Quantitative criticism of literary relationships

Joseph P. Dexter^a,b,2, Theodore Katz^b,c,d,1, Nilesh Tripuraneni^e,h,1, Tathagata Dasgupta^a,1, Ajay Kannan^f, James A. Brofos^f, Jorge A. Bonilla Lopez^g, Lea A. Schroeder^h, Adriana Casarez^i, Maxim Rabinovich^j, Ayelet Haimson Lushkov^k, and Pramit Chaudhuri^g,b,2

^aDepartment of Systems Biology, Harvard Medical School, Boston, MA 02115; ^bThe Dalton School, New York, NY 10128; ^cResearch Science Institute, Center for Excellence in Education, McLean, VA 22102; ^dDepartment of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; ^eDepartment of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom; ^fDepartment of Computer Science, Dartmouth College, Hanover, NH 03755; ^gDepartment of Classics, Dartmouth College, Hanover, NH 03755; ^hAustin Independent School District, Austin, TX 78703; ^iDepartment of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720; and ^jDepartment of Classics, University of Texas, Austin, TX 78712

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved February 27, 2017 (received for review July 20, 2016)

Authors often convey meaning by referring to or imitating prior works of literature, a process that creates complex networks of literary relationships ("intertextuality") and contributes to cultural evolution. In this paper, we use techniques from stylometry and machine learning to address subjective literary critical questions about Latin literature, a corpus marked by an extraordinary concentration of intertextuality. Our work, which we term "quantitative criticism," focuses on case studies involving two influential Roman authors, the playwright Seneca and the historian Livy. We find that four plays related to but distinct from Seneca’s main writings are differentiated from the rest of the corpus by subtle but important stylistic features. We offer literary interpretations of the significance of these anomalies, providing quantitative data in support of hypotheses about the use of unusual formal features and the interplay between sound and meaning. The second part of the paper describes a machine-learning approach to the identification and analysis of citation material that Livy loosely appropriated from earlier sources. We extend our approach to map the stylistic topography of Latin prose, identifying the writings of Caesar and his near-contemporary Livy as an inflection point in the development of Latin prose style. In total, our results reflect the integration of computational and humanistic methods to investigate a diverse range of literary questions.

Significance

Famous works of literature can serve as cultural touchstones, inviting creative adaptations in subsequent writing. To understand a poem, play, or novel, critics often catalog and analyze these intertextual relationships. The study of such relationships is challenging because intertextuality can take many forms, from direct quotation to literary imitation. Here, we show that techniques from authorship attribution studies, including stylometry and machine learning, can shed light on inexact literary relationships involving little explicit text reuse. We trace the evolution of features not tied to individual words across diverse corpora and provide statistical evidence to support interpretive hypotheses of literary critical interest. The significance of this approach is the integration of quantitative and humanistic methods to address aspects of cultural evolution.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission. Freely available online through the PNAS open access option.

1J.P.D., T.K., N.T., and T.D. contributed equally to this work.
2To whom correspondence may be addressed. Email: jdexter@fas.harvard.edu or pramit.chaudhuri@utexas.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1611910114/-/DCSupplemental.

Tell me the significance of this approach.

The significance of this approach is the integration of quantitative and humanistic methods to address aspects of cultural evolution. It has been a longstanding goal of research in the digital humanities to integrate quantitative methods with the aims of literary study. Following the lead of Burrows’ 1987 book, *Computation into Criticism,* more recent attempts have involved the theorization and implementation of methods of “distant reading” (7, 8), “algorithmic criticism” (9), “macroanalysis” (10), and “literary pattern recognition” (11). This work has been augmented by additional theoretical analyses (12, 13) and empirical studies that exploit specific methodological innovations, such as topic modeling, often for large-scale profiling of genres or periods (10, 14, 15). Quantitative methods have been especially valuable for the characterization of intertextuality both classical and modern. Computational searches for lexically similar phrases, similarity between the works. We take as a case study two problems in classical Latin literature that are of substantial current interest to literary critics and historians. The literature of the Roman Republic and Empire contains an extraordinary density and diversity of intertextual parallels. Intertextuality has become an essential focus of modern critics of Latin literature, and detailed qualitative taxonomies of Latin intertextuality have been constructed (3–5). Another advantage of our focus on classical literature is the near-complete digitization of extant texts in searchable, high-quality databases (6).

It has been a longstanding goal of research in the digital humanities to integrate quantitative methods with the aims of literary study. Following the lead of Burrows’ 1987 book *Computation into Criticism,* more recent attempts have involved the theorization and implementation of methods of “distant reading” (7, 8), “algorithmic criticism” (9), “macroanalysis” (10), and “literary pattern recognition” (11). This work has been augmented by additional theoretical analyses (12, 13) and empirical studies that exploit specific methodological innovations, such as topic modeling, often for large-scale profiling of genres or periods (10, 14, 15). Quantitative methods have been especially valuable for the characterization of intertextuality both classical and modern. Computational searches for lexically similar phrases,
exemplified by the work of the Tesseræ and Perseus projects on Greek and Latin literature, are useful for the high-throughput identification of local verbal intertexts (16–19). Such work was highlighted in a 2016 special issue of the journal *Digital Humanities Quarterly* devoted entirely to digital methods and classical studies (20). Digitization of enormous corpora, such as Google Books and the Project Gutenberg Digital Library, has enabled “culturomic” analyses of global linguistic trends (21–24). A notable recent application of such methods was a large-scale study of stylistic influence in English literature based on use patterns of “content-free” words (25). Finally, quantitative stylistic analyses have long been used to clarify gross relationships between texts. Standard applications of stylometry include dating literary works and resolving questions of attribution (26–30). Both ad hoc stylometric analysis and supervised machine learning with stylometric features have proven successful for such applications (31–33), including for cases in Latin literature (34).

Whether an entire work is spurious or authentic, however, is a coarser question than typically posed in literary criticism. Of greater interest is how the spurious work differs from authentic writings and how its composition was influenced by the larger tradition. Recent studies have begun to repurpose stylometry to answer such literary critical questions (10, 35–39). Much of this research relies on the suitability of techniques of authorship attribution for addressing broader literary questions (40). Here, we show that complex relationships between partially similar texts, exemplified at short scales by literary paraphrase and large scales by creative imitation of entire works, can be characterized through the application of stylometry and machine learning, core methods in computational attribution studies. Although much work in computational text analysis has focused on the word or phrase as the principal unit of analysis, some recent research has shown the utility of other kinds of units, such as character and rhythm, in both large- and small-scale quantitative analyses of literature (41, 42). Our work quantifies a selection of subverbal, syntactic, and prosodic features, which have also been used for authorship attribution. We redeploy these techniques to resolve multiple literary problems of interest to classicists and other humanists.

The philosopher and statesman Seneca (4 BC to AD 65) (Fig. 1B) wrote tragic plays, 10 of which have been transmitted under his name via the medieval manuscript tradition and hugely influenced later dramatists, such as Shakespeare and Racine (43, 44); 2 of these 10 (the *Octavia* and the *Hercules Oetaeus*) are spurious, however, the work of careful imitators writing in the years after Seneca’s death. Despite considerable attention, the precise literary and stylistic relationships among both the 8 works attributed to Seneca and the entire corpus of 10 transmitted texts remain unclear. Our computational analysis identifies several subtle but significant differences in poetic style between the *Octavia* and the *Hercules Oetaeus* and the eight authentic tragedies. We extend these methods to contrast typical Senecan style with that of the *Proone*, a neo-Latin tragedy influenced by Seneca but written centuries after his death, and the *Phoenissae*, an authentic but incomplete play. Although easily tabulated computationally, the differentiating features cannot be studied using traditional means without substantial repetitive effort.

The historian Livy (64 or 59 BC to AD 17) (Fig. 1C) wrote a monumental history of Rome covering the period from the city’s foundation and the rise of the Roman empire to Livy’s contemporary world. The work consisted of 142 books (~2 million words), of which only 35 survive. Livy makes frequent reference to previous works of history, but his citational practices to Seneca and the entire corpus of 10 transmitted texts remain unclear. Our computational analysis identifies several subtle but significant differences in poetic style between the *Octavia* and the *Hercules Oetaeus* and the eight authentic tragedies. We extend these methods to contrast typical Senecan style with that of the *Proone*, a neo-Latin tragedy influenced by Seneca but written centuries after his death, and the *Phoenissae*, an authentic but incomplete play. Although easily tabulated computationally, the differentiating features cannot be studied using traditional means without substantial repetitive effort.

