PNAS PLUS

Quantitative criticism of literary relationships

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

[a]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; [b]The Dalton School, New York, NY 10128; [c]Research Science Institute, Center for Excellence in Education, McClean, VA 22102; [d]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; [e]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom; [f]Department of Computer Science, Dartmouth College, Hanover, NH 03755; [g]Department of Classics, Dartmouth College, Hanover, NH 03755; [h]Austin Independent School District, Austin, TX 78703; [i]Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720; and [j]Department 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 citational 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.

authorship attribution | cultural evolution | intertextuality | machine learning | stylometry

The study of literature relies on mapping interactions between texts. Ancient Greek critics understood the tragedies of Aeschylus in part through their relation to Homeric epic, and ancient Roman commentators interpreted words and phrases in texts by citing parallels in other works. Much of literary criticism today rests on understanding these vast networks of intertextuality, which often have profound consequences for the meaning of both individual texts and larger groupings by genre or period (1). Through quantitative analysis of formal elements and their change over time, the study of intertextuality can shed light on the cultural evolution of literature (2).

A central challenge in the study of intertextuality is its heterogeneous nature. Literary parallels differ widely in both similarity and scope (Fig. 1*A*). The relationship between the associated texts can range from obvious (direct quotation) to extremely subtle (artfully constructed indirect references, often referred to as allusions in literary study). Furthermore, parallels can operate on the level of individual words or phrases, short passages, or entire works and can involve verbal, syntactic, phonetic, or metrical features. As illustrated in Fig. 1*A*, intertexts can be of comparable similarity but very different scope; an adaptation of an entire work, for instance, can be thought of as a collection of many (local) allusions.

In this paper, we focus on the quantitative characterization of intertextual relationships that involve some (but not extensive) 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,

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.

Author contributions: J.P.D., T.K., N.T., T.D., M.R., A.H.L., and P.C. designed research; J.P.D., T.K., N.T., A.K., J.A.B., J.A.B.L., L.A.S., A.C., and P.C. performed research; J.P.D., T.K., N.T., T.D., A.K., J.A.B., J.A.B.L., L.A.S., A.C., A.H.L., and P.C. analyzed data; and J.P.D., A.H.L., and P.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]J.P.D., T.K., N.T., and T.D. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jdexter@fas.harvard.edu or pramit.chaudhuri@austin.utexas.edu.

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

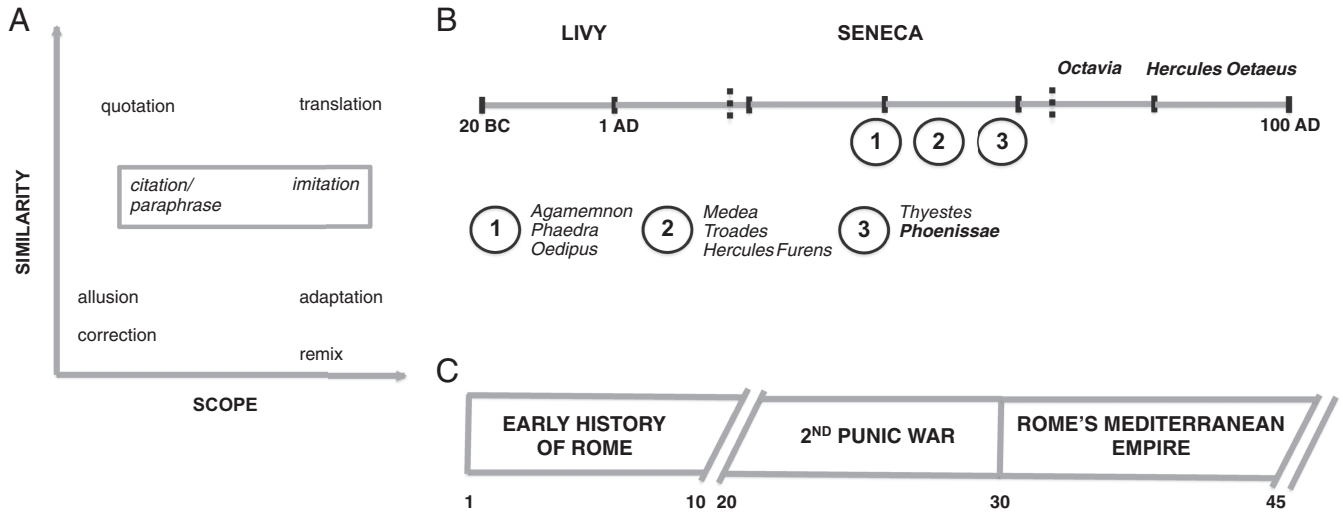

Fig. 1. Intertextuality in Seneca and Livy. (*A*) Categories of intertextuality. Instances of intertextuality can be characterized according to the similarity between the source text and intertext and the scope of the association. For instance, a short quotation (upper left) exhibits higher similarity 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.

