectively using just three top features. From examination of the highly ranked features, we observe that measures of the hypotactic style favored in Latin prose (i.e. subordinating constructions in complex sentences, such as relative clauses) are especially useful for classification.

Correspondence:
Joseph P. Dexter, 200 Longwood Avenue, Boston, MA 02115, USA.
E-mail:
jdexter@fas.harvard.edu

Interrogator: *In the first line of your sonnet which reads 'Shall I compare thee to a summer's day,' would not 'a spring day' do as well or better?*

Witness: *It wouldn't scan.*
Interrogator: *How about 'a winter's day,' That would scan all right.*
Witness: *Yes, but nobody wants to be compared to a winter's day.*
Interrogator*: Would you say Mr. Pickwick reminded you of Christmas?*
Witness*: In a way.*
Interrogator*: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.*
Witness*: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.*

A. M. Turing, 'Computing Machinery and Intelligence' (1950)

1 Introduction

The differences between prose and verse are often both numerous and straightforward. Meter, rhyme, form, tone, appearance on the page—any one of these features can be a decisive, instantaneous indicator of a text's poetic quality. Nor is advanced

Digital Scholarship in the Humanities, Vol. 34, No. 4, 2019.
doi:10.1093/llc/fqy070 Advance Access published on 24 December 2018

training in the humanities or creative writing typically required to tell poetry from prose; almost everyone has an intuitive appreciation that (for instance) rap lyrics and a political speech are quite different, over and above any differences in content. It is less easy, however, to explain precisely how poetry differs from prose, especially when standard formal features such as meter or rhyme are set aside, as in much free verse, or when prose writing favors rhetorical techniques typically associated more closely with poetry. For many readers, the distinction will come down to 'I know it when I see it,' rather than any ironclad criteria. Indeed, the question of what makes poetry poetic is one almost as old as literary theory itself, preoccupying critics from Horace to the Russian Formalists and many others besides. As the epigraph above attests, however, the question is of broader interest than to scholars of literature alone. Appearing in a well-known paper by Alan Turing, the passage—an imaginary dialogue between a computer ('Witness') and human ('Interrogator')—suggests that an understanding of the nature and function of poetry is paradigmatic for any convincing claim to artificial intelligence. For a machine to qualify as genuinely intelligent, it needs to do more than merely understand metrical rules ('it wouldn't scan'); rather, the computer would require the suppleness to grasp and appropriately respond to emotions, literary references, and other elements of meaning.

Against this background, it may come as less of a surprise that the rarefied realm of poetry, with its manifold hermeneutical challenges, has furnished a variety of problems of interest to contemporary computer scientists. In particular, the classification of prose and verse using machine learning has attracted attention as an initial avenue for integrating literary study and sophisticated computation. Although the basic task of distinguishing poetry and prose would seem to have a greater affinity with the rudimentary beginning of Turing's dialogue than its more ambitious end, classification promises more than an early step along the path toward artificial intelligence. As this article shows, the use of machine learning can also tease out certain subtle characteristics underlying prose or verse and provide a quantitative profile of these different forms of expression.

A preliminary approach to prose/verse classification might focus on metrical features. The problem with such a line of attack, however, is the general absence of meter in prose texts, so that any classifier would trivially recapitulate manual annotations or the results of automated scansion programs.[3] More promising strategies can be divided into two broad categories: image-based and text-based. Image-based approaches rely on the typically distinct appearance of poetry on the page; although they have proven useful for certain tasks such as data mining and document organization (Hanauer, 1996; Lorang *et al.*, 2015), image-based classifiers do not invite follow-up investigation of more intricate literary questions or integration with traditional modes of criticism. In contrast, text-based approaches are useful not only for binary classification but also for analysis (and can even enable novel modes of composition). Classification of prose and verse written in English has been accomplished using a range of machine learning algorithms and linguistic features (Hanauer, 1996; Tizhoosh and Dara, 2006; Tizhoosh *et al.*, 2008; Kumar and Minz, 2014). Successful analyses have also been reported for various premodern and non-Western literary traditions, including ecclesiastical Latin and Malay poetry (Manjavaces *et al.*, 2017; Jamal *et al.*, 2012). The documentation for the technical software package *Mathematica* even includes a workflow (with Shakespeare as a case study) for prose/verse classification without extensive user intervention or judgement.[4] Particular attention has been devoted to poetic composition aided by machine learning on literary corpora, with notable recent examples including Swift-Speare (for generating pseudo-Shakespearean poetry) and DeepBeat (for rap lyrics) (Matias, 2010; Malmi *et al.*, 2016).

To the best of our knowledge, however, no similar analysis has been done for a classical literary tradition. In this article we report a large-scale characterization of Latin prose and verse using stylometric analysis and supervised machine learning. We train a set of classifiers that can differentiate prose and poetry very effectively (i.e. with accuracy values

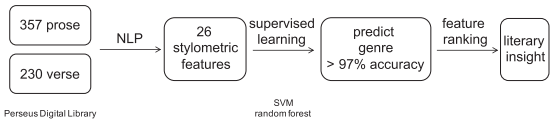

Fig. 1 Workflow for prose/verse classification. Twenty-six stylometric features (Table1) were calculated for 587 Latin text files (drawn from the Perseus Digital Library and further processed by the Tesserae Project) using custom heuristics. Two supervised learning algorithms (RF and linear SVM) were used to classify the text files as prose or verse, and statistical feature ranking was performed to gain insight into the stylometric features that best distinguish the genres.

of >97%) using a set of twenty-six text-based features (Fig. 1). We then rank the relative statistical importance of the features to identify a low-dimensional set still sufficient for classification and offer detailed literary interpretations of the possible significance of those features. In particular, we find that measures of hypotactic style favored in much Latin prose (i.e. the use of subordinating constructions in complex sentences, such as relative clauses) often rank highly.