Although much work in computational text analysis has focused on the word or phrase as the principal unit of analysis, some recent research has shown the utility of other kinds of units, such as character and rhythm, in both large- and small-scale quantitative analyses of literature (41, 42). Our work quantifies a selection of subverbal, syntactic, and prosodic features, which have also been used for authorship attribution. We redeploy these techniques to resolve multiple literary problems of interest to classicists and other humanists.

Fig. 1. Intertextuality in Seneca and Livy. (A) Categories of intertextuality. Instances of intertextuality can be characterized according to the similarity between the source text and intertext and the scope of the association. For instance, a short quotation (upper left) exhibits higher similarity and narrower scope than a loose adaptation of an entire play (lower right). The primary focus of the paper is imitation of Seneca and citation/paraphrase in Livy (gray box). (B) Timeline indicating the dates of composition of the texts analyzed. The eight tragedies of Seneca are often divided into early (1), middle (2), and late (3) groups. The two pseudo-Senecan tragedies were composed shortly after his death. Dotted lines indicate the dates of death of Livy and Seneca. (C) Schematic of Livy’s history of Rome, which contained 142 books. Books 11–20 and 46–142 have been lost; the subject matter of the surviving books is summarized.
This variation stands in contrast to the monotony of an unbroken series of end-stopped lines (i.e., those lines in which the meaning is complete by the end of the line and marked by firm punctuation). One plausible explanation of the unusually high incidence of enjambment in the Procle is the desire of the young author—only 18 years old at the time—to display his virtuosity in Latin verse composition in part through the use of a feature that signified confident poetic technique. Although we possess no direct evidence of Correr’s intent with respect to enjambment in particular, the playwright did preface his drama with a discussion of the varied meters used in the course of the text, including explicit discussion of meters that are rare in tragedy but more commonly found in comedies. Correr’s frequent exploitation of enjambment can thus be considered complementary to his similar exploitation of the full array of Latin metrical forms, which went well beyond the range of meters used in Seneca’s Thyestes (his primary classical model). The intertextual relationship between the Procle and its Senecan predecessors thus consists partly of similarities that highlight the tradition in which Correr is working and partly of differences (in this case, a difference in verse composition) that highlight Correr’s distinctiveness within that tradition.

To investigate another potential stylistic difference, we next examined the use of relative clauses across the Senecan corpus. The relative clause, constructed using the relative pronoun who or which, is a standard method of subordinating one thought to another within a sentence. In Latin, relative pronouns are the various inflected forms of qui (Materials and Methods and SI Appendix, Text and Table S3 have details and error analysis). We computed the fraction of noninterrogative sentences with at least one relative clause for the 10 Senecan and pseudo-Senecan tragedies; interrogative sentences were excluded to obviate the need for semantic parsing of relative and interrogative pronouns, which are often identical morphologically. The count revealed that almost one-quarter of sentences in the Octavia contain a relative clause (Fig. 2B and SI Appendix, Fig. S1B), whereas the fraction for all other tragedies is below 20%. The Octavia stands out from the remainder of the corpus as a drama on a historical subject—the divorce and death of Nero’s wife and the event’s political context—in contrast to the mythological subjects of the other nine plays. The combination of non-Senecan authorship and historical subject matter has led critics to look for stylistic differences in the language and syntax of the work. With varying degrees of persuasiveness, claims have been made for the tragedy’s comparatively less elaborate style, more colloquial speech, and features typically avoided in poetry (48). Our identification of the enrichment of relative clauses provides systematic, quantitative evidence that the Octavia’s syntax is distinctive from that of the other plays. The reason for this more hypotactic style is unclear. One possible explanation is that subordinating constructions of this kind indicate a more prosaic style, which could be an authorial habit or reflective of a more specific consideration. Partial corroboration of such a style can be found in specific instances identified by literary critics, such as the concatenation of relative clauses at lines 111 and 113 (48). The literary influence of Seneca’s prose writing, especially the De Clementia, might also account for the Octavia’s more prosaic style (49).

**Results**

Quantitative Criticism Identifies Literary Differences Across the Senecan Corpus and Tradition. We profiled a broad range of stylistic features across the whole Senecan and pseudo-Senecan corpus and in Gregorio Correr’s Procle, a 15th century neo-Latin tragedy deeply influenced by Seneca. We first considered sense pauses (interruptions in speech indicated by any punctuation mark other than a comma), which have proven useful in manual studies of Senecan style. We observed almost no variation in the length-normalized number of sense pauses across the eight authentic Senecan tragedies (Fig. 2A, i). In contrast, total sense pauses were significantly reduced (Octavia or enriched (Hercules Oetaeus and Procle) in the Senecan-influenced tragedies (Fig. 2A, i and SI Appendix, Fig. S1A, i), suggesting that the imitators either deliberately disregarded or failed to replicate a typical, if likely unconscious, aspect of Senecan style.

We then recapitulated a seminal literary critical study that used manual tabulation of sense-pause statistics to establish a relative chronology for the eight authentic tragedies (46). In contrast to total sense pauses, the ratio of intraline (sense pauses that do not coincide with line breaks in the iambic trimeter verse) to total sense pauses is more heterogeneous across the tragedies, as reported by Fitch (46) and supported by our computational analysis (Fig. 2A, ii and SI Appendix, Fig. S1A, ii). On the basis of this variation, Fitch (46) divided the tragedies into three groups, which we confirmed differ significantly in intraline to total sense-pause ratio (SI Appendix, Fig. S2). By analogy with the stylistic development of other playwrights, Fitch (46) further suggested that the ratio is higher in Seneca’s later tragedies as the playwright became more skillful at exploiting tension between the basic units of meaning and meter. This relative chronology of Seneca’s plays has been widely influential in classics, and even critics who disagree with Fitch’s placements (46) of individual works have tended to retain the majority of his ordering (47). Fitch (46) excluded from his study the two tragedies in the corpus considered spurious. Ferri (48) has applied Fitch’s method and in Gregorio Correr’s Dexters et al. PNAS | Published online April 3, 2017 | E3197

**Phonetic and Thematic Analyses of the Octavia and the Phoenix sae.** Functional n-grams are short, syllable-length strings of characters, which can reflect ingrained authorial style and capture patterns of sound in poetry. Analysis of functional n-grams has proven useful for authorship attribution studies and addressed literary questions in the postclassical reception history of the Roman poet Catullus (37). Although critics have long paid attention to specific aural effects and sound play in poetry, systematic studies have been infeasible without computational tabulation of n-grams.
Fig. 2. Quantitative comparison of Senecan and pseudo-Senecan literary style. (A, i) Total sense pauses in each tragedy. (A, ii) Ratio of intraline to total sense pauses. (A, iii) Frequency of enjambment. (B) Fraction of noninterrogative sentences containing at least one relative clause. The Octavia is at Q3 + 1.46IQR, where Q is the quartile and IQR is the interquartile range. Frequencies of the five most common (C, i) three and (C, ii) four grams in the Octavia (dark gray bars). Light gray bars show the mean frequencies of each n-gram across the tragedies. (D, i) Frequency of the four-gram ente. (D, ii) Frequency of ente in choral and nonchoral passages. Each circle denotes the frequency in one tragedy. The Phoenissae lacks choral odes and was, therefore, excluded from the group on the right. The difference is nonsignificant (p = 0.10 by a two-tailed unpaired t test). (D, iii) Spatial distribution of ente in 10 tragedies. Each vertical line denotes one or more instances of ente at that position. (D, iv) Fraction of instances of ente that occur within clusters in each tragedy. The dark gray bars indicate instances within one line of each other, and the light gray bars indicate instances within three lines of each other. All frequencies are per character. In all plots, the dotted lines denote the mean of the relevant quantity across all tragedies, except the Procne. Error bars denote 1 SD across the tragedies. Senecan and pseudo-Senecan tragedies are referred to by abbreviations given in the Oxford Classical Dictionary: Ag, Agamemnon; HF, Hercules Furens; HO, Hercules Oetaus; Med, Medea; Oct, Octavia; Oed, Oedipus; Pha, Phaedra; Phoen, Phoenissae; Tro, Troades; Thy, Thyestes. The Procne is a neo-Latin tragedy written in 1428 by Gregorio Correr. *Outliers (defined as >Q3 + 1.5IQR or <Q1 − 1.5IQR).
We initially examined the most common functional bigrams (two-letter strings) in the Octavia and the Hercules Octaeus and found that their frequency was comparable in both the spurious and authentic tragedies (SI Appendix, Fig. S4). This result prompted us to repeat the analysis for the Octavia with functional trigrams, for which we observed clear differences (Fig. 2C, i). Of particular interest, two of the six most common trigrams in the Octavia (tri and ris) are elevated compared with the authentic tragedies. The enrichment of particular n-grams points to the author’s disposition toward a particular sound and possibly words containing those n-grams. In the case of the Octavia, those words are the various inflected forms of tristis (sad, stern) and noster (our), which together appear 69 times in the Octavia and account for more than 60% of the instances of tri and ris. The frequent use of tristis and noster is also reflected in the enrichment of the four-grams tris, nost, ostr, and stra (Fig. 2C, ii).

