A Stylometry Toolkit for Latin Literature

Thomas J. Bolt,[1] **Jeffrey H. Flynt,**[2] **Pramit Chaudhuri,**[1] **Joseph P. Dexter**[3†]
[1] Department of Classics, University of Texas at Austin
[2] Department of Molecular Biosciences, University of Texas at Austin
[3] Neukom Institute for Computational Science, Dartmouth College
[†] Corresponding author: joseph.p.dexter@dartmouth.edu

Abstract

Computational stylometry has become an increasingly important aspect of literary criticism, but many humanists lack the technical expertise or language-specific NLP resources required to exploit computational methods. We demonstrate a stylometry toolkit for analysis of Latin literary texts, which is freely available at www.qcrit.org/stylometry. Our toolkit generates data for a diverse range of literary features and has an intuitive point-and-click interface. The features included have proven effective for multiple literary studies and are calculated using custom heuristics without the need for syntactic parsing. As such, the toolkit models one approach to the user-friendly generation of stylometric data, which could be extended to other premodern and non-English languages underserved by standard NLP resources.

1 Introduction

Stylometry, the quantitative analysis of writing style, is an longstanding yet active area of research in literary studies. Traditional applications of stylometry in both classical and modern literary scholarship have focused on authorship attribution and establishing relative chronology (Mosteller and Wallace, 1964; Marriott, 1979; Fitch, 1981; Vickers, 2004; Jockers and Witten, 2010; Stover et al., 2016). In recent years, new digital tools and computational methods, especially machine learning (Long and So, 2016; Dexter et al., 2017), have allowed researchers to address more fine-grained literary critical questions and have also given rise to novel frameworks for literary analysis, such as 'distant reading' and 'macroanalysis' (Moretti, 2013; Jockers, 2013; Piper, 2018; Underwood, 2019).

Much research in computational stylometry has focused on English literature due in part to the rich NLP resources available for the English language, especially high-quality syntactic parsing. NLP resources for many premodern and non-English languages are, by contrast, at an earlier stage of development or entirely lacking. Moreover, many of the academic disciplines studying these languages are smaller than for English, and thus the community of potential developers is correspondingly reduced. These factors suggest the need for user-friendly stylometric tools, which can provide a wide range of literary data for under-resourced languages and are suitable for use by humanists lacking a computational background.

Syntactic parsing, which remains at an early stage of development for Latin,[1] is not a prerequisite for the successful application of computational stylometry to literary problems. Our prior work has shown that custom heuristics can enable extraction of a wide range of features useful for the study of Latin literature, in particular syntactic markers, non-content words, and elements of sound and rhythm (Dexter et al., 2017; Chaudhuri et al., 2018). Here we report development of a point-and-click stylometry toolkit to enable easy generation of such data for a corpus containing almost all major classical Latin texts.

Other recently developed stylometry packages, such as the "stylo" R package and Lexomics, are aimed at audiences with a range of computational expertise (Eder et al., 2016; Drout et al., 2007). These packages, however, have typically been developed for general-purpose application to multiple languages instead of a single language. Focusing on the latter creates opportunities for targeting language-specific features, which often play a crucial role in literary style.

[1] See, for example, the recent progress of the Classical Language Toolkit (CLTK) (www.cltk.org) and StanfordNLP (https://stanfordnlp.github.io/stanfordnlp/index.html).

Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations), pages 205–210
Hong Kong, China, November 3 – 7, 2019.

The need for a point-and-click toolkit is particularly acute in classical studies. Although classical philologists have long applied stylometry to shed light on questions of authorship, relatively few studies have employed digital tools. Exceptions have tended to focus on a restricted set of features, such as relative word frequency (Stover et al., 2016) or average sentence length (Marriott, 1979; Clayman, 1981). Such limitations may be due in part to the absence of an accurate method for syntactic parsing, and in part to a more general lack of collaboration to date between classical philologists and NLP specialists. By improving the accessibility of rich philological data, our toolkit should further promote the adoption of quantitative approaches by literary critics. At the same time, the toolkit bridges the gap between classical studies and research on English, in which computational approaches are more common and are supported by a more extensive technical apparatus.

2 Toolkit

Our toolkit provides researchers working with Latin literature access to large-scale stylometric data difficult to acquire by non-computational methods and enables humanists without specialist digital training to construct custom datasets.