Our research leverages the pioneering work of the Perseus Digital Library, further facilitated by the Tesserae Project, to digitize almost all extant Greek and Latin literary texts (Crane, 1996; Coffee *et al.*, 2012). We obtained a set of 587 digitized Latin text files, which we divided into prose (357 files, ca. 112 works) and verse (230 files, ca. 94 works) following standard generic conventions. Each file typically contains either a whole work (e.g. an oratorical speech, such as Cicero's *Pro Caelio*) or, in cases where a work is divided into individual books, one book (e.g. one of the three books of Caesar's *De Bello Gallico* or one of the twelve of Vergil's *Aeneid*). The set of texts (a full list of which is provided in the Appendix) includes almost all major surviving works of classical Latin literature, with the exception of substantially prosimetric works such as Boethius' *Consolatio Philosophiae* or Seneca's *Apocolocyntosis*, the latter of which is discussed below. (Petronius' *Satyricon* was labeled as prose given that its prose content significantly exceeds its verse content.) The chronological scope of the material is expansive, ranging from the comic plays of Plautus and fragments of Ennius' epic *Annales* (c. 200 BCE) to the poems of Ennodius (c. 500 CE) for verse, and from Cicero's early speeches (c. 80 BCE) to Jerome's letters (c. 400 CE) for prose. The generic scope is also broad. Represented in the corpus is epic (Vergil's *Aeneid*, Ovid's *Metamorphoses*) and didactic poetry (Lucretius' *De Rerum Natura*), tragedy (Seneca), comedy (Plautus, Terence), elegy (Propertius, Tibullus), historiography (Caesar, Livy, Tacitus, Suetonius), oratory (Cicero), philosophy (Cicero, Seneca), and technical writing (Vitruvius' *De Architectura*), to highlight only a handful of famous authors and works.

An important aspect of our approach is that it does not rely on syntactic parsing for feature extraction. Although development of a syntactic parser for Latin is an active area of research,[5] natural language processing (NLP) research for Latin and other classical languages lags well behind efforts for English. We therefore devised a set of twenty-six features that could be computed without recourse to general-purpose syntactic parsing, for instance by tabulation of a signal word (e.g. a pronoun or conjunction) or signal *n*-gram (e.g. morphological endings and infixes indicating a particular grammatical function, such as the *-issim-* element characteristic of regular superlative adjectives and adverbs). Where a syntactical marker might have a homonym, we devised heuristics to disambiguate between them as far as possible. Certain features were deliberately made especially capacious or selective in response to the challenges posed by Latin morphology and complex syntax. In related research, we used a similar feature set to analyze and identify citations of fragmentary early historians in Livy's monumental history of Rome, a problem of major interest in Latin historiography (Dexter *et al.*, 2017). In total, the feature set is intended to provide a thorough (though inevitably incomplete) picture of Latin literary style, including many items that are of standard philological interest (e.g. usage of relative and other subordinate clauses) or that have proven useful for computational analysis of genre in other languages (e.g. prepositions) (Adams *et al.*, 2005; Jockers, 2013), and builds on

Burrows' pioneering use of function words in literary stylometry (Burrows, 1987). Our approach was thus eclectic, derived from no single source and exploiting a range of features that have been applied to various languages and problems. In addition, we devised other features, including ones that were partial or noisy, to incorporate multiple types of evidence that could collectively capture diverse aspects of style. Table 1 lists the feature set divided into five broad grammatical categories: pronouns, non-content adjectives, conjunctions, subordinate clauses, and miscellaneous. We anticipate that our approach to feature extraction and literary machine learning will be applicable to other languages for which advanced NLP methods have not yet been developed.

A further goal of our research is to explore how stylometry and machine learning can support the practice of literary criticism. There is a long history of using stylometric analysis to address questions of authorship attribution and the dating of literary works in both classical and modern literary traditions (Mosteller and Wallace, 1964; Morton and Winspear, 1971; Marriott, 1979; Fitch, 1981; Holmes *et al.*, 2001; Vickers, 2004; Stamatatos, 2009; Forstall and Scheirer, 2010; Jockers and Witten, 2010; Stover *et al.*, 2016). Some recent work, however, has focused on the reapplication of stylometry, often involving machine learning methods, to address subtler literary critical questions. Notable examples include a statistical study of Shakespearean genre, integration of machine learning with traditional modes of criticism in the context of haiku, and an analysis of stylistic intertextuality in Latin tragedy and historiography (Hope and Witmore, 2010; Long and So, 2016; Dexter *et al.*, 2017), in addition to the development of pioneering frameworks of 'distant reading' (Moretti, 2013) and 'macroanalysis' (Jockers, 2013). Here we demonstrate that machine learning can decisively address a well-posed literary question, enabling us to identify large-scale stylistic signatures characteristic of Latin prose and verse genres. Moreover, we introduce methods for systematic ranking of feature importance, which are widely used in applications of machine learning outside of the digital humanities, to literary study (Chapelle and Vapnik, 1999; Guyon *et al.*, 2002; De la Torre and Vinyals, 2007; Grissa *et al.*, 2016).

Table 1 Full set of Latin stylometric features

	Pronouns
1	Frequency of personal pronouns
2	Frequency of demonstrative pronouns
3	Frequency of *quidam*
4	Frequency of third-person reflexive pronouns
5	Frequency of *iste*
6	Frequency of *ipse*
7	Frequency of *idem*
	Non-content adjectives
8	Frequency of *alius*
	Conjunctions
9	Aggregate frequency of conjunctions
10	Frequency of *atque* followed by consonant
	Subordinate clauses
11	Frequency of conditional clauses
12	Frequency of *cum* clauses
13	Frequency of *quin* clauses
14	Frequency of *quominus* clauses
15	Frequency of *antequam* clauses
16	Frequency of *priusquam* clauses
17	Frequency of *dum* clauses
18	Fraction of sentences containing relative clause
19	Mean length of relative clauses
	Miscellaneous
20	Frequency of interrogative sentences
21	Frequency of selected vocatives
22	Frequency of superlatives
23	Frequency of *ut*
24	Frequency of selected gerunds and gerundives
25	Mean sentence length
26	Aggregate frequency of prepositions

Note: The twenty-six features are divided into five broad grammatical categories (pronouns, non-content adjectives, conjunctions, subordinate clauses, and miscellaneous).

2 Methods

2.1 Texts

All analyses were performed on a set of 587 Latin text files, most of which were originally digitized by the Perseus Digital Library (Crane, 1996) and further processed by the Tesserae Project. The set includes 357 prose files (ca. 112 works) and 230 verse files (ca. 94 works) and is publicly available at https://github.com/tesserae/tesserae/tree/master/texts/la, with the exception of six text files for the poet Phaedrus, which

were obtained directly from Perseus. The vast majority of the works are classical Latin. The full list of texts is provided in an Appendix at the end of the article (unless otherwise noted, all features were calculated for each individual book).