exemplified by the work of the Tesserae 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 stylometric 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 the authorship of most of the texts under consideration is not in dispute, these methods allow us to characterize similarities and differences between them in great detail. Our experiments thus provide a richer profile of known intertextual relationships by showing continuity of certain stylometric features within a tradition as well as individual or collective departures from that tradition, and by enabling exploration of the interplay between style and theme.

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. 1*B*) 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 *Procne*, 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. 1*C*) 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 are poorly understood. He cites and quotes both named and unnamed sources, he blends paraphrase and direct quotation, and he freely composes passages in ways likely informed by his reading of sources (45). This complex combination of text reuse has posed particular challenges for literary critics seeking to understand Livy's relationship to his sources. We use an anomaly

detection algorithm trained with a set of 25 stylometric features to classify most material in a curated database of possible citations as differing in style from the rest of Livy. We then apply a similar method to profile the development of Latin prose style across several centuries, which identifies the histories of Caesar and Livy as marking the start of a pronounced shift in literary style that extends across multiple genres.

Results

Quantitative Criticism Identifies Literary Differences Across the Senecan Corpus and Tradition. We profiled a broad range of stylometric features across the whole Senecan and pseudo-Senecan corpus and in Gregorio Correr's *Procne*, 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. 2*A*, *i*). In contrast, total sense pauses were significantly reduced (*Octavia*) or enriched (*Hercules Oetaeus* and *Procne*) in the Senecan-influenced tragedies (Fig. 2*A*, *i* and *SI Appendix*, Fig. S1*A*, *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. 2*A*, *ii* and *SI Appendix*, Fig. S1*A*, *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 (46) to the *Octavia* but likewise used a manual count. In addition to rapidly confirming Fitch's three groupings (46), we also verified Ferri's discovery (48) that the *Octavia* has a relatively low ratio, similar to that expected for an early Senecan tragedy (Fig. 2*A*, *ii*). This result holds across multiple editions of Seneca, despite variations in absolute value of the ratio caused by differences in editorial practice (*SI Appendix*, Fig. S3). The stylistic resemblance of the *Octavia* to early Senecan tragedies is consistent with traditional critical assessments of the play as showing less technical virtuosity than most Senecan drama (48).

Enjambments are a special class of poetic sense pause, in which a sentence or clause "runs over" the end of a line of verse to the first word of the following line. We computationally tabulated enjambments in the tragedies by counting, in lines not starting with a new sentence, every punctuation mark (including commas) immediately after the first word. Counting punctuation is an effective heuristic for the identification of enjambments; for Correr's *Procne*, the precision was 0.97, and the recall was 1.0 (details are in *SI Appendix*, *Text* and Tables S1 and S2). Our analysis revealed a substantial (approximately threefold) enrichment of enjambments in Correr's *Procne* above any Senecan or classical pseudo-Senecan text (Fig. 2*A*, *iii* and *SI Appendix*, Fig. S1*A*, *iii*). As noted above, flexibility in the shape of the verse is typically considered as a mark of skillful poetic composition. 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 *Procne* 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 *Procne* 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. 2*B* and *SI Appendix*, Fig. S1*B*), 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).

Phonetic and Thematic Analyses of the *Octavia* and the *Phoenissae*. 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.