The design goal for the toolkit is to provide an intuitive and easy-to-use interface hosted in a web browser. The interface is point-and-click and can be used by researchers with no prior programming or NLP experience. Users can choose from over 700 Latin texts, which comprise almost all of the surviving corpus of classical Latin. The texts were originally digitized by the Perseus Digital Library and further developed by the Tesserae Project (Crane, 1996; Coffee et al., 2012). Texts can be selected by author, text, or book (roughly the ancient equivalent of a chapter). Searches can be as fine-grained as examining a single book, or as large-scale as analyzing the entire built-in corpus in one go (Figure 1).

Next, users select the stylometric features to analyze for their chosen corpus. They can run analyses using any combination of the twenty-six features (Figure 2 shows a sample output). The results are displayed on a spreadsheet in the web browser and can be downloaded as a CSV file. In addition, a user can produce simple visualizations (e.g., a bar chart comparing the values of a particular feature across a set of texts) inside the toolkit.

The ease with which the toolkit can be used does not limit its versatility. A user can create a custom corpus of texts preselected from the existing database, which is close to comprehensive for canonical material, or upload texts of their own for analysis. This latter functionality is especially important for understudied texts, such as those produced in Late Antiquity and during the Renaissance, the sum total of which far exceeds the quantity of extant classical Latin. While digital versions are available for many post-classical texts, for the most part the later periods are not well served by the prominent tools or repositories in the field, which maintain a classical focus. Our toolkit allows users to analyze any text available in electronic form. Furthermore, if a work is not available online, a user may upload a plain text file or transcribe it directly into the upload interface.

3 Features

Our feature set comprises twenty-six stylometric features across four broad syntactic and grammatical categories (pronouns and non-content adjectives, subordinate clauses, conjunctions, and miscellaneous, as listed in Table 1) and is described in detail in a previous publication (Chaudhuri et al., 2018). Some features are lexical (e.g., prepositions), while others are syntactic (e.g., sentence length) or address semantic and rhetorical aspects of the texts (e.g., superlatives and interrogative sentences). Taken together, the features offer a rich and diverse, albeit necessarily partial, profile of Latin literary style.

An important aspect of our toolkit is that it does not depend on syntactic parsing, named entity recognition, or other NLP methods that have not been developed fully for classical Latin (Erdmann et al., 2016). We employ three strategies to circumvent current technical limitations. The majority of features (*Alius*, *Idem*, *Ipse*, *Iste*, *Quidam*, Demonstrative Pronouns, Personal Pronouns, Third-Person Pronouns, *Atque* + Consonant, *Antequam*, *Cum*, *Dum*, *Priusquam*, *Quin*, *Quominus*, *Ut*, and Prepositions) are computed using hard-coded lists of almost all possible forms of the relevant Latin words. While some features (e.g., *Quin* or *Dum*) are frequencies of a single form, others (e.g., Demonstrative Pronouns) involve long lists of morphological variants. Other features are estimated based on the frequency of a

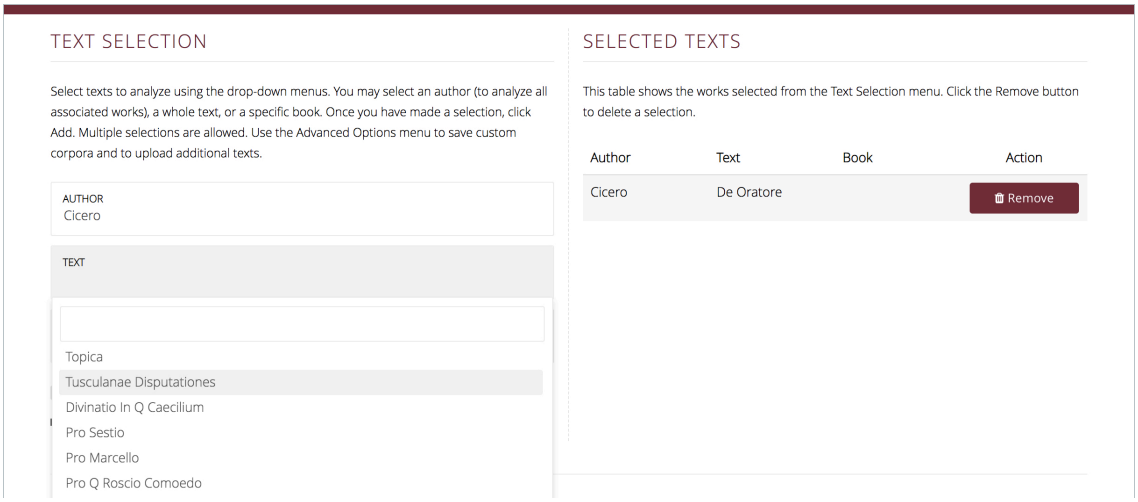

Figure 1: On the left, drop-down menu for text selection; on the right, a list of texts that have been selected.