As an example of the kind of literary critical hypotheses that can be supported by analysis of functional n-grams, we might interpret the frequency of the appearance of tristis as substantiating the mood of lament and pessimism that pervades much of the Octavia, over and above what is typical even for Senecan tragedy. The enrichment of inflected forms of noster suggests a different but compatible hypothesis. Although the date and possible performance context of the Octavia are unknown, on the basis of its negative characterization of Nero scholars have argued that it was composed in the wake of Nero’s death, either during or shortly after the period of civil wars known as the Year of the Four Emperors (AD 69). Much of the drama is concerned with Nero’s tyrannical behavior and removal of opposition, and the play ends with mention of a popular uprising in support of Octavia. It thus dwells on various claims on political authority. The frequent use of the word noster (our) in the play repeatedly emphasizes the ownership that various parties feel over, for instance, the city (nostra urbs) or the imperial household (nosta domus). Resolving these rival claims is both the plot of the drama and a stimulus for the post-Neronian audience to reflect on the significance of such claims for their own time (discussed in detail in SI Appendix, Text).

Although written by Seneca, the Phoenissae has long been recognized as distinct from the remainder of the corpus (S0). It is several hundred lines shorter than any other tragedy and obviously incomplete. Another distinctive aspect of the Phoenissae is that it does not contain any odes sung by a chorus, which are a standard component of Roman tragedy and present in all other Senecan and pseudo-Senecan tragedies. In our analysis of functional n-grams across the Senecan corpus, we found that the four-gram ente is significantly enriched in the Phoenissae (Fig. 2D, i and SI Appendix, Fig. S1C, i). This enrichment is specific to ente; related four grams, in which “nt” is immediately preceded and succeeded by any vowel, are not enriched in the Phoenissae (SI Appendix, Fig. S5). The enrichment of “vowel + nt + vowel” four grams in the Thyestes is a consequence of frequent references to Tantalus, an important character in that tragedy (SI Appendix, Fig. S5). Furthermore, there is no significant difference between the frequency of ente in choral and nonchoral passages across the Senecan corpus (Fig. 2D, ii), suggesting that the concentration of ente in the Phoenissae cannot be explained by its peculiar structure.

We examined the spatial distribution of instances of ente in the tragedies (Fig. 2D, iii), which revealed that the four gram is often repeated in close proximity in the Phoenissae. This effect, as measured by the fraction of instances of ente occurring within three-line clusters, is specific to the Phoenissae (Fig. 2D, iv). Additionally, clusters of the generic vowel + nt + vowel four gram are not enriched in any tragedy other than the Thyestes (SI Appendix, Fig. S6). As such, variations in its frequency might reflect some stylistic choice by the author, especially when clustered to create a partial echo.

Repetition of words for stylistic effect is a common feature of Senecan tragedy and the Phoenissae in particular, which exhibits frequent instances of exact repetition (e.g., sequor, sequor at 40 and ibo, ibo at 12 and 407) and morphological variation (e.g., patris ... pater at 55, frater ... fratrem at 355, and pectus ... pectori at 470). These formal repetitions often possess literary significance. In the Phoenissae, for instance, clusters of familial terms highlight the play’s thematic focus on a civil war fought between two brothers (51). The repetitions cited by critics, however, operate at the level of the word (whether exact or a morphological variant) rather than purely phonetic elements, such as ente. Traditional critical approaches, based on reading or word searches, are thus poorly equipped to detect subtler forms of repetition manifested in smaller units.

The clusters of ente in the Phoenissae include repetitions of both whole words and morphological endings. Repetitions often serve to emphasize ideas or feelings important to the drama. At 368 and 369, for instance, Jocasta uses the word nocentes (guilty) in successive lines to amplify her sense of her own wrongdoing; n-gram analysis is especially useful for the identification of clusters of nonidentical, even etymologically unrelated words. To give one example, at 98–100, nolentem and cupientem (desiring) are paired in opposition to each other, a contrast highlighted by the aural echo of the ending. Other clusters of nonidentical words containing ente highlight themes of sexual aberration (467–469) and moral responsibility (451–454) that are important to the subject matter of the play (SI Appendix, Text).

Furthermore, we suggest that Senecan’s greater propensity to exploit the repetition of this sound is consistent with the word-level repetitions already observed by critics as part of a larger stylistic aim. Seneca seems to use repeated words and sounds in close proximity in a systematic way. In dramatizing the mythological war between the twins Polynices and Eteocles, the Phoenissae is especially concerned with repetition, doubling, and assimilation—features that suffuse the speech, themes, and structure of the play. Although impossible to determine with any certainty, our inference about the frequent clustering of adjectival or participle endings in the Phoenissae, which are often used to signal apparent contrasts or amplifications, is that they embody at the level of sound a larger concern with repetition that defines the drama as a whole.

Anomaly Detection Differentiates Suspected Citations from Other Livian Material. We next considered citation and paraphrase, a class of intertextuality of comparable similarity but narrower scope than creative imitation of entire works (Fig. L4) and potentially amenable to techniques of authorship attribution. We took as a case study the use of source material in Livy’s enormous history of Rome. The scope of Livy’s writings required that he consult a wide variety of sources, mostly earlier historians but also published speeches and other texts. Like other historians, the manner in which Livy used his sources was equally varied, ranging from direct quotation and referential citation (“I found these numbers in X”) to vague indications of a source (“some say,” “I read somewhere”) (45, 52, 53). Literary critics have also shown that, in certain places, Livy uses a specific source without explicitly saying so (54). The nature of Livy’s source use is made even more opaque by the loss of most of the source texts in addition to the loss of the majority of his own history. Classical scholars have debated inconclusively the extent to which the text of earlier sources can be reconstructed from Livy’s citational passages (i.e., passages that include a citational gesture, whether a reference to a specific author or a more indirect suggestion of source use) (55, 56). The paucity of extant source material poses an extreme challenge for standard stylometric identification (whether manual or computational) of Livian citations. Following our approach with pseudo-Senecan tragedy, we used a combination of computational and literary critical approaches.
to achieve an improved understanding of Livy's citational practice. Our main result is the development of an anomaly detection algorithm that can differentiate Livian citations from noncitational material (i.e., the vast majority of the text) using stylometric features.

Our analysis relied on a database previously developed by one of the authors (A.H.L.) for use in literary research, which catalogs citational passages in the extant parts of Livy's history. The database was compiled by noting all passages (in an English translation) in which Livy suggests use of source material, whether by explicit identification of a source or through citational language. In total, the database contains 439 citational passages.

We first performed a simple computational test to confirm the linguistic basis for the citation database. We compared the frequency of four representative citational phrases (fama est, it is rumored that; annalibus, in the annals; scribit, he writes; tradit, he reports) between the citation database and the rest of Livy and found, as expected, that these terms are enriched significantly in the database (Fig. 3A, i). We also examined the distribution of citations across Livy (Fig. 3A, ii). Over 50% of entries in the database occur in the first decade of Livy. Consistent with this enrichment of citations, the frequency of the citational phrase annalibus is significantly higher in the first decade (SI Appendix, Fig. S7).