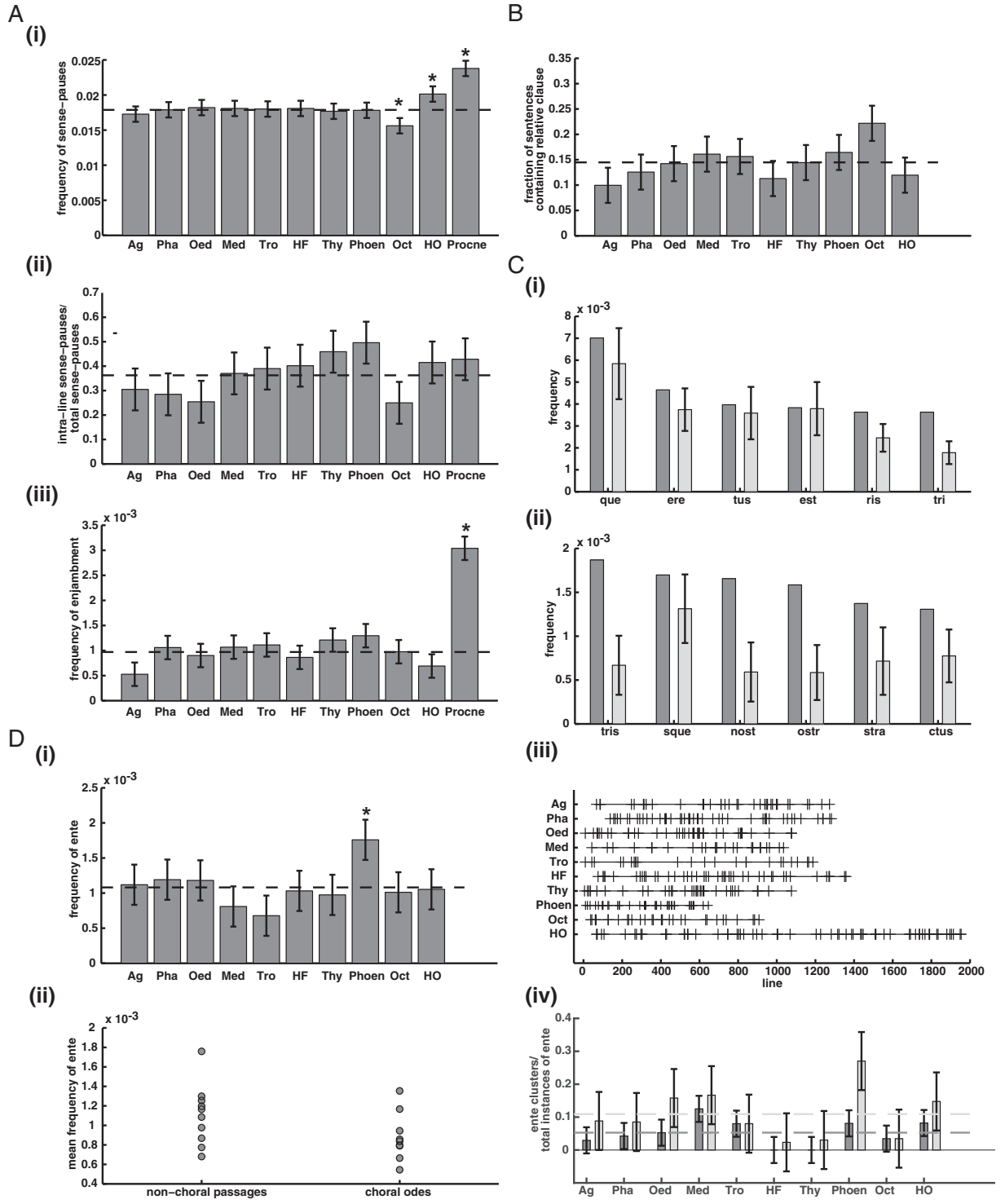

Fig. 2. Quantitative comparison of Senecan and pseudo-Senecan literary style. (*A*, *i*) Total sense pauses in each tragedy. (*A*, *ii*) Ratio of intraline to total sense pauses. (*A*, *iii*) Frequency of enjambment. (*B*) Fraction of noninterrogative sentences containing at least one relative clause. 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 Oetaeus* 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. 2*C*, *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. 2*C*, *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 (nostra 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 (50). 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. 2*D*, *i* and *SI Appendix*, Fig. S1*C*, *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. 2*D*, *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. 2*D*, *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. 2*D*, *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 (unwilling) 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 Seneca'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. 1*A*) 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. 3*A*, *i*). We also examined the distribution of citations across Livy (Fig. 3*A*, *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. 3*B*). 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, Yilmazel 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 "Livian." 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 results have a shared implication for Livy specialists—that critical attention should focus less on the question of whether Livy quoted Antias and more on the question of the potential stylistic irregularities in the first account within the narrative.