2.2 Computation of stylometric features

All NLP tasks were performed using JavaScript (ES2015). We computed a set of twenty-six Latin stylometric features for use in the prose/verse classifiers. All features are continuous, were computed without use of syntactic parsing, and fall into five broad categories (Table 1). The features in the first two categories (pronouns and non-content adjectives) were calculated by counting instances of the various inflected forms of the indicated Latin word(s). Tables of the inflected forms can be found in any standard textbook or reference grammar for Latin.[6] A small number of feature calculations rely on modern editorial conventions, in particular punctuation, which has been exploited successfully in previous quantitative studies of classical literature (Clayman, 1981). In most cases the relevant punctuation is firm (e.g. a period or question mark) and is clearly implied by the syntax of the text, thereby reducing the likelihood of significant editorial differences, especially at scale.

Counts include either whole words or sequences of characters within words. For example, if counting instances of the polysemous word *ut*, which is both an adverb and a conjunction, we computed all appearances of the *n*-gram as a single word (e.g. ***ut*** *geniti*, ***ut*** *educati*, ***ut*** *cogniti essent*, not *Turnus rex R**ut**ulorum*.) If counting (for instance) standard superlative forms, however, we computed all appearances of the relevant *n*-gram as a part of a word (*opulent**issim**a*). Counts of whole words include both capitalized and lowercase forms, as well as instances with the enclitic *-que* (e.g. *utque* and *perque*), unless the enclitic produced another common Latin word (e.g. the number *quinque* instead of the conjunction *quin* with the enclitic *-que*) or the word was already included in the feature list (e.g. *atque* was not double-counted as *atque* and *at* + *que*). All frequencies are per character.

2.3 Conjunctions

- Conjunctions were computed by counting all instances of *ac, ast, at, atque, aut, autem, donec, dum, dummodo, enim, et, etenim, etiam, etiamtum, etiamtunc, nam, namque, nanque, neque, postquam, quamquam, quanquam, -que, quia, quocirca, sed, set, tamen, uel, uerumtamen, ueruntamen, utrumnam, vel, verumtamen*, and *veruntamen*.
- Frequency of *atque* followed by a consonant was computed by counting all instances of *atque* immediately followed by a word that begins with a consonant other than h (as h does not prevent elision).

2.4 Subordinate clauses

- Conditional clauses were computed by counting all instances of the conditional conjunctions *dummodo*, *nisi*, *quodsi*, *si*, and *sin*.
- *cum* clauses (where *cum* is an adverb or conjunction, but not a preposition) were computed by counting all instances of *cum* that are not immediately followed by a word ending in *-a*, *-e*, *-i*, *-o*, *-u*, *-is*, *-ibus*, *-ebus*, *-obus*, or *-ubus*. These restrictions were applied to exclude instances of *cum* as a preposition (which is followed by nouns in the ablative case, the inflected endings of which are listed above).
- *quin* clauses were computed by counting all instances of *quin*.
- *quominus* clauses were computed by counting all instances of *quominus* and *quo minus*.
- *antequam* clauses were computed by counting all instances of *antequam* and *ante quam*.
- *priusquam* clauses were computed by counting all instances of *priusquam* and *prius quam*.
- *dum* clauses were computed by counting all instances of *dum*.
- The fraction of non-interrogative sentences containing at least one relative clause was computed by identifying sentences that are both non-interrogative (i.e. ending with a punctuation mark other than '?') and have at least one form of the Latin relative pronoun (*qui*, *cuius*, *cui*, *quem*, *quo*, *quae*, *quam*, *qua*, *quod*, *quorum*,

quibus, *quos*, *quarum*, or *quas*). Interrogative sentences were excluded to obviate the need for semantic parsing of relative and interrogative pronouns, which are often morphologically identical.
- The average length of relative clauses was computed by counting the number of characters, excluding spaces and punctuation, in relative clauses, identified as beginning with a relative pronoun and ending at the next punctuation mark.

2.5 Miscellaneous

- (Direct) interrogative sentences were computed by counting all instances of a sentence ending in a question mark.
- Vocatives were computed by counting all instances of 'o' followed by a single word ending in *-a*, *-e*, *-i*, *-u*, *-ae*, *-es*, *-um*, or *-us*. The limitations were applied to exclude stand-alone instances of 'o' (an exclamation of surprise as well as part of a direct address), but to include instances of 'o' followed by a word with a standard vocative case ending.
- Regular superlative adjectives and adverbs were computed by counting all instances of *-issim-* within a word. The method excludes certain common superlatives such as *maximus* or *optimus*, which would be difficult to capture precisely without also incorporating proper names (e.g. Fabius Maximus or Jupiter Optimus Maximus).
- Frequency of *ut* (where *ut* is an adverb or a conjunction) was computed by counting all instances of *ut*.
- The limited subset of gerunds and gerundives was computed by counting all instances of *-ndus, -ndum*, *-ndarum*, and *-ndorum*. The restriction was designed to exclude the many verb forms that share the same letter sequence as the characteristic gerundival ending (e.g. *defendo* and *pendo*), though at the cost of also excluding the majority of the inflected forms of the gerund and gerundive. The common adverb *nondum* was excluded from this count. Erroneous inclusion of words such as the adjective *blandus* or noun *mundus* was assumed not to vitiate the count.
- The average length of sentences was computed by counting the number of characters, excluding spaces and punctuation, in sentences ending in a '.', '?', or '!'. We excluded any periods occurring after a single stand-alone character, since such instances are typically an abbreviation of a proper name rather than a sentence-end, and periods occurring after other common abbreviations (e.g. Aug., Cn., Kal., or common multi-letter Roman numerals followed by a period).
- Prepositions were computed by counting all instances of *ab*, *abs*, *absque*, *apud*, *cis*, *de*, *e*, *erga*, *ex*, *inter*, *ob*, *penes*, *per*, *praeter*, *pro*, *propter*, *sub*, *tenus*, and *trans*. Prepositions that may function as adverbs were excluded.