We next assembled a large set of Latin stylometric features that might be useful for distinguishing citational and noncitational material. The set consists of 25 features encompassing many items of stylistic interest, including noncontent words, specific syntactic constructions, and length of sentences and clauses (SI Appendix, Table S4). As discussed above, Livy's source texts are largely not extant, which precludes the application of binary classification. As an alternative, we used a one-class support vector machine (SVM) as an anomaly detection algorithm. The one-class SVM was trained on the Livian corpus (with some material excluded for cross-validation) and used to classify material in the citation database as anomalous (non-Livian) or nonanomalous (Livian). A primary challenge in the analysis of the citation database is the length of individual entries, many of which include only a few sentences. To generate meaningful feature statistics, we aggregated multiple citations into “bins” randomly and analyzed each bin as if it were a single passage (37). We set the bin size at 35 sentences, which was the minimum passage length for which we obtained consistent results (SI Appendix, Fig. S8). To maintain consistency, we also binned test material from Livy and other authors studied, even if extensive material was available.

For the citation database, we found that the fraction of bins classified as Livian was very low (less than 10%), regardless of the Livian material used for training (Fig. 3B). In contrast, ∼80% of bins from Livian material withheld for cross-validation were classified as Livian. The correct identification of most of the cross-validation material as Livian and the substantial difference between the cross-validation material and the citation database validate the model as an effective tool for the analysis of citations. The fact that a small amount of Livian material was classified as anomalous likely reflects the well-known heterogeneity of Livy’s style across 35 books of his history (57) and the general tendency of one-class anomaly detection methods to classify some test material as anomalous (58). For instance, Yilmaz et al. (59) used a one-class SVM to analyze a corpus of government documents and reported false negative rates between 29 and 47% (substantially higher than we obtained for Livy), depending on the features used.

We then investigated which of the stylometric features were most effective for differentiating citational material. We reasoned that markers of hypotactic style (extensive use of subordinate clauses) might be particularly important, because the earlier historians on whom Livy drew are generally held to have favored a simpler sentence structure (parataxis) in contrast to Livy’s more varied and hierarchical syntax (60). Consistent with this hypothesis, we identified five features (mean sentence length, variance of sentence length, fraction of noninterrogative sentences containing at least one relative clause, mean length of relative clauses, and mean number of relative clauses per sentence) sufficient to establish a clear difference between citational and noncitational material (SI Appendix, Fig. S9). All five of these features relate to various aspects of the organization of sentences and together reflect tendencies toward hypotactic or paratactic style. Use of this low-dimensional feature set also enabled reduction of the bin size to 20 sentences (SI Appendix, Fig. S8) and a correspondingly finer-grained characterization of the citation database.

We applied our anomaly detection procedure with the reduced feature set to a passage that has provoked particular controversy over Livy’s use of source material. Toward the end of Book 38, Livy describes a complicated sequence of events in the late career of Scipio Africanus, the famous Roman general. Focused primarily on the legal tribulations of Scipio and his brother, Livy’s narrative is divided into two contrasting accounts, with the second largely undermining the first (61). The first account follows that of an earlier historian, Valerius Antias, whom Livy explicitly cites as a source. The second follows a number of other sources, including records of various speeches made by some of the principal participants in the events. Modern commentators have disagreed in particular on the extent to which Livy reused Valerius Antias, with judgments ranging from minimal reuse to extensive quotation (62). We applied our method to this narrative to ascertain whether there is a meaningful stylistic difference between the two accounts and determine which account, if either, differs from Livy’s typical style. We divided the whole narrative into two sections large enough to include a substantial portion of text: the first (38.50.1–51.14) putatively more indebted to Valerius Antias, and the second (38.54.1–60.10) indebted to other sources. The one-class SVM classified the first section as “non-Livian” and the second section as “Livan.” The result corroborates the view that Livy’s first account was substantively influenced by Valerius Antias. However, it does not indicate whether such influence amounts to quotation, imitation, or a subtler stylistic effect. Both Livy and his sources may have a shared interest in some historical or political theme, and together reflect tendencies toward hypotactic or paratactic style. Use of this low-dimensional feature set also enabled reduction of the bin size to 20 sentences (SI Appendix, Fig. S8) and a correspondingly finer-grained characterization of the citation database.

Profiling the Development of Latin Prose Style. Given the clear difference observed between bulk Livy and the citation database, we next hypothesized that post-Livian historiography, and perhaps even imperial prose in general, would resemble bulk Livy more closely than citational material. The hypothesis was based on an assumption that Livy’s sources would show traces of an earlier prose style, whereas Livy’s own style was part of a more generally influential movement that would be reflected in later authors. Our approach was to assess the “Livaniness” of 17 non-Livian texts using the reduced feature set and the same methodology applied to the citation database. We chose a wide-ranging corpus consisting of prose and poetry from a variety of genres and periods. The poetry was used as a control group. As expected, all five works—including comedy, tragedy, epic, and philosophical poetry from times before, after, and contemporaneous with Livy—scored as extremely non-Livian. The prose texts were also of various genres, including speeches, letters, and technical treatises in addition to historiography.

We observed a clear difference between most pre- and post-Livian prose. Of the pre-Livian material, the nonhistorical texts registered as very non-Livian, quite unlike Caesar’s historiographical accounts of his wars in Gaul and a few years later.
Fig. 3. Anomaly detection differentiates cited material from the rest of Livy. (A, i) Comparison of the frequency of four “signal words” indicating potential instances of citation (fama est, annalibus, scribit, and tradit) between all of Livy (left) and the citation database (right). *p < 0.05 by a two-tailed unpaired t test. (A, ii) Frequency of entries in the citation database across 35 extant books of Livy. (B) Fraction of bins (random aggregates of 35 sentences) classified as Livian from bulk Livian material (left) and the citation database (right) by a one-class SVM using a set of 25 stylometric features. Results are the mean ± 1 SD of 35 leave-one-out cross-validation experiments. ***p < 0.001 by a two-tailed unpaired t test. (C) Fraction of 20-sentence bins from a range of Latin literature classified as Livian using a reduced set of five stylometric features. Works are referred to by abbreviations given in the Oxford Classical Dictionary: Agr, Cato’s De Agri Cultura; Ann, Tacitus’ Annals; Conf, Augustine’s Confessions; De or, Cicero’s De oratore; De rep, Cicero’s De republica; Cat, Sallust’s De coniuratione Catilinae; G, Vergil’s Georgics; Gal, Caesar’s Bellum Gallicum; Ger, Tacitus’ Germania; HF, Seneca’s Hercules Furens; Inst 1, Quintilian’s Institutio Oratoria 1; Iug, Sallust’s Bellum Iugurthinum; Lucr, Lucretius’ De rerum natura; Mur, Cicero’s Pro Murena; Ps, Plautus’ Pseudolus; Theb, Statius’ Thebaid; Vit, Vitruvius’ De architectura. Genres represented include historiography (Gal, Cat, Iug, Ger, and Ann), nonhistoriographical prose (Agr, De or, Mur, De rep, Vitr, Inst 1, and Conf), comedy (Ps), tragedy (HF), and poetry in dactylic hexameter (G, Lucr, and Theb). Prose and poetic texts are arranged chronologically. (D) Proposed outline of the development of Latin prose style; + indicates similarity to the style of Caesar and Livy.
Sallust’s two monographs on historical topics, the *De coniuratione Catilinae* and the *Bellum Iugurthinum*. The result for Caesar’s text, in particular, corroborates standard scholarly views about the resemblance between Caesar’s and Livy’s sentence structures and may reflect similarities in subject matter (57). The intermediate similarity of Cicero’s *De re publica* suggests that content indeed plays a part in style. Unlike the two other Cicero-rian works, a speech (*Pro Murena*) and a rhetorical treatise (*De oratore*), the *De re publica* contains more explicit discussions of history and politics in a narrative style. This fact may account for the work’s greater resemblance to Livy’s history. In the case of the later prose writers, however, even rhetorical (*Institutio Oratoria*) and technical (*De architecture*) treatises score as Livian, extending to Augustine’s autobiographical *Confessions* written almost 400 years later. We note that two historiographical works by Tacitus (the *Germania* and the *Annales*) both seem particularly Livian in style (even slightly more so than bulk Livy). The difference between bulk Livy and Tacitus is far smaller than that between bulk Livy and the citation database or between early and later prose. The strong similarity, however, does suggest that Tacitus might have been influenced by Livy’s syntax to a greater extent than has been appreciated previously (63).