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 "Livianness" 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

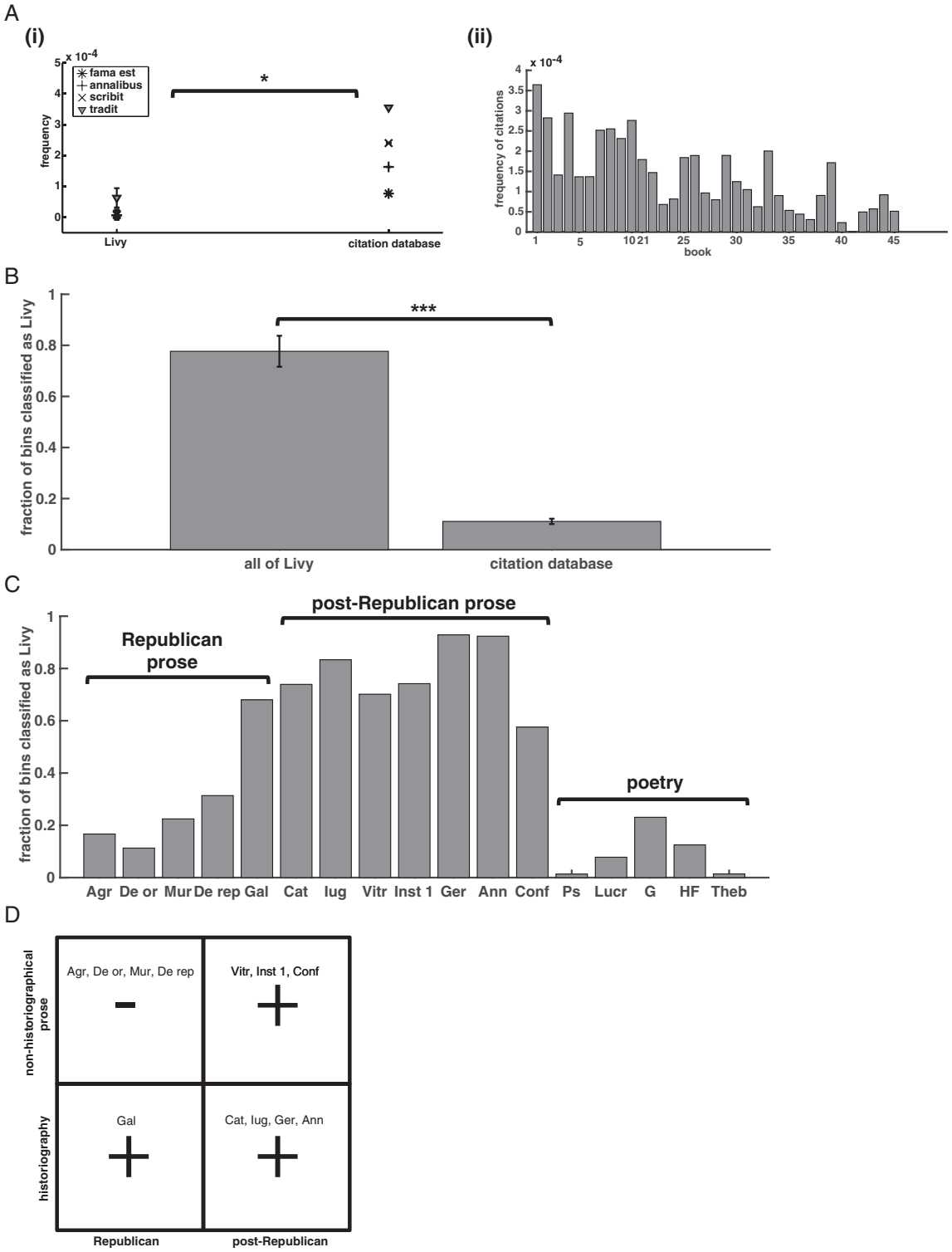

Fig. 3. Anomaly detection differentiates cited material from the rest of Livy. (*A*, *i*) Comparison of the frequency of four "signal words" indicating potential instances of citation (fama est, annalibus, scribit, and tradit) between all of Livy (left) and the citation database (right). *$p < 0.05$ 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*; Vitr, 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 Ciceronian 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 1*) and technical (*De architectura*) 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. 3*D*). 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 citational 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 stylometric 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 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.