Stylometry

New Search | Export Results | View Graphs | Compare Features

	Text	Words	Relative Clauses	Mean Length Relative Clauses	Interrogative Sentences	Vocatives	S
403	Lucan Bellum Civile Part 9	7,077	0.29915	27.04348	0.00136	0.00017	0
404	Lucan Bellum Civile Part 10	3,508	0.27322	27.77778	0.00045	0.00005	0
405	Lucretius De Rerum Natura Part 1	7,223	0.66812	38.8	0.00055	0	0
406	Lucretius De Rerum Natura Part 2	7,562	0.6087	37.29661	0.00054	0.00002	0
407	Lucretius De Rerum Natura Part 3	7,398	0.54018	39.57635	0.00071	0.00005	0
408	Lucretius De Rerum Natura Part 4	8,622	0.56508	36.77955	0.00046	0	0
409	Lucretius De Rerum Natura Part 5	9,508	0.51299	42.15234	0.00082	0.00002	0
410	Lucretius De Rerum Natura Part 6	8,517	0.5	36.67045	0.00053	0	0
411	Manilius Astronomicon Part 1	5,880	0.44076	42.97101	0.00032	0	0
412	Manilius Astronomicon Part 2	6,244	0.44578	34.25	0.00025	0	0
413	Manilius Astronomicon Part 3	4,465	0.43646	36.16807	0.00016	0.00004	0
414	Manilius Astronomicon Part 4	6,059	0.30435	39.15686	0.00072	0	0

Figure 2: Sample output from the toolkit for a selection of Latin literary texts.

signal n-gram. For instance, all regular superlative adjectives include the n-gram *-issim-* (e.g., *largissimus*, "most abundant" or *clarissima*, "clearest"). As this n-gram is extremely rare outside of superlatives, we could curate a near-comprehensive list of exclusions (e.g., *dissimilis*, "unlike"). We use a similar strategy to capture the instances of selected gerunds and gerundives, which contain the n-grams *-ndus*, *-ndum*, *-ndarum*, or *-ndorum*. A third class of features are determined using punctuation (e.g., question marks to assess the frequency of direct interrogative sentences or to filter interrogative pronouns, which have many forms in common with relative pronouns, from relative clause counts).

The precision and recall of each of these heuristics is discussed in detail in (Chaudhuri et al., 2018). We emphasize that these approaches are not intended as a substitute for NLP, but rather as a stopgap for philologists until more substantial resources become available for classical languages. We expect that the overall usefulness of the toolkit will increase as our heuristics are rendered obsolete by improvements in part-of-speech tagging and dependency parsing for Latin.

Our features are drawn from a wide array of sources in order to maximize the capture of information pertinent to Latin literary style. Some features, such as prepositions, are inspired by studies of other languages, where they have proven useful for the characterization of genres or subgenres (Jockers, 2013). Most features, however, are based on previous studies of Latin style and are designed to capture aspects specific to the Latin language (Adams, 1972; Adams et al., 2005). For example, *atque* ("and") followed by a word beginning with a consonant is a stylistic feature that is associated with certain influential figures writing early in the tradition. When later authors employ *atque* + consonant, they do so either in imitation of these figures specifically, or to recall an archaizing style more generally.

	Feature
	Pronouns and non-content adjectives
1	*Alius*
2	*Idem*
3	*Ipse*
4	*Iste*
5	*Quidam*
6	Demonstrative Pronouns
7	Personal Pronouns
8	Third-Person Pronouns
	Conjunctions
9	*Atque* + Consonant
10	Conjunctions
	Subordinate clauses
11	*Antequam*
12	*Cum*
13	*Dum*
14	*Priusquam*
15	*Quin*
16	*Quominus*
17	Conditional Markers
18	Fraction of Sentences with Relative Clauses
19	Mean Length of Relative Clauses
	Miscellaneous
20	*Ut*
21	Interrogative Sentences
22	Mean Length of Sentences
23	Prepositions
24	Regular Superlatives
25	Selected Gerunds & Gerundives
26	Selected Vocatives

Table 1: Full set of Latin stylometric features.