2.6 Error analysis

As described above, certain features could not be computed exactly and instead were estimated using various heuristics. To assess the effectiveness of these heuristics, we performed a manual error analysis of three features: frequency of regular superlatives, frequency of *cum* clauses, and frequency of the enclitic conjunction *-que*, which was a subset of our aggregated conjunctions feature. We analyzed the features in a small set of Latin prose and verse texts: Seneca's *Phoenissae* and the *Octavia* (verse) and Livy 22.1-15 (prose) for relative clauses, *Aeneid* 1 (verse) and Livy 22.1-15 for superlatives and *-que*, and Livy 22.1-15 for *cum* clauses. Table 2 lists the precision and recall of each heuristic in the texts analyzed.

2.7 Supervised machine learning

Prior to classification all features were rescaled to have minimum value 0 and maximum value 1. All supervised learning tasks were performed using Python 2.7. We used the scikit-learn implementations of the binary support vector machine (SVM) and random forest (RF) classifiers (Pedregosa *et al.*, 2011). For the linear SVM, we set $C = 0.5$, which is the default value in the scikit-learn package. For the RF classifier, feature ranking was according to Gini importance (Breiman and Cutler, 2008); ranking for

Table 2 Error analysis of selected features

Feature	Text	TP	FP	FN	Precision	Recall
Relatives	*Phoenissae* (V)	73	9	19	0.89	0.79
Relatives	*Octavia* (V)	103	7	7	0.94	0.94
Relatives	Livy 22.1-15 (P)	90	28	9	0.76	0.91
Superlatives	*Aeneid* 1 (V)	5	1	0	0.83	1
Superlatives	Livy 22.1-15 (P)	5	0	0	1	1
cum clauses	Livy 22.1-15 (P)	14	0	13	1	0.52
-que	*Aeneid* 1 (V)	280	19	0	0.94	1
-que	Livy 22.1-15 (P)	154	41	0	0.79	1

Notes: Table 2 summarizes the performance of four heuristics used for feature extraction on a sample of Latin texts. P and V indicate prose and verse, respectively, and TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. Data on relative pronouns are derived from the feature 'mean length of relative clause' (i.e. the error being analyzed is the incorrect identification of the relative pronoun using the heuristics underlying the feature). *-que* is not a stand-alone feature but is the only item in the aggregate frequency of conjunctions that could not be calculated exactly.

the linear SVM was determined by inspection of the set of weights of each support vector associated with the features' contributions (Chang and Lin, 2008). Unless otherwise noted, all results are based on five-fold cross-validation.

2.8 Code availability

All code is freely and publicly available at https://github.com/qcrit.

3 Results

We report a set of supervised learning classifiers that can distinguish Latin prose and verse with high accuracy and a literary critical examination of the stylometric features most useful for prose/verse classification. Table 1 lists the full set of twenty-six stylometric features, and an outline of the computational workflow is shown in Fig. 1.

3.1 Classification of prose and verse using the full set of stylometric features

We first attempted classification of Latin prose and poetry using all twenty-six features and two classification algorithms (RF and SVM with a linear kernel). Table 3 summarizes the results with five-fold cross-validation. With RF a total of fourteen texts were misclassified, and the mean accuracy across the folds was 97.6%; with the linear SVM, a total of thirteen texts were misclassified, and the mean accuracy across the folds was 97.8%. These results demonstrate that Latin prose and verse can be differentiated using stylometric features and prompted us to examine the relative importance of individual features, as discussed below. In addition, we confirmed that the RF model could classify texts artificially partitioned into 500-word chunks as prose or verse (mean cross-fold accuracy >90%), indicating that classifier performance is not strongly influenced by text length.[7]

3.2 Feature ranking identifies a small set of stylometric features sufficient for high-accuracy classification

We ranked the twenty-six features according to their contribution to classifier performance using both RF and linear SVM; Table 4 lists the top ten features by Gini importance (RF) or by the absolute value of the feature weight (linear SVM). Six of the top ten features are common between the models, as are three of the top four (fraction of sentences containing a relative clause, frequency of regular superlatives, and frequency of *quidam*), suggesting that our analysis identified a reproducible set of critical features. Furthermore, the three and five highest-ranked features alone are sufficient for prose/verse classification with >95% accuracy (Table 5). The literary and linguistic significance of the top features is reviewed in detail in the Discussion.

3.3 Classification of the Apocolocyntosis

The *Apocolocyntosis* ('Pumpkinification') is a satire on the deification of the emperor Claudius. Traditionally

Table 3 Performance of prose/verse classification using full feature set with five-fold cross-validation

	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Mean (%)	SD (%)	Accuracy (%)	*F*1 score (%)
RF	97.5	98.3	96.6	99.1	96.6	97.6	1.1	97.6	97.5
SVM	98.3	100	97.4	96.6	96.6	97.8	1.4	97.8	97.7

Note: Table 3 lists the accuracy for each fold, along with the mean and SD across folds, overall accuracy, and overall *F*1 score (macro-averaged).

Table 4 List of highly ranked features

Ranking	RF	Gini importance	SVM (linear kernel)	$\lvert w \rvert$
1	Frequency of superlatives	0.223	Frequency of superlatives	2.69
2	Fraction of sentences containing relative clause	0.184	Frequency of *antequam* clauses	1.65
3	Frequency of *quidam*	0.144	Fraction of sentences containing relative clause	1.56
4	Frequency of *idem*	0.0922	Frequency of *quidam*	1.53
5	Aggregate frequency of prepositions	0.0678	Frequency of *alius*	1.50
6	Frequency of selected gerunds and gerundives	0.0634	Frequency of *idem*	1.39
7	Frequency of *dum* clauses	0.0334	Frequency of personal pronouns	1.28
8	Frequency of selected vocatives	0.0331	Frequency of *iste*	1.22
9	Frequency of *ut*	0.0312	Frequency of *dum* clauses	1.07
10	Frequency of *cum* clauses	0.0257	Aggregate frequency of prepositions	1.06

Note: For RF the features are ranked by Gini importance; for the linear SVM, they are ranked by the absolute value of the weight of the support vector associated with the contribution of the feature.