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. To that end, we seek to use quantitative data to understand literary relationships and literary interpretation to suggest quantitative 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 *Procne*, 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. S3 were determined manually using Giardina's 1966 text (70). All texts used in the comparison of Latin literary style reported in Fig. 3*C* 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/qcrit. 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 [., ?, !, ;, :, (,), -, ', ', ", and "]. Enjambments were identified by noting instances of punctuation (including commas) that occurred after the first word of a line not immediately preceded by an end-line sense pause. A sentence was scored as having a relative clause if it was both noninterrogative (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). We performed a manual error analysis of the procedures for enjambment and relative clause counting, which is discussed in *SI Appendix, Text* and Tables S1–S3).

For analysis of Livian citations, we considered a set of 25 stylometric features divided into five broad categories: pronouns, noncontent adjectives, conjunctions, subordinate clauses, and miscellaneous. The feature set is listed in *SI Appendix*, Table S4, and the methods used for calculating the features are described in *SI Appendix, Text*.

Assembly of Database of Possible Livian Citations. The database of Livian citations was constructed previously by one of the authors (A.H.L.). The method used to compile the database involved reading the entirety of Livy's history in English translation and noting all passages in which Livy names a source or uses citational language. Manual checks of portions of the Latin text found no instances of passages erroneously included. The database contains 439 distinct entries. The final corpus used for our analysis was created computationally by aggregating all passages of Livy mentioned in the database from the XML file of Weissenborn and Müller's text (68).

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 $\gamma = 1/25$ or $\gamma = 1/5$ (for the full and reduced feature sets, respectively) and $\nu = 1/5$ (71). 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. S8).

We trained the one-class SVM on the whole Livian corpus except for Book 1 using the full set of 25 stylometric features. We then classified all bins in the citation database and Book 1 as nonanomalous (Livian) or anomalous (non-Livian). This procedure was repeated 34 times, with one of the other extant books of Livy withheld for cross-validation each time. Fig. 3*B* reports mean fraction of bins classified as Livian over these 35 experiments. We then identified by direct experimentation a reduced set of five stylometric features for which we obtained comparable classifier performance (*SI Appendix*, Fig. S9). This reduced feature set was used for the analysis of Latin prose style reported in Fig. 3*C*.

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 interdisciplinary 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.