4 Literary Importance

The stylometric data generated by the toolkit sheds light on a variety of literary problems. The simplest type of analysis involves a single feature calculated across a small number of texts. Past research in Ancient Greek stylometry, for instance, has shown that sentence length constitutes one meaningful difference between the early Homeric hexameter tradition and the Hellenistic tradition, since later writers use longer sentences even as they retain other core aspects such as formulaic language and meter (Clayman, 1981). Figure 3 shows the mean sentence length of most of the surviving classical Latin epics as calculated by the toolkit. Three texts, *De Rerum Natura* by Lucretius, *Astronomicon* by Manilius, and the *Georgics* by Vergil, have noticeably longer sentences on average (mean length >140 characters, compared to <125 characters for the other epics). An attractive explanation for the three anomalous texts is that they are all identified with a subgenre of epic known as "didactic," a specific class which purports to teach its readers philosophy or a specialized technical skill, such as astrology or farming. The sentences are longer plausibly because detailed treatment of intricate philosophical or technical issues requires more complex sentences than typically more straightforward narrative action or direct speech, which represent the principal content of the other epics.

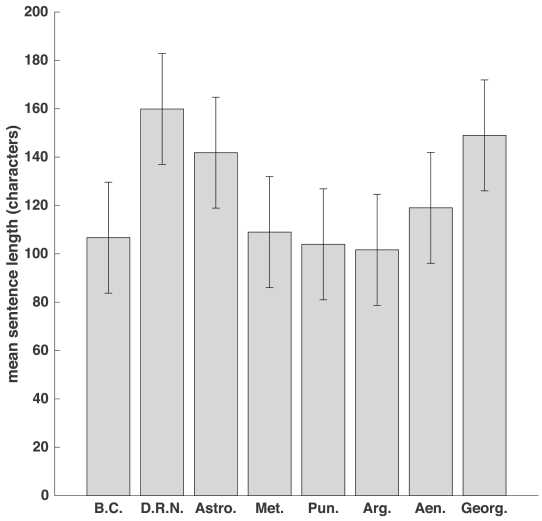

Figure 3: Mean sentence length of Latin epic poems (in characters). Error bars denote one s.d. across the eight poems.

The toolkit also reveals that Latin drama has a higher frequency of personal pronouns than other verse genres, as shown in Figure 4. This is no doubt due to drama's dialogic form: characters speak to each other directly, often employing first ("I" or "we") and second person ("you") pronouns. Many other literary genres primarily employ a narrative structure in which a narrator describes the action. This narrative type often uses third person pronouns ("he,""she," "it"), but rarely uses first or second person pronouns. Accordingly, the frequency of personal pronoun use is higher in drama. While this difference may be intuitive to a reader, the large-scale data generated by the toolkit offers quantitative evidence of a genre's formal style, which would otherwise be difficult if not im-

possible to calculate by hand.

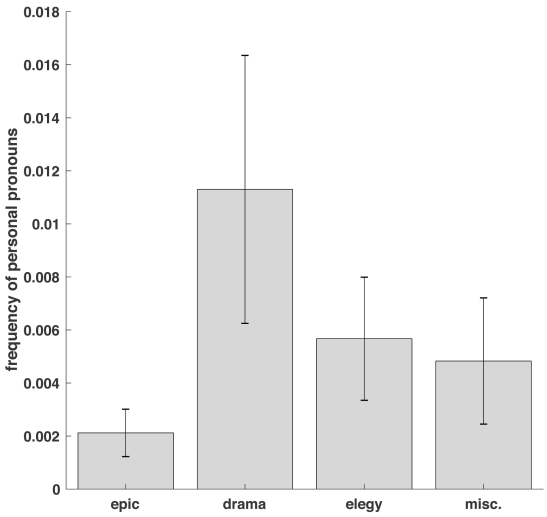

Figure 4: Mean per-character frequency of personal pronouns in the major genres of Latin literature. Error bars denote one s.d. across the texts within each of the four genres.