Table 5 Accuracy of prose/verse classification using RF with reduced feature sets

	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Mean (%)	SD (%)	Accuracy (%)	*F*1 score (%)
All	97.5	98.3	96.6	99.1	96.6	97.6	1.1	97.6	97.5
Top 10	95.8	97.5	96.6	96.6	95.7	96.4	0.7	96.4	96.2
Top 5	98.3	97.5	95.7	97.4	94.0	96.6	1.7	96.8	96.6
Top 3	96.6	94.1	94.9	96.6	94.0	95.2	1.3	95.2	95.0

Note: For the three reduced feature sets, Table 5 lists the accuracy for each fold, along with the mean and SD across folds, overall accuracy, and overall *F*1 score (macro-averaged). The top features (by Gini importance with RF classification) are given in Table 4.

attributed to the statesman and philosopher Seneca, its date of composition is uncertain but likely to be soon after Claudius' death in 54 CE. The work is written in both prose and verse: the prose sections narrate the emperor's journey and encounters in the afterlife and are interspersed with passages of poetry in a high linguistic register. The combination makes the text an especially attractive candidate for testing automated differentiation of prose and verse. The various sections of the text (again drawn from the Perseus Digital Library) were aggregated into two bins, one of prose and the other of verse, for classification using the full set of twenty-six features. Both bins were correctly classified using an RF classifier trained on the full set of 587 texts.

4 Discussion

Our stylometric features collectively capture different aspects of the prosaic or poetic quality of a text, which accounts for their effectiveness as a combined set. Beyond this collective effectiveness, certain specific features stand out as having an intuitively greater suitability for one form of expression over the other. Two of the top five features by both rankings, for instance, point toward the relatively fuller and less restrictive scope of prose sentences in comparison to verse. Although not all Latin prose fits that characterization—prose writings, especially in the first-century AD, sometimes exploited the short, pointed sentence, and certain genres in any period,

Table 6 List of texts misclassified by both RF and linear SVM

Author	Text	Book
Lucretius	*De rerum natura*	1
Lucretius	*De rerum natura*	2
Lucretius	*De rerum natura*	3
Lucretius	*De rerum natura*	4
Lucretius	*De rerum natura*	5
Lucretius	*De rerum natura*	6
Manilius	*Astronomica*	1
Manilius	*Astronomica*	2
Sallust	*Historiae*	All
Tacitus	*Annales*	12

Note: Ten texts (eight verse and two prose) were misclassified by both models.

such as legal writings and literary commentary, could be fairly compact in expression (Kennedy, 1994; Adams *et al.*, 2005)—the general trend is nevertheless sufficient for texts to be classified effectively using only the highest-ranked features.

One of the most important features among the twenty-six for distinguishing between prose and verse is the fraction of non-interrogative sentences containing at least one relative clause (ranked second using RF, third using SVM). This finding reflects a common difference in the syntactical structures of the two forms. Prose tolerates longer sentences, which can be broken down into smaller clauses using a variety of subordinating constructions. The most common such construction is the relative clause, which hinges on the relative pronoun, 'who' or 'which' (in Latin the various inflected forms of *qui*). Relative clauses—like other subordinating constructions, such as conditional clauses, purpose clauses ('in order to'), and many further types—organize a thought into main and subsidiary elements, prior and latter actions, actual and contingent events, etc. With little restriction on sentence length, prose authors could regularly avail themselves of complex subordination to qualify a thought, or, more simply, to avoid the monotony of a series of parallel clauses. It is prose authors' customary favoring of this hypotactic, as opposed to paratactic, style that partly accounts for the prominence of the relative clause feature. This is not to say that verse is paratactic—far from it—but rather that the shorter sentences more often used by poets (mean length 102.3 ± 38.1 characters for the verse files, compared to 128.8 ± 37.1 for prose) reduced the need for extensive subordination, and hence relative clauses. Furthermore, one specific use of the relative pronoun may favor prose over verse: *qui* (or its inflections) often appears at the beginning of a sentence where it refers to an antecedent in the previous sentence, a usage known as the connecting relative. Although common to both prose and poetry, it is a frequent feature of Caesarian and other prose texts (Mayer, 2005; Spevak, 2010). Additional corroboration of the importance of hypotactic markers can be found among the top ten features: frequency of *ut* (ranked ninth using RF) and of *antequam* clauses (ranked second using SVM). In the former case, the word *ut* does have some non-subordinating uses, but a very large number of occurrences introduce one of a range of dependent clauses indicating purpose, result, or command, among other types. In the latter case, the adverb *antequam* ('sooner than', 'before') introduces certain temporal clauses. The feature rankings thus point to three potential markers of hypotaxis (*qui*, *ut*, and *antequam*) as playing an important role in the differentiation of prose and verse.

A different kind of expansiveness (or, conversely, compactness) is likely to account for the top ranking of superlative adjectives and adverbs, which again are enriched in prose texts. In this case, the relevant unit of analysis is not the sentence but rather the verse line. Latin poetry was metrical or quantitative and was structured according to patterns of syllable weight, primarily, and stress, secondarily. These patterns allowed for a certain number of syllables in the verse line, with twelve to seventeen syllables, for example, allowed in epic. At a mechanical level, then, verse composition partially consisted of the art of fitting choice words into interesting configurations while conforming to the metrical rules of the particular genre or subgenre (epic, tragedy, elegy, choral ode, etc.). The most common form of the superlative adjective in Latin adds the combined infix and ending *-issimus* (or its various inflections) to the stem of the base adjective, so that the resulting word is at a minimum four syllables long

(e.g. *fortissimus*, 'very brave') and in many cases five or more. Such a word can be accommodated into most verse meters, but it occupies a significant proportion of the line. Some words are especially long and would take up almost half of an entire verse line. The heptasyallabic word *tumultuosissimum* ('very tumultuous'), for instance, can be used by the historian Livy (2.10) without repercussion, but it presents a serious challenge to any poet, over and above the fact that its syllabic pattern prevents the word from being used at all in certain meters. In contrast to the subordinating features above, the frequency of superlative adjectives and adverbs is thus likely to be affected by meter rather than syntax. Both types of feature, however, are more characteristic of prose. Mostly unrestricted by metrical rules and often highly elastic and hypotactic in syntax, prose could accommodate several of the top-ranked features with somewhat greater ease than poetry.