1. Kristeva J (1980) Word, dialogue, and novel. *Desire and Language*, ed Roudiez LS (Columbia Univ Press, New York), pp 64–91.
2. Juvan M (2009) *History and Poetics of Intertextuality* (Purdue Univ Press, West Lafayette, IN).
3. Thomas RF (1986) Virgil's Georgics and the art of reference. *Harv Stud Class Philol* 90:171–198.
4. Hinds SE (1998) *Allusion and Intertext: Dynamics of Appropriation in Roman Poetry* (Cambridge Univ Press, Cambridge, UK).
5. Edmunds L (2001) *Intertextuality and the Reading of Roman Poetry* (Johns Hopkins Univ Press, Baltimore).
6. Crane G (1996) Building a digital library: The Perseus Project as a case study in the humanities. *Proceedings of the First ACM International Conference on Digital Libraries* (Association for Computing Machinery, New York), Vol 1, pp 3–10.
7. Moretti F (2005) *Graphs, Maps, Trees: Abstract Models for Literary History* (Verso, London).
8. Moretti F (2013) *Distant Reading* (Verso, London).
9. Ramsay S (2011) *Reading Machines: Toward an Algorithmic Criticism* (Univ of Illinois Press, Champaign, IL).
10. Jockers M (2013) *Macroanalysis: Digital Methods and Literary History* (Univ of Illinois Press, Champaign, IL).
11. Long H, So R (2016) Literary pattern recognition: Modernism between close reading and machine learning. *Crit Inq* 42:235–267.
12. Hammond A, Brooke J, Hirst G (2013) A tale of two cultures: Bringing literary analysis and computational linguistics together. *Proceedings of the Second Workshop on Computational Linguistics for Literature* (Association for Computational Linguistics, Stroudsburg, PA), pp 1–8.
13. Underwood T (2014) Theorizing research practices we forgot to theorize twenty years ago. *Representations* 127:64–72.
14. Jockers ML, Mimno D (2013) Significant themes in 19th-century literature. *Poetics* 41:750–769.
15. Piper A (2015) Novel devotions: Conversional reading, computational modeling, and the modern novel. *New Lit Hist* 46:63–98.
16. Bamman D, Crane G (2008) The logic and discovery of textual allusion. *Proceedings of the 2008 LREC Workshop on Language Technology for Cultural Heritage Data*. Available at citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.366.8213.
17. Coffee N, Koenig JP, Porrnima S, Ossewaarde R, Jacobson S (2012) Intertextuality in the digital age. *Trans Am Philol Assoc* 142:383–422.
18. Coffee N, Koenig JP, Porrnima S, Ossewaarde R, Jacobson S (2013) The Tesserae Project: Intertextual analysis of Latin poetry. *Lit Linguist Comput* 28:221–228.
19. Bernstein N, Gervais K, Lin W (2015) Comparative rates of text reuse in classical Latin hexameter poetry. *Digit Humanit Q* 9(3).
20. Coffee N, Bernstein N (2016) Digital methods and classical studies. *Digit Humanit Q* 10(2).
21. Michel JB, et al. (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182.
22. Sieber R, Wellen C, Jin Y (2011) Spatial cyberinfrastructures, ontologies, and the humanities. *Proc Natl Acad Sci USA* 108:5504–5509.
23. Aiden E, Michel JB (2013) *Uncharted: Big Data as a Lens on Human Culture* (Riverhead Books, New York).
24. Lansdall-Welfare T, et al. (2017) Content analysis of 150 years of British periodicals. *Proc Natl Acad Sci USA* 114:E457–E465.
25. Hughes JM, Fotia NJ, Krakauer DC, Rockmore DN (2012) Quantitative patterns of stylistic influence in the evolution of literature. *Proc Natl Acad Sci USA* 109:7682–7686.
26. Lyu S, Rockmore D, Farid H (2004) A digital technique for art authentication. *Proc Natl Acad Sci USA* 101:17006–17010.
27. Koppel M, Schler J, Argamon S (2009) Computational methods in authorship attribution. *J Am Soc Inf Sci Technol* 60:9–26.
28. Stamatatos E (2009) A survey of modern authorship attribution methods. *J Am Soc Inf Sci Technol* 60:538–556.
29. Hughes J, Graham D, Rockmore D (2010) Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proc Natl Acad Sci USA* 107:1279–1283.
30. Stamou C (2008) Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Lit Linguist Comput* 23:181–199.
31. Mosteller F, Wallace DL (1964) *Inference and Disputed Authorship: The Federalist* (Addison-Wesley, Reading, MA).
32. Holmes DI, Robertson M, Paez R (2001) Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Comput Humanit* 35:315–331.
33. Vickers B (2004) *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays* (Oxford Univ Press, Oxford).
34. Stover J, Winter Y, Koppel M, Kestemont M (2016) Computational authorship verification method attributes a new work to a major 2nd century African author. *J Assoc Inf Sci Technol* 67:239–242.
35. Hoover DL (2007) Corpus stylistics, stylometry, and the styles of Henry James. *Style* 41:174–203.
36. Hoover DL (2014) Modes of composition in Henry James: Dictation, style, and what Maisie knew. *Henry James Rev* 35:257–277.
37. Forstall C, Jacobson S, Scheirer W (2011) Evidence of intertextuality: Investigating Paul the Deacon's Angustae Vitae. *Lit Linguist Comput* 26:285–296.
38. Bulson E (2014) Ulysses by numbers. *Representations* 127:1–32.
39. Hope J, Witmore M (2010) The hundredth psalm to the tune of 'Green Sleeves': Digital approaches to the language of genre. *Shakespeare Q* 61:357–390.
40. Eder M (2016) A bird's-eye view of early modern Latin: Distant reading, network analysis, and style. *Early Modern Studies After the Digital Turn*, eds Estill L, Jakacki D, Ullyot M (Iter and ACMRS, Toronto), pp 63–89.
41. Altmann E, Cristadoro G, Esposti MD (2012) On the origin of long-range correlations in texts. *Proc Natl Acad Sci USA* 109:11582–11587.
42. Clement T, Tcheng D, Auvil L, Capitanu B, Monroe M (2013) Sounding for meaning: Using theories of knowledge representation to analyze aural patterns in texts. *Digit Humanit Q* 7(1).
43. Levitan W (1989) Seneca in Racine. *Yale Fr Stud* 76:185–210.
44. Miola RS (1992) *Shakespeare and Classical Tragedy: The Influence of Seneca* (Clarendon, Oxford).
45. Haimson Lushkov A (2013) Citation and the dynamics of tradition in Livy's AUC. *Histos* 7:21–47.
46. Fitch J (1981) Sense-pauses and relative dating in Seneca, Sophocles and Shakespeare. *Am J Philol* 102:289–307.
47. Dingel J (2009) *Die Relative Datierung der Tragödien Senecas* (De Gruyter, Berlin).
48. Ferri R (2003) *Octavia: A Play Attributed to Seneca* (Cambridge Univ Press, Cambridge, UK).
49. Braund S (2009) *Seneca, De Clementia* (Oxford Univ Press, Oxford).
50. Frank M (1995) *Seneca's Phoenissae: Introduction and Commentary* (Brill, Leiden, The Netherlands).
51. Wills J (1996) *Repetition in Latin Poetry: Figures of Allusion* (Clarendon, Oxford).
52. Fehling D (1989) *Herodotus and His 'Sources': Citation, Invention and Narrative Art* (Francis Cairns, Leeds, UK).