On the whole, the two key observations are the difference between Livy and both pre-Livian prose and the material in the citation database and the similarity between Livy and Caesar and post-Livian prose. These results show in a quantitative and large-scale fashion a development in Latin prose style, namely that a stylistic shift occurred with Caesar, continued with Sallust and Livy, and exerted a critical influence on later prose literature (Fig. 3D). We find the effect of that influence even on genres, such as treatises, that had previously looked more unlike historiography. The results also reveal the extent to which Livy’s citation material—whether in the form of imitations, quotations, or stylistic modulations—differs from later prose style.

**Discussion**

**High-Throughput Data Generation for the Study of Literature and Culture.** Numbers and statistics have long played an important, if underappreciated, role in literary criticism. Commentators often cite tabulations of particular words or formal features to bolster their arguments; in the mid-20th century, Duckworth (64) published a detailed quantitative study of meter in Latin poetry that, despite some issues of methodology, has had broad influence in the field of classics. In this regard, one obvious application of computation to literature is the replication, at larger scale and with greater efficiency, of standard stylistic studies. In our computational analysis of sense pauses in Senecan tragedy, we were able to both recapitulate Fitch’s core results (46) efficiently and extend the scope of the original investigation. Accordingly, high-throughput methods are likely to have particular influence on the study of noncanonical material, such as the neo-Latin *Procne*, which receives negligible attention compared with famous classical authors, such as Vergil and Livy.

We find that frequency statistics on syllable-length n-grams can support literary criticism in two distinct but complementary ways. Highly enriched n-grams can point to patterns of word use that have thematic significance, as exemplified by our examination of tristis and noster in the *Octavia*. For such applications, the key advantage of functional n-gram analysis over simple word searches is that the former is untargeted, allowing for studies of diction even when the researcher does not have a specific hypothesis in mind. Additionally, functional n-grams enable the convenient investigation of colocalizations of sounds. Although criticism of poetry routinely reflects an intuitive understanding of aural effects, sound play and phonetic patterns are difficult to quantify using conventional methods. We suggest that analysis of short n-grams, an established technique in attribution studies and computational linguistics (65, 66), can inform literary critical studies of poetry’s aural quality. Functional n-grams are likely to be particularly useful when integrated with other computational approaches, such as the use by Forstall et al. (37) of functional bigrams as features for anomaly detection in literary texts.

**Quantitative Criticism: Attribution, Interpretation, and the Digital Humanities.** Computation has long been used for attribution and dating of literary works, problems that are unambiguous in scope and invite binary or numerical answers (27, 28). The recent explosion of interest in the digital humanities, however, has led to the key insight that similar computational methods can be repurposed to address questions of literary significance and style, which are often more ambiguous and open-ended. This turn from attribution to interpretation has been exemplified by the work of Jockers (10), who has pursued an approach to large-scale literary analysis termed “macroanalysis” (in analogy to macroeconomics). To this end, Jockers (10) has applied machine learning with stylometric features to trace patterns of influence across a substantial corpus of Latin prose, identifying Caesar, Sallust, and Livy as the key point in the development of Latin prose style. These results suggest that later prose authors were influenced by the style of Caesar and the writers in Caesar’s wake, including Livy, to a greater extent than has been previously acknowledged, even when writing about very different subject matter. Analogous phenomena have also been observed for the evolution of genres and literary style in English and other Latin corpora (7, 10, 25, 40). Throughout our work, we show the usefulness of incorporating syntactic and metrical features in addition to diction, noncontent words, and punctuation marks, which have been considered previously by Jockers (10) and others (25), into such comparative analyses.

Our approach, which we have termed “quantitative criticism,” relies on a productive fusion of humanistic and computational methods. Although indebted to much groundbreaking work in the fields of computational text analysis and authorship attribution, we intend the reference to “criticism” to signal an equal debt to literary study’s traditional concern with aesthetics and meaning. That said, we seek to use quantitative data to understand literary relationships and interpretative suggestion to evaluate experiments, so that the computational work of the scientist and the critical work of the humanist operate in symbiosis.

**Materials and Methods**

**Editions of Texts.** We used Peiper and Richter’s 1921 edition of Seneca (67) and Weissenborn and Müller’s 1911 edition of Livy (68) for all computational analyses. Both texts are freely and publicly available in searchable form through the Perseus Digital Library. For computational analysis of the *ProÌœne*, we scanned Grund’s 2011 text (69), applied optical character recognition, and manually corrected errors in the output. Sense-pause counts for the *Octavia* reported in SI Appendix, Fig. 53 were determined manually using Giardina’s 1966 text (70). All texts used in the comparison of Latin literary style reported in Fig. 3C are available through the Perseus Digital Library.

**Computation of Stylometric Features.** All natural language processing tasks were done using Python 2.7, and the code is freely and publicly available at https://github.com/qcrn. Copies of the relevant texts were obtained from the Perseus Digital Library as extensible markup language (XML) files and first stripped of all XML tags.

Following the definition of Fitch (46), sense-pause counts were determined by tabulation of punctuation marks other than commas (., ?, !, ;, :, ’, ‘, “, “) in a large English literary corpora, such as Victorian novels, and identify stylistic signatures of particular genres. Our analysis of the evolution of Latin prose style builds on such work in important ways. We repurpose anomaly detection to trace resemblances in a substantial corpus of Latin prose, identifying Caesar, Sallust, and Livy as a key point in the development of Latin prose style. These results suggest that later prose authors were influenced by the style of Caesar and the writers in Caesar’s wake, including Livy, to a greater extent than has been previously acknowledged, even when writing about very different subject matter. Analogous phenomena have also been observed for the evolution of genres and literary style in English and other Latin corpora (7, 10, 25, 40). Throughout our work, we show the usefulness of incorporating syntactic and metrical features in addition to diction, noncontent words, and punctuation marks, which have been considered previously by Jockers (10) and others (25), into such comparative analyses.
Anomaly Detection of Livian Citations. For anomaly detection, we used the scikit-learn implementation of a one-class SVM with a nonlinear (radial basis function) kernel and hyperparameters set to $\gamma = \frac{1}{5\sigma}$ or $\gamma = \frac{5}{\sigma}$ (for the full and reduced feature sets, respectively) and $\nu = \frac{1}{7}$ (T1). As described in the text, experiments were performed on randomly aggregated bins constructed from the texts analyzed. The bin size was determined empirically (SI Appendix, Fig. 3B).

ACKNOWLEDGMENTS. We thank Sarah Heiter for assistance with the error analysis of stylometric features and Krithika Iyer for help with natural language processing. We also thank Neil Coffee, Joe Farrell, Stephen Hinds, Dan Rockmore, and Ariane Schwartz for comments on the manuscript. This work was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org), an interdiscipliary project codirected 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 National Endowment for the Humanities Digital Humanities Start-Up Grant HD-248410-16. J.P.D. was supported by National Science Foundation Graduate Research Fellowship Grant DGE1144152, and P.C. was supported by an American Council of Learned Societies Digital Innovation Fellowship.
Quantitative criticism of literary relationships

Joseph P. Dexter,1,2 Theodore Katz,1 Nilesh Tripuraneni,1 Tathagata Dasgupta,1 Ajay Kannan, James A. Brofos, Jorge A. Bonilla Lopez, Lea A. Schroeder, Adriana Casarez, Maxim Rabinovich, Ayelet Haimson Lushkov, and Pramit Chaudhuri2

SUPPORTING INFORMATION APPENDIX

SI Appendix, Text

This text has three primary objectives. It discusses validation of the heuristics used for computation of some of the stylometric features (SI Appendix, Tables S1-S3), it provides a more detailed literary critical interpretation of the Seneca data (Fig. 2; SI Appendix, Figs. S1-S6), and it describes the full set of features used for analysis of Livian citations (SI Appendix, Table S4). It should be read in conjunction with the Results section of the main paper.