Of the 587 texts in the corpus, a mere ten were misclassified (eight verse, two prose) by both models using the full feature set (Table 6). These shared misclassifications merit some explanation. The largest and most interesting category of misclassified texts contains didactic poetry, in particular the philosophical poems of Lucretius and Manilius. Both works are naturally influenced by philosophical prose, which plausibly explains features such as their longer-than-average sentence length compared to other hexameter works. The effects of that influence may partially account for the misclassifications. Consistent with that hypothesis, Vergil's *Georgics*, a didactic poem on an agricultural rather than conventionally philosophical theme, was correctly classified. In contrast, there is no clear reason for the misclassification of Sallust's *Historiae* and Book 12 of Tacitus' *Annales*; none of the other extant books of the *Annales*, and no other work by either author, was misclassified. In such mysterious cases it may be unprofitable to speculate on an explanation: the addition of a new feature or even the enhancement of an existing feature might be all that is required for a correct classification.

Insofar as the core objective of the experiment was to distinguish between Latin prose and verse, the achievement of >97% accuracy is a notable success. Since Latin poetry comprises a highly structured set of generic norms defined by meter, thematic content, and other features, it may seem to be a more straightforward candidate for classification than, say, much English poetry, especially those types using looser or non-existent meter and prosaic vocabulary. As discussed earlier, however, for all its structure Latin verse is also characterized by a remarkable diversity of metrical forms and subject matter, ranging from tightly constrained lyric to highly flexible drama. Moreover, poetry and prose often overlap in content: epic and historiography share long narratives about kings and battles, while letters are written in both prose and elegiac verse. Despite those challenges, our approach shows that Latin poetry—with extensive formal diversity and topical breadth—is nevertheless highly amenable to computational differentiation from prose. Also notable is the basis of this differentiation: mostly syntactic features, which were calculated without the aid of syntactic parsing. Attention to lexical markers as a proxy for syntactical constructions, coupled with heuristics to disambiguate between words of similar form, may offer a productive path forward for researchers working on literary traditions that lack the technical resources of English and other commercially important modern languages.

Furthermore, the introduction of feature ranking enabled us to identify the most salient features for the differentiation of prose and verse. Although routinely employed in other applications of machine learning, such as bioinformatics, it is an underutilized component of the toolkit of the digital humanist. Leveraging high-dimensional calculations well beyond the capacity of a human researcher, feature ranking can bring to light subtle yet important relationships between individual features and the data set as a whole. We hope that our application of feature ranking can provide a useful model for digital humanists seeking to extend the power of classification as a critical method.

Acknowledgements

The authors thank Jeffrey Flynt, Thomas Bolt, and Elizabeth Adams for assistance with development of

the Latin feature set. This work was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org), an interdisciplinary project co-directed by J.P.D. and P.C. and supported by seed funding from the Office of the Provost at Dartmouth College, a Neukom Institute for Computational Science CompX Faculty Grant, and a National Endowment for the Humanities Digital Humanities Start-Up Grant (grant number HD-248410-16). J.P.D. was supported by a National Science Foundation Graduate Research Fellowship (grant number DGE1144152) and a Neukom Fellowship, and P.C. was supported by an American Council of Learned Societies Digital Innovation Fellowship and a New Directions Fellowship from the Andrew W. Mellon Foundation. K.I. was a Research Science Institute Scholar during the summer of 2016.

Appendix

The full list of texts is as follows (unless otherwise noted, all features were calculated for each individual book):

Verse texts: Anonymous, *Laudes Domini*; Catullus, *Carmina* (divided into three files: epic, elegy, and miscellaneous); Claudian, *Carmina Minora*, *De Bello Gildonico*, *De Bello Gothico*, *De Consulatu Stilichonis*, *De Raptu Proserpinae*, *Epithalamium De Nuptiis Honorii Augusti*, *Panegyricus Dictus Probino et Olybrio Consulibus, In Eutropium*, *In Rufinum*, *Panegyricus De Tertio Consulatu Honorii Augusti*, *Panegyricus De Quarto Consulatu Honorii Augusti*, *Panegyricus De Sexto Consulatu Honorii Augusti*, and *Panegyricus Dictus Manlio Theodoro Consuli*; Dracontius, *Romulea* 10; Ennius, *Annales*; Ennodius, *Carmina* (Books 1 and 2 combined); Horace, *Ars Poetica*, *Carmen Saeculare*, *Epistles* (Books 1 and 2 combined), *Epodes*, *Odes*, and *Satires*; Italicus, *Ilias Latina*; Juvenal, *Satires*; Juvencus, *Historia Evangelica*; Lucan, *Bellum Civile*; Lucretius, *De Rerum Natura*; Manilius, *Astronomica*; Martial, *Epigrams*; Ovid, *Amores*, *Ars Amatoria*, *Epistulae Ex Ponto*, *Fasti*, *Heroides*, *Ibis*, *Medicamina Faciei Femineae*, *Metamorphoses*, *Remedia Amoris*, and *Tristia*; Persius, *Satires*; Phaedrus, *Fabulae*; Plautus, *Amphitruo*, *Asinaria*, *Aulularia*, *Bacchides*, *Captivi*, *Casina*, *Cistellaria*, *Curculio*, *Epidicus*, *Menaechmi*, *Mercator*, *Miles Gloriosus*, *Mostellaria*, *Persa*, *Poenulus*, *Pseudolus*, *Rudens*, *Stichus*, *Trinummus*, and *Truculentus*; Propertius, *Elegies* (Books 1–4 combined); Prudentius, *Apotheosis*, *Contra Symmachum*, *Dittochaeon*, *Epilogus*, *Hamartigenia*, and *Psychomachia*; Rutilius Namatianus, *De Reditu Suo*; Seneca, *Agamemnon*, *Hercules Furens*, *Hercules Oetaeus*, *Medea*, *Octavia*, *Oedipus*, *Phaedra*, *Phoenissae*, *Thyestes*, and *Troades*; Silius Italicus, *Punica*; Statius, *Achilleid* (Books 1 and 2 combined), *Silvae* and *Thebaid*; Terence, *Adelphi*, *Andria*, *Eunuchus*, *Heautontimorumenos*, *Hecyra*, and *Phormio*; Tibullus, *Elegies* (Books 1–3 combined); Valerius Flaccus, *Argonautica*; Vergil, *Aeneid*, *Eclogues*, and *Georgics*.