53. Grafton A (1997) *The Footnote: A Curious History* (Harvard Univ Press, Cambridge, MA).
54. Levene DS (2010) *Livy on the Hannibalic War* (Oxford Univ Press, Oxford).
55. Walsh PG (1961) *Livy: His Historical Aims and Methods* (Cambridge Univ Press, Cambridge, UK).
56. Forsythe G (1999) *Livy and Early Rome: A Study in Historical Method and Judgment* (Franz Steiner Verlag, Stuttgart).
57. Oakley S (1997) *A Commentary on Livy Books VI-X. Volume I, Introduction and Book VI* (Clarendon, Oxford).
58. Jain LP, Scheirer WJ, Boult TE (2014) Multi-class open set recognition using probability of inclusion in computer vision. *Proceedings of the ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014*, eds Fleet D, Pajdla T, Schiele B, Tuytelaars T (Springer, Cham, Switzerland), Part III, pp 393–409.
59. Yilmazel O, Symonenko S, Balasubramanian N, Liddy ED (2005) Leveraging one-class SVM and semantic analysis to detect anomalous content. *Intelligence and Security Informatics. ISI 2005*. Lecture Notes in Computer Science, eds Kantor P, Muresan G, Roberts F, Zeng D, Wang F-Y, Chen H, Merkle R (Springer, Berlin), Vol 3495, pp 381–388.
60. Briscoe J (2005) The language and style of the fragmentary republican historians. *Aspects of the Language of Latin Prose*, eds Reinhardt T, Lapidge M, Adams J (Oxford Univ Press, Oxford), pp 53–72.
61. Haimson Lushkov A (2010) Intertextuality and source criticism in the Scipionic trials. *Livy and Intertextuality*, ed Polleichtner W (Wissenschaftlicher Verlag Trier, Trier, Germany), pp 93–133.
62. Briscoe J (2008) *A Commentary on Livy Books 38-40* (Oxford Univ Press, Oxford).
63. Oakley S (2009) Style and language. *Cambridge Companion to Tacitus*, ed Woodman A (Cambridge Univ Press, Cambridge, UK), pp 195–211.
64. Duckworth GE (1969) *Vergil and Classical Hexameter Poetry: A Study in Metrical Variety* (Univ of Michigan Press, Ann Arbor, MI).
65. Cavnar W, Trenkle J (1994) N-gram based text categorization. *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*. Available at citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.53.9367&type=ab.
66. Houvardas J, Stamatatos E (2006) N-gram feature selection for authorship identification. Proceedings of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA), eds Euzennat J, Domingue J (Springer, Berlin), pp 77–86.
67. Peiper R, Richter G (1921) L. *Annaei Senecae Tragoediae* (Teubner, Leipzig, Germany).
68. Weissenborn W, Müller HJ (1880–1911) *Titi Livi ab urbe condita libri* (Weidmann, Berlin).
69. Grund GR (2011) *Humanist Tragedies* (Harvard Univ Press, Cambridge, MA).
70. Giardina GC (1966) *L. Annaei Senecae Tragoediae* (Editrice Compositori, Bologna, Italy).
71. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830.