Finally, the toolkit can also generate input data for supervised and unsupervised machine learning analyses. In our recent study of Latin prose and verse, we trained a random forest classifier using all 26 features to distinguish the two genres with high (>97%) accuracy (Chaudhuri et al., 2018). The underlying data can now be produced easily using the toolkit, and similar datasets can be constructed for other machine learning applications.

5 Conclusion and Future Work

This paper introduces a stylometry toolkit for Latin literature, which incorporates a diverse feature set demonstrably useful for literary criticism. The toolkit includes a point-and-click interface to maximize usage among core domain specialists, principally researchers in the humanities, who may not have specialized computational training. Future versions of the toolkit will further diversify the feature set, incorporating high-frequency n-grams and sense-pauses alongside the existing categories (Fitch, 1981; Dexter et al., 2017), and will leverage expected advances in Latin NLP to improve the methods for calculation of existing features.

In related work, we have developed a similar feature set for Ancient Greek, which has been used to classify prose and verse and, at a more fine-grained level, epic and drama (Gianitsos et al., 2019). Our work on Old English has demonstrated the utility of related features for various literary and attribution studies (Neidorf et al., 2019). After extension of the current toolkit to Ancient Greek and Old English, we plan in due course to incorporate other underserved languages, in particular Bengali.

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 a National Endowment for the Humanities Digital Humanities Start-Up Grant (grant number HD-10 248410-16) and an American Council of Learned Societies (ACLS) Digital Extension Grant. T.J.B. was supported by an Engaged Scholar Initiative Fellowship from the Andrew W. Mellon Foundation, P.C. by an ACLS Digital Innovation Fellowship and a Mellon New Directions Fellowship, and J.P.D. by a Neukom Fellowship.

References

J.N. Adams. 1972. The language of the later books of Tacitus' *Annals*. *The Classical Quarterly*, 22(2):350–373.

J.N. Adams, M. Lapidge, and T. Reinhardt. 2005. Introduction. In J.N. Adams, M. Lapidge, and T. Reinhardt, editors, *Aspects of the Language of Latin Prose. Proceedings of the British Academy, 129*, pages 1–36. Oxford University Press, Oxford.

P. Chaudhuri, J.P. Dexter, T. Dasgupta, and K. Iyer. 2018. A small set of stylometric features differentiates Latin prose and verse. *Digital Scholarship in the Humanities*.

D.L. Clayman. 1981. Sentence length in Greek hexameter poetry. In R. Grotjahn, editor, *Hexameter Studies. Quantitative Linguistics 11*, pages 107–136. Brockmeyer, Bochum.

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

G. Crane. 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*, pages 3–10.

J.P Dexter, T. Katz, N. Tripuraneni, T. Dasgupta, A. Kannan, J.A. Brofos, J.A. Bonilla Lopez, L.A.

Schroeder, A. Casarez, M. Rabinovich, A. Haimson Lushkov, and P. Chaudhuri. 2017. Quantitative criticism of literary relationships. *Proceedings of the National Academy of Sciences USA*, 114(16):E3195–204.

M.D.C. Drout, M.J. Kahn, M.D. LeBlanc, and C. Nelson. 2007. Of dendrogrammatology: Lexomic methods for analyzing relationships among Old English poems. *Journal of English and Germanic Philology*, 110(3):301–336.

M. Eder, J. Rybicki, and M. Kestemont. 2016. Stylometry with R: A package for computational text analysis. *The R Journal*, 8(1):107–121.

A. Erdmann, C. Brown, B. Joseph, M. Janse, P. Ajaka, M. Elsner, and M.-C. de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan.

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

E.T. Gianitsos, T.J. Bolt, P. Chaudhuri, and J.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.

M. Jockers. 2013. *Macroanalysis*. University of Illinois Press, Champaign, IL.

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

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

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

F. Moretti. 2013. *Distant Reading*. Verso, London.

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

L. Neidorf, M.S. Krieger, M. Yakubek, P. Chaudhuri, and J.P. Dexter. 2019. Large-scale quantitative profiling of the Old English verse tradition. *Nature Human Behaviour*, 3(6):560–567.

A. Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press, Chicago.

J. Stover, Y. Winter, M. Koppel, and M. Kestemont. 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):239–243.

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

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