Prose texts: Ammianus, *Rerum Gestarum*; Apuleius, *Apologia*, *Florida*, and *Metamorphoses*; Augustine, *Epistulae* (Books 1–10, 11–20, 21–30, 31–40, 41–50, and 51–62); Julius Caesar, *De Bello Civili* and *De Bello Gallico*; Augustus Caesar, *Res Gestae Divi Augusti*; Celsus, *De Medicina*; Cicero, *Academica*, *Brutus*, *Cum Populo Gratias Egit*, *De Amicitia*, *De Divinatione*, *De Domo Sua*, *De Fato*, *De Finibus Bonorum et Malorum*, *De Haruspicum Responso*, *De Imperio Cn. Pompei*, *De Inventione*, *De Lege Agraria Contra Rullum*, *De Natura Deorum*, *De Officiis*, *De Optimo Genere Oratorum*, *De Oratore*, *De Partitione Oratoria*, *De Provinciis Consularibus*, *De Republica*, *De Senectute*, *Divinatio in C. Verrem* (Books 1, 2.1, 2.2, 2.3, 2.4, and 2.5), *Divinatio in Q. Caecilium*, *Epistulae ad Familiares*, *In Catilinam* (Books 1–4 combined), *In L. Pisonem*, *In Vatinium*, *Epistulae ad Atticum*, *Epistulae ad Brutum*, *Epistulae ad Quintum Fratrem*, *Lucullus*, *Orator*, *Paradoxa Stoicorum ad M. Brutum*, *Philippicae*, *Post Reditum in Senatu*, *Pro A. Caecina*, *Pro A. Cluentio*, *Pro Archia*, *Pro Balbo*, *Pro C. Rabirio*, *Pro C. Rabirio Postumo*, *Pro Fonteio*, *Pro L. Flacco*, *Pro Ligario*, *Pro M. Caelio*, *Pro Marcello*, *Pro Milone*, *Pro Murena*, *Pro Plancio*, *Pro Publio Quinctio*, *Pro Q. Roscio Comoedo*, *Pro Rege Deiotaro*, *Pro S. Roscio*, *Pro Scauro*, *Pro Sestio*, *Pro Sulla*, *Pro Tullio*, *Timaeus*, *Topica*, and *Tusculanae Disputationes*; Q. Cicero, *Commentariolum Petitionis*; Columella, *De Re*

Rustica (Books 1–9 only); Curtius Rufus, *Historiae Alexandri Magni*; Florus, *Epitomae De Tito Livio Bellorum Omnium Annorum DCC*; Gellius, *Noctes Atticae*; Jerome, *Epistulae* selections (Letters 1, 7, 14, 22, 38, 40, 43, 44, 45, 52, 54, 60, 77, 107, 117, 125, 127, and 128 combined); Livy, *Ab Urbe Condita* (Books 1–10, 21–30, 31–40, and 41–45); Marcus Minucius Felix, *Octavius*; Nepos, *Vitae*; Petronius, *Satyricon*; Pliny the Elder, *Naturalis Historia* (Books 1–5, 6–10, 11–15, 16–20, 21–25, 26–30, and 31–37); Pliny the Younger, *Epistulae*; Pseudo-Cicero, *In Sallustium*; Pseudo-Quintilian, *Declamationes Maiores*; Quintilian, *Institutio Oratoria*; Sallust, *Catilina*, *Historiae*, and *Jugurtha*; Scriptores Historiae Augustae, *Historia Augusta* (Books 1–5, 6–10, 11–15, and 16–21); Seneca the Younger, *Epistulae ad Lucilium* (Letters 1–10, 11–20, 21–30, 31–40, 41–50, 51–60, 61–70, 71–80, 81–90, 91–100, 101–110, 111–120, and 121–124), *De Beneficiis*, *De Brevitate Vitae*, *De Clementia*, *De Consolatione ad Helviam*, *De Consolatione ad Marciam*, *De Consolatione ad Polybium*, *De Constantia*, *De Ira*, *De Otio*, *De Providentia*, *De Tranquillitate Animi*, and *De Vita Beata*; Seneca the Elder, *Controversiae* (Books 1, 2, 7, 8, and 10), *Excerpta Controversiae*, *Fragmenta*, and *Suasoriae*; Suetonius, *De Vita Caesarum*; Tacitus, *Agricola*, *Annales*, *De Origine et Situ Germanorum*, *Dialogus de Oratoribus*, and *Historiae*; Tertullian, *Apologeticum* and *De Spectaculis*; Valerius Maximus, *Facta et Dicta Memorabilia*; and Vitruvius, *De Architectura*.

References

Adams, J. N., Lapidge, M., and Reinhardt, T. (2005). Introduction. In Reinhardt, T., Lapidge, M., and Adams, J. N. (eds), *Aspects of the Language of Latin Prose. Proceedings of the British Academy*, vol. **129**. Oxford: Oxford University Press, pp. 1–36.

Breiman L. and Cutler A. (2008). Random forests–classification manual. http://www.math.usu.edu/~adele/forests/ (accessed 30 August 2017).

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Chang, Y.-W. and Lin, C.-J. (2008). Feature ranking using linear SVM. In *JMLR: Workshop and Conference Proceedings*, vol. 3, pp. 53–64. http://proceedings.mlr.press/v3/chang08a/chang08a.pdf (accessed 30 August 2017).

Chapelle, O. and Vapnik, V. (1999). Model selection for support vector machines. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Denver, vol. 12, pp. 230–6.

Clayman, D. L. (1981). Sentence length in Greek hexameter poetry. In Grotjahn, R. (ed.), *Hexameter Studies. Quantitative Linguistics 11*. Bochum: Brockmeyer, pp. 107–36.

Coffee, N., Koenig, J.-P., Poornima, S., Ossewaarde, R., Forstall, C., and Jacobson, S. (2012). Intertextuality in the digital age. *Transactions of the American Philological Association*, **142**(2): 383–422.

Crane, G. (1996). Building a digital library: the Perseus Project as a case study in the humanities. In *Proceedings of the First ACM International Conference on Digital Libraries*, Bethesda, vol. 1, pp. 3–10.

De la Torre, F. and Vinyals, O. (2007). Learning kernel expansions for image classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4270176 (accessed 30 August 2017).