Error analysis of enjambment calculations. The computational identification of enjambments relied on punctuation. As described in the Materials and Methods section of the main paper, we counted any sense-pause (including commas) that occurred after the first word of a line as an enjambment unless there was also a sense-pause at the end of the previous line. However, punctuation after the first word in a verse line occasionally is used not to mark a sense-pause of any literary significance, but rather to set off a subsequent address to a named individual or entity (in grammatical terms, the name typically appears in the vocative case). We manually tabulated enjambments in two sample plays, Seneca's Phoenissae and Correr's Procne, and compared the results with the computational tallies. We found no instances of false negatives (i.e., true enjambments missed by the punctuation counting procedure) and a small number of false positives, all but one of which involved a vocative at the beginning of the line. We counted 27 true enjambments and three false positives for the Phoenissae, and 89 true enjambments and three false positives for the Procne. As such, the precision of the enjambment heuristic is 0.9 and the recall 1.0 for the Phoenissae; for the Procne, 0.97 and 1.0. Sentences containing misidentified enjambments are listed in SI Appendix, Tables S1 (Phoenissae) and S2 (Procne).

Enjambment in Correr's Procne. There are numerous examples of Correr's sensitivity to the attention-grabbing effects made possible by enjambment. In an address to the god Mars, for instance, the character Tereus refers to himself in an enjambed line: inclitum cernis, pater, / gnatum. (“you, father, behold your famous son,” Procne 142-143). Tereus thus draws attention to his divine birth and his relationship to Mars through the enjambment, which places emphasis on the word “son” (gnatum) occurring immediately after “father” (pater) and yet on the next line, marked by a firm pause (the period following gnatum). The arrangement of words makes adjacent and yet separates two familial terms that intuitively belong together in a way that cannot easily be replicated in English translation. Correr’s interest in the relationship between Tereus and Mars, highlighted here in the disposition of the words “father” and “son,” is corroborated by the preface to the play, which explicitly mentions the mythical genealogy linking the two figures.

The striking frequency of enjambment in the Procne compared with Senecan and pseudo-Senecan tragedy of the classical period may point to further, and necessarily more speculative, literary critical hypotheses. Both of Correr’s classical models, Ovid’s account of the myth in the Metamorphoses and especially Seneca’s Thyestes, are explicitly concerned with the idea of surpassing one’s predecessors and of excessiveness in general. It is possible, then, that the preponderance of enjambment in the Procne reflects Correr’s youthful exuberance to outdo his classical forebears in the context of a play that itself thematizes one-upmanship. On this view, Correr’s frequent use of enjambment has semantic as well as stylistic value: through its repeated deployment, the technique evokes the idea of exceeding a limit (represented by the end of the verse line), which in turn reflects the thematic concerns of the play and its prior tradition. Although it is impossible to prove the interpretation, the example nevertheless illustrates the productive combination of quantitative and literary critical approaches. The rapid computational calculation of a standard poetic feature such as enjambment can lead directly to the generation of interesting, albeit speculative, literary critical hypotheses.

Error analysis of relative clause calculations. Latin relative pronouns and interrogative pronouns/adjectives/adverbs have very similar forms. For instance, quem can mean either “whom” (relative pronoun), “whom?” (interrogative pronoun), or “which [person or thing]?” (interrogative adjective) depending on the syntax of the sentence. Our aim was to investigate complex subordination of sentences (indicated by relative pronouns) as a marker of authorial style. This goal entailed computationally counting instances of relative pronouns, but not interrogative pronouns or adjectives, without recourse to semantic parsing. Our approach, described in the Materials and Methods section of the main paper, was to exclude all direct interrogative sentences (i.e., those ending in a question mark), since interrogative sentences are much more likely than non-interrogative sentences to contain an interrogative pronoun that could be misidentified as a relative pronoun. We performed a manual error analysis of our relative pronoun counts using a sample corpus that consisted of two tragedies (Phoenissae, Octavia) and a quarter of one book of Livy (22.1-15).

1J.P.D., T.K., N.T., and T.D. contributed equally to this work.
2To whom correspondence may be addressed. Email: jdexter@fas.harvard.edu or pramit.chaudhuri@austin.utexas.edu.
We first checked indirect interrogative sentences (questions reported by the author or a speaker rather than being posed directly, which therefore do not end in a question mark and were not excluded) for instances of interrogative pronouns and adjectives (i.e., false positives). Our manual tabulation found no instances of an interrogative pronoun or adjective within an indirect question in the *Phoenissae* and *Octavia*, and only two instances in the sample of Livy (SI Appendix, Table S3). We then checked for instances of relative pronouns within direct interrogative sentences (i.e., false negatives). The number of false negatives exceeds the number of false positives, but remains low compared with the total number of relative clauses in non-interrogative sentences (SI Appendix, Table S3). As reported in SI Appendix, Table S3, the precision for our heuristic ranged from 0.97 to 1.0 depending on the text examined, and the recall from 0.77 to 0.88. The analysis of the sample texts therefore suggests that the method is sufficient to support our inferences regarding syntactical style in Seneca, Livy, and other Latin authors.

Instances of the adverb *quam* (typically meaning “than” in comparisons or “how” in questions) or of the conjunction *quod* (meaning “because”) are also likely to have been miscounted as an identical form of the relative pronoun. However, such uses are considerably less frequent than the relative pronoun and hence are unlikely to have a substantial impact on the calculated relative clause frequencies.

**Diction, style, and theme in the Octavia.** As described in the main text, our analysis of functional n-grams in the *Octavia* identified two words both often frequent in and thematically important for the play, *noster* (“our”) and *tristis* (“sad,” “stern”). The main objectives of this supplementary discussion are to cite additional literary evidence in support of our analysis, and to elaborate on the implications of our findings for understanding the themes of the drama.

First- and second-person possessive pronouns (*meus*, “my”; *tuus*, “your”) are unusually common in the *Octavia*. A longstanding argument explains the prevalence of such words in terms of versification and compositional style rather than semantic significance (1). On this view, the poet takes over a reasonably common Ovidian and Senecan disyllabic line-ending and uses it excessively. This habit contributes to a more general critique of the competent though not outstanding abilities of the poet, who is able to follow Senecan style but falls short of his exemplar’s level.

The first-person plural possessive *noster* (“our”), already highlighted as an important term using our functional n-gram analysis, is not deployed by the poet in the same way as *meus, tuus,* and other disyllabic possessives (e.g., *suus,* “his/her/its own”). The overwhelming majority of instances of the latter words and their grammatical inflections appear at line-end (*meus*: 44 line-end / 10 mid-line, *tuus*: 30 / 12, *suus*: 32 / 6). In marked contrast, *noster,* which differs prosodically from *meus* and *tuus,* appears far more commonly mid-line, with almost no line-end examples (3 line-end / 39 mid-line). In other words, whatever motivates the poet to use *noster* with great frequency, it is not the same habit of versification that plausibly underlies the placement of other possessives.

Even if a large proportion of the possessive pronouns are best explained as the product of the poet’s versifying tendencies, their collective prevalence bears on the themes of the drama. The plot of the *Octavia* concerns the divorce and exile of the emperor Nero’s wife (the eponymous Octavia), Nero’s marriage to his mistress Poppea, and the tyrannical excesses of his character. On a literary analysis, possessive pronouns - especially first-person (*noster, meus*) and second-person (*tuus*) pronouns - are directly connotative of ownership and suggest of a personal perspective on events. The *Octavia* is a play in which rival claims to possession are perhaps more central, and are certainly more numerous, than in other Senecan tragedies: the first wife lies a prior and more important question: whether these entities should be seen as belonging to one person or another, or even to a group. This question of ownership drives the struggle between the imperial household and, at a larger scale, between the emperor himself, the city, or even the age defined by Nero’s reign. In some cases the word is used as a genuine plural (e.g., by Nero). But beyond such linguistic parsing of *noster* lies a prior and more important question: whether these entities should be seen as belonging to one person or another, or even to a group. This question of ownership drives the struggle between members of the imperial household and, at a larger scale, between tyrant and people. Nero was notorious for treating (and mistreating) as his own what should belong to others or to a wider constituency (cf. Tacitus, *Annales* 15.45.1). This attitude is precisely characteristic of tyranny, and the critique of it is highly appropriate subject matter for a follower of Seneca writing some years in the wake of Nero’s fall.