Dexter, J. P., Katz, T., Tripuraneni, N., Dasgupta, T., Kannan, A., Brofos, J. A., Bonilla Lopez, J. A., Schroeder, L. A., Casarez, A., Rabinovich, M., Haimson Lushkov, A., and Chaudhuri, P. (2017). Quantitative criticism of literary relationships. *Proceedings of the National Academy of Sciences United States of America*, **114**(16): E3195–204.

Fitch, J. (1981). Sense-pauses and relative dating in Seneca, Sophocles and Shakespeare. *American Journal of Philology*, **102**(3): 289–307.

Forstall, C. W. and Scheirer, W. J. (2010). Features from frequency: authorship and stylistic analysis using repetitive sound. In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, 1.2. Chicago, IL: University of Chicago.

Grissa, D., Pétéra, M., Brandolini, M., Amedeo, N., Comte, B., and Pujos-Guillot, E. (2016). Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. *Frontiers in Molecular Biosciences*, **3**: 30.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1–3): 389–422.

Hanauer, D. (1996). Integration of phonetic and graphic features in poetic text categorization judgements. *Poetics*, **23**(5): 363–80.

Holmes D. I., Robertson, M., and Paez, R. (2001). Stephen Crane and the *New-York Tribune*: a case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, **35**(3): 315–31.

Hope, J. and Witmore, M. (2010). The hundredth psalm to the tune of 'Green Sleeves': digital approaches to the language of genre. *Shakespeare Quarterly*, **61**(3): 357–90.

Jamal, N., Mohd, M., and Noah, S. A. (2012). Poetry classification using support vector machines. *Journal of Computer Science*, **8**(9): 1441–6.

Jockers, M. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–23.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Kennedy, G. A. (1994). *A New History of Classical Rhetoric*. Princeton, NJ: Princeton University Press.

Kumar, V. and Minz, S. (2014). Poem classification using machine learning approach. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012)*, December 28-30, 2012, Jaipur, India, pp. 675–82. https://link.springer.com/chapter/10.1007/978-81-322-1602-5_72 (accessed 30 August 2017).

Long, H. and So, R. J. (2016). Literary pattern recognition: modernism between close reading and machine learning. *Critical Inquiry*, **42**(2): 235–67.

Lorang, E., Soh, L. K., Datla, M. V., and Kulwicki, S. (2015). Developing an image-based classifier for detecting poetic content in historic newspaper collections. *D-Lib Magazine*, **21**: 7–8. http://www.dlib.org/dlib/july15/lorang/07lorang.html (accessed 30 August 2017).

Malmi, E., Takala, P., Toivonen, H., Raiko, T., and Gionis, A. (2016). DopeLearning: a computational approach to rap lyrics generation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco*, pp. 195–203. http://dl.acm.org/citation.cfm?doid=2939672.2939679 (accessed 30 August 2017).

Manjavaces, E., De Gussem, J., Daelemans, W., and Kestemont, M. (2017). Assessing the stylistic properties of neurally generated text in authorship attribution. In *Proceedings of the Workshop on Stylistic Variation*, Copenhagan, pp. 116–125. http://aclweb.org/anthology/W17-4914 (accessed 4 June 2018).

Marriott, I. (1979). The authorship of the *Historia Augusta*: two computer studies. *Journal of Roman Studies*, **69**: 65–77.

Matias, J. N. (2010). *Swift-Speare: Statistical Poetry*. http://natematias.com/portfolio/DesignArt/Swift-SpeareStatisticalP.html (accessed 30 August 2017).

Mayer, R. (2005). The impracticability of Latin Kunstprosa. In Reinhardt, T., Lapidge, M., and Adams, J. N. (eds), *Aspects of the Language of Latin Prose. Proceedings of the British Academy*, vol. 129. Oxford: Oxford University Press, pp. 195–210.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Morton, A. Q. and Winspear, A. D. (1971). *It's Greek to the Computer*. Montreal: Harvest House.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, **12**: 2825–30.

Spevak, O. (2010). *Constituent Order in Classical Latin Prose*. Amsterdam: John Benjamins Publishing Company.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.

Stover, J., Winter, Y., Koppel, M., and Kestemont, M. (2016). Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology*, **67**(1): 239–42.

Tizhoosh, H. R. and Dara, R. A. (2006). On poem recognition. *Pattern Analysis and Applications*, **9**(4): 325–38.

Tizhoosh, H. R., Sahba, F., and Dara, R. (2008). Poetic features for poem recognition: a comparative

study. *Journal of Pattern Recognition Research*, **3**(1): 24–39.

Vickers, B. (2004). *Shakespeare, Co-author: A Historical Study of Five Collaborative Plays*. Oxford: Oxford University Press.

Notes

1 Present address: 4D Path, Inc., Newton, MA, USA.

2 The authors are listed alphabetically, and the order of the author list does not reflect relative contributions to the work reported.

3 For instance, *Pede certo* is a tool for computational scansion of Latin dactylic hexameter (the meter of epic poetry) and elegiacs developed by the Università di Udine and the *Musisque Deoque* digital archive (http://www.pedecerto.eu/).

4 See https://www.wolfram.com/mathematica/new-in-10/highly-automated-machine-learning/determine-if-a-text-is-prose-or-poetry.html.

5 See, for instance, the ongoing development of the Classical Language Toolkit (CLTK), an extension of the Python NLTK library to Latin and Greek and, in due course, other ancient languages (www.cltk.org).

6 For example, *Allen and Greenough's New Latin Grammar*, which is available electronically through the Perseus Project (http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0001&redirect=true). For the sake of completeness, we also included several archaic forms and alternative spellings of inflected forms.

7 We selected a diverse subset of prose and verse texts from the full corpus (Horace's *Odes* 1, Ovid's *Metamorphoses* 1, Plautus' *Amphitruo*, Vergil's *Aeneid* 1, Caesar's *De Bello Gallico* 1, Cicero's *Pro Archia*, and Vitruvius' *De Architectura*) and partitioned them into chunks of 500 words each, with any remaining material set aside. We then classified the chunks by genre using the full feature set and the same workflow as for the main experiments. The mean cross-fold accuracy was >90% as were the overall accuracy and *F*1 score.