Although traditional scholarship attributes the frequency of possessives to the poet’s crude versification, a combination of n-gram analysis (which highlighted *noster*) and philological study (which highlighted several possessive pronouns as a class) led to alternative hypotheses about the importance of such words. These hypotheses were in turn corroborated and fleshed out in a qualitative fashion using the techniques of literary criticism. The author of the *Octavia* may have been a more formulaic poet than Ovid or Seneca, but a more charitable interpretation of his diction is enabled by the use of quantitative analysis applied in tandem with traditional critical practices.

Our attention to possessive pronouns also has a bearing on interpretation of the adjective *tristis* (“sad” or “stern”), the other word besides *noster* highlighted by the n-gram analysis as being especially enriched in the *Octavia*. Based on the n-gram analysis alone, we postulated that the word’s frequency might create a mood of melancholy, lament, or suffering. That notion appears to find orthogonal support from other aspects of the play’s diction. In surveying Octavia’s uses of *meus,* we observe that many instances refer to her *fortuna* (“fortune”), *casus* (“misfortune”), *mala* (“evils”), *luctus* (“grief”), and *fata* (“fate”). These
moments of unhappy self-reflection bolster our claim about the heightened mood of lament due to the frequent appearances of *tristis*. These various expressions attribute an unusually pessimistic cast to the *Octavia*, even in comparison to a Senecan corpus generally characterized by harshness and gloom.

**Phonetic clustering in the Phoenissae.** Three examples of clusters of “ente” four-grams in the *Phoenissae* illustrate the potential literary significance of this anomalous feature within the Senecan corpus.

*Significant repetitions need not be adjacent.* “Ente” clusters at one- or two-line intervals are especially enriched in the *Phoenissae* compared with the rest of the corpus. It may seem counterintuitive, especially for readers accustomed to poetry characterized by rhyming endings of successive or alternating verse lines (as in much English poetry), that a writer might exploit echoes of sound at greater intervals. *Phoen*. 314-319, which contains a triple repetition of the phrase *iubente te* (“if you give the order”) at the beginning of the verse line, illustrates Seneca’s exploitation of sound echoes both in adjacent lines (318-319) and at greater intervals (314): *iubente te . . . / iubente te, praebet alitibus licer; / iubente te, vel vivet* (“if you give the order . . . / if you give the order, he will offer his liver to the birds; if you give the order, he will even live”). Although 318-319 contain an adjacent repetition, the first instance of *iubente te* occurs several lines earlier at 314. The effect of the word arrangement is to shorten the period of repetition, first felt at 318 as a distant echo of the initial phrase four lines earlier, only to become closer and more emphatic with the third occurrence in the immediately following line, which is the climax and culmination of Oedipus’ speech.

*Significant repetitions need not be restricted to whole words.* Perhaps the most striking instance of a repetition of “ente” occurs when Jocasta urges her exiled son Polynices to put down his weapons and end the siege of his home city, Thebes. In the context of a play about the effects of an incestuous marriage, a play that literary critics have often identified as sexually suggestive, Jocasta uses perhaps the most jarring innuendo in Latin literature: *clauder vagina impium / ensen, et trementem iamque cupientem excuti / hastam solo defige* (“Sheathe your impious sword in its scabbard, and plant your trembling spear, which already desires to be cast down, in the ground.” *Phoen*. 467-469) (2). The language of sheathes, weapons, and desire leaves almost no room for ambiguity, and in this already erotically charged context it may even be that the audience is supposed to hear in the sound of the word *trementem* an allusion to the Latin word for penis, *mentula* (3). With specific regard to the repetition of “ente,” the jingle of participle endings would here seem to draw further emphasis to the psychological push and pull (“trembling” and “desiring”) characteristic of this most Freudian of dramas.

*Significant repetitions can be both non-adjacent and not restricted to whole words.* Our third and final example, though less spectacular than the previous one, best encapsulates the interest of the four-gram data (*Phoen*. 451-454):

```latex
\begin{align*}
\text{error invitos adhuc} \\
\text{feci nocentes omne Fortunae fuit} \\
\text{peccantis in nos crimen: hoc primum nefas} \\
\text{inter scientes geritur.}
\end{align*}
```

Error has made me, though unwilling, nonetheless guilty: the crime was all Fortune’s, doing us wrong: this is the first sin committed knowingly.

Here we see non-adjacent clustering of non-identical words that share the same morphological ending. The meaning of the clauses is contrasted but the words themselves are not antonyms, as are *nolentem* (“unwilling”) and *cupientem* (“desiring”) at 98-100. It is in part the similar sound of the two words *nocentes* and *scientes*, perhaps augmented by *peccantis*, which reinforces the comparison, and ultimately the opposition, between the two clauses. Here is a simple yet effective instance of local sound repetition - identified computationally - contributing to the structure and semantics of a passage.

**Computation of stylistic features.** We computed a set of 25 Latin stylometric features for use in the anomaly detection experiments, which was subsequently narrowed to a reduced set of five features. All features are continuous, were computed without use of syntactic parsing, and fall into five broad categories (*SI Appendix, Table S4*). 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, such as *Allen and Greenough’s New Latin Grammar* (freely available through the Perseus Project at [http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0001&redirect=true](http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0001&redirect=true)).

Some features comprised whole words, others comprised 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 genit*, *ut educati*, *ut cogniti essent*, *nec Turnus rex Rutulorum*). When counting morphological forms such as superlative endings, however, we computed all instances of the relevant n-gram within a word (e.g., *opulentissima*, where the n-gram -issim- is common to all standard superlative endings). All frequencies in the feature set are per-word.

We selected a diverse range of grammatical and syntactical categories to increase the chance of capturing stylistic patterns of different kinds. Although some features could be calculated with perfect accuracy (e.g., counts of n-grams), without the aid of syntactic parsing other features could only be approximated using heuristics. Error analysis was performed for a small
sample of these features (SI Appendix, Table S3). In general, the accuracy or comprehensiveness of the feature counts is not uniform, and some features were chosen with the understanding that only a small subset of instances were being counted (e.g., gerunds and gerundives).

Conjunctions:

- Conjunctions were computed by counting all instances of *et, -que, atque, ac, neque, aut, vel, at, autem, sed, tamen, postquam.*
- 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.

Subordinate clauses:

- Conditional clauses were computed by counting all instances of the words *si, nisi, quodsi.*
- *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, -is, -e, -ibus, -ebus.* The limitations were applied to exclude instances of *cum* as a preposition (which is followed by nouns in the ablative case, several inflected endings of which are listed above).
- *Quin* clauses were computed by counting all instances of *quin.*
- *Antequam* clauses were computed by counting all instances of *antequam.*
- *Priusquam* clauses were computed by counting all instances of *priusquam.*
- *Dum* clauses were computed by counting all instances of *dum.*
- The fraction of non-interrogative sentences containing at least one relative clause was calculated as follows: a sentence was scored as having a relative clause if it was both non-interrogative (i.e., ending with a punctuation mark other than “?”) and had 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 identical morphologically.
- The mean length of relative clauses was calculated by counting the number of characters in relative clauses identified as above.
- The number of relative pronouns per non-interrogative sentence was calculated by dividing the total number of relative pronouns in non-interrogative sentences by the total number of non-interrogative sentences. Interrogative sentences were excluded for the reasons given above.

Miscellaneous:

- (Direct) interrogative sentences were computed by counting all instances of a sentence ending in a question mark.
- Standard superlative adjectives and adverbs were computed by counting all instances of *-issim-* within a word. The method excluded certain common superlatives such as *maximus* or *optimus,* which would be difficult to capture precisely without also incorporating proper names (e.g., Fabius Maximus, Jupiter Optimus Maximus).
- *Ut* clauses (where *ut* is an adverb or a conjunction) were computed by counting all instances of *ut.*
- The limited subset of gerunds and gerundives was computed by counting all instances of *-ndus* and *-ndum.* The restriction was designed to exclude the many verb forms that share the same letter sequence as the characteristic gerundival ending (e.g., *defendo, pendo*), though at the cost of also excluding the majority of the inflected forms of the gerund and gerundive. Erroneous inclusion of adjectives of the form *blandus* were assumed not to vitiate the count.
- The mean length of sentences was calculated by counting the number of characters in sentences ending in a “.” “?” or “!” and computing the mean. We excluded from the count any periods occurring after a single standalone character, since such instances typically indicate an abbreviation of a proper name rather than a sentence-end.
- Sentence length variance was calculated by counting the number of characters in sentences ending in a “.” “?” or “!” and computing the variance. We excluded from the count any periods occurring after a single standalone character for the reason given above.

### SI Appendix, Tables

<table>
<thead>
<tr>
<th>Reference</th>
<th>Misidentified Enjambment</th>
</tr>
</thead>
<tbody>
<tr>
<td>74-75</td>
<td>non deprecor, non hortor, ( \text{extingui cupidis} ) votumque, genitor, maximum mors est tibi?</td>
</tr>
<tr>
<td>232-233</td>
<td>( \text{\ldots et aures ingerunt quicquid mihi donastis, oculi, cur caput tenebris grave} )</td>
</tr>
<tr>
<td>520-521</td>
<td>quantum daturus: ( \text{\textquote{quando pro te desinam} dixi timere?} ) dixit inridens deus:</td>
</tr>
</tbody>
</table>

**Table S1.** Specific instances of misidentified enjambments in Seneca's *Phoenissae.*
<table>
<thead>
<tr>
<th>Reference</th>
<th>Misidentified Enjambment</th>
</tr>
</thead>
<tbody>
<tr>
<td>517-518</td>
<td>Bacchis lampade nos vocat</td>
</tr>
<tr>
<td></td>
<td>Euboe, Oggigie, adveni!</td>
</tr>
<tr>
<td>542-543</td>
<td>Mundus serta decentia</td>
</tr>
<tr>
<td></td>
<td>munus, Bacche, tuum tuit.</td>
</tr>
<tr>
<td>751-752</td>
<td>Discer ex marito denique insigne facinus audere, Progne!</td>
</tr>
</tbody>
</table>

Table S2. Specific instances of misidentified enjambments in Correr’s *Procne*. 
<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Octavia</td>
<td>77</td>
<td>0</td>
<td>14</td>
<td>1.0</td>
<td>0.85</td>
</tr>
<tr>
<td>Phoenissae</td>
<td>43</td>
<td>0</td>
<td>13</td>
<td>1.0</td>
<td>0.77</td>
</tr>
<tr>
<td>Livy 22.1-15</td>
<td>67</td>
<td>2</td>
<td>9</td>
<td>0.97</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table S3. Error analysis of relative clause frequency. The table lists the true positives, false positives, false negatives, precision, and recall for identification of relative clauses in the three sample texts.
Table S4. Full feature set for stylometric analysis of Livian citation. The 25 features are divided into five broad grammatical and syntactical categories.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pronouns</td>
<td>frequency of personal pronouns</td>
</tr>
<tr>
<td></td>
<td>frequency of demonstrative pronouns</td>
</tr>
<tr>
<td></td>
<td>frequency of quidam</td>
</tr>
<tr>
<td></td>
<td>frequency of third-person reflexive pronouns</td>
</tr>
<tr>
<td></td>
<td>frequency of iste</td>
</tr>
<tr>
<td>non-content adjectives</td>
<td>frequency of alius</td>
</tr>
<tr>
<td></td>
<td>frequency of ipse</td>
</tr>
<tr>
<td></td>
<td>frequency of idem</td>
</tr>
<tr>
<td>conjunctions</td>
<td>aggregate frequency of conjunctions</td>
</tr>
<tr>
<td></td>
<td>frequency of atque followed by a consonant</td>
</tr>
<tr>
<td>subordinate clauses</td>
<td>frequency of conditional clauses</td>
</tr>
<tr>
<td></td>
<td>frequency of cum clauses</td>
</tr>
<tr>
<td></td>
<td>frequency of quin clauses</td>
</tr>
<tr>
<td></td>
<td>frequency of antequam clauses</td>
</tr>
<tr>
<td></td>
<td>frequency of priusquam clauses</td>
</tr>
<tr>
<td></td>
<td>frequency of dum clauses</td>
</tr>
<tr>
<td></td>
<td>fraction of sentences containing a relative clause</td>
</tr>
<tr>
<td></td>
<td>mean length of relative clauses</td>
</tr>
<tr>
<td></td>
<td>number of relative clauses per sentence</td>
</tr>
<tr>
<td>miscellaneous</td>
<td>frequency of interrogative sentences</td>
</tr>
<tr>
<td></td>
<td>frequency of superlatives</td>
</tr>
<tr>
<td></td>
<td>frequency of ut clauses</td>
</tr>
<tr>
<td></td>
<td>frequency of selected gerunds and gerundives</td>
</tr>
<tr>
<td></td>
<td>mean sentence length</td>
</tr>
<tr>
<td></td>
<td>variance of sentence length</td>
</tr>
</tbody>
</table>
Fig. S1. Outliers in Senecan stylometric data. Box plots of the data presented in Fig. 2. (A, i-iii) correspond to Fig. 2A. (ii) corresponds to Fig. 2B, and (C, i-iii) correspond to Fig. 2D, i and iv. (C, ii) is for clusters within one line, C, iii for clusters within five lines. The red line denotes the median, the top and bottom of the blue box denote the 25th and 75th percentile, respectively, and the whiskers extend to the furthest non-outlier points. Outliers (black crosses) are defined as $> Q_3 + 1.5 IQR$ or $< Q_1 - 1.5 IQR$, where $Q$ is the quartile and $IQR$ is the interquartile range.
Fig. S2. Statistical analysis of Fitch's proposed groupings. Ratio of intra-line to total sense pauses for putatively early (group A), middle (group B), and late (group C) tragedies. Groupings follow Fitch 1981; sense-pauses were tabulated computationally using Peiper and Richter’s text. At least one group is significantly different; \( p < 0.001 \) by a one-way ANOVA. Pairwise comparisons were made using a post-hoc Tukey HSD test; * \( p < 0.05 \), ** \( p < 0.01 \).
Fig. S3. Sense-pauses in Giardina's Seneca. Ratio of intra-line to total sense-pauses. Statistics for the eight authentic tragedies are reprinted from Fitch 1981. The ratio in the Octavia was determined by manual tabulation using Giardina's text. The dotted line denotes the mean; error bars denote one SD.
Fig. S4. Bigram analysis of the *Octavia* and *Hercules Oetaeus*. Per-character frequencies of the five most common bigrams in (A) *Octavia* and (B) *Hercules Oetaeus* (gray bars). Beige bars show the mean frequency of each n-gram across the 10 tragedies; error bars denote one SD.
Fig. S5. Co-occurrences of “nt” with vowels. (A) Per-character frequency of the 20 combinations of the form “vowel + nt + vowel.” Error bars indicate one SD across the 10 tragedies. (B) Box plot of the data in A.
Fig. S6. Clusters of “vowel + nt + vowel” four-grams. Fraction of instances of “vowel + nt + vowel” four-grams that occur in clusters within each tragedy. The beige bars indicate instances within one line of each other, the gray bars within three.
Fig. S7. Distribution of annalibus in Livy. Frequency of annalibus between the first decade (left) and subsequent (right) books of Livy. In multiple books of Livy the frequency of annalibus was 0 (indicated by the superscripts). *** $p < 0.001$ by a two-tailed unpaired t-test.
Fig. S8. Bin size and classifier performance. Fraction of bins from bulk Livy (dotted line) and the citation database (solid line) classified as Livian for bins of five sentences to 50 sentences using (A) the full set of 25 features and (B) the reduced set of five features.
Fig. S9. Analysis of Livian citations using reduced feature set. Fraction of bins (random aggregates of 20 sentences) classified as Livian from bulk Livian material (left) and from the citation database (right) by a one-class SVM. Results are the mean ± one SD of 35 leave-one-out cross-validation experiments. *** $p < 0.001$ by a two-tailed unpaired t-test.