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 Chaudhuri[2]

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 oneupmanship. 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).

[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.

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 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 Poppaea, 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 suggestive 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 vs. the second, Nero vs. the stepbrother he has murdered, Nero vs. his political advisor Seneca (who appears as a character within the drama), Nero vs. the chorus of Roman people (who favor Octavia), to mention only the largest contentions. In addition, there are multiple struggles over the sites that various parties lay claim to: the city, the household, the bedroom.

The combination of ownership and personal perspective takes on an especially political coloring in several of the phrases in which *noster* appears. Consider the following words used with *noster*, with the speaker or speakers noted in parentheses: *domus* ("household;" Octavia, Nero), *princeps* ("emperor;" Chorus, Octavia), *dux* ("leader;" Chorus), *urbs* ("city;" Nero, Chorus), *saeculum* ("age;" Nero). In each case *noster* is attached to a political or politicized entity, whether the imperial household, 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 the chorus), in other cases as a royal "our" (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, praebebit alitibus iecur; / 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: *claude vagina impium / ensem, 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):

error invitos adhuc
fecit nocentes: omne Fortunae fuit
peccantis in nos crimen: hoc primum nefas
inter scientes geritur.

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 stylometric 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).

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*** *geniti*, ***ut*** *educati*, ***ut*** *cogniti essent*, not *Turnus rex R**ut**ulorum.*) 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.

1. Ferri R (2003) *Octavia: A Play Attributed to Seneca.* (Cambridge Univ Press, Cambridge, UK).
2. Ginsberg L (2015) Don't stand so close to me: Antigone's *pietas* in Seneca's *Phoenissae. Trans Am Philol Assoc* 145:199–230.
3. Adams J (1982) *The Latin Sexual Vocabulary.* (Johns Hopkins Univ Press, Baltimore, MD).

SI Appendix, Tables

Reference	Misidentified Enjambment
74-75	*non deprecor, non hortor, extingui cupis* *votumque, genitor, maximum mors est tibi?*
232-233	*. . . et aures ingerunt quicquid mihi* *donastis, oculi, cur caput tenebris grave*
520-521	*quantum daturus: ‘quando pro te desinam’* *dixi ‘timere?’ dixit inridens deus:*

Table S1. Specific instances of misidentified enjambments in Seneca’s *Phoenissae*.

Reference	Misidentified Enjambment
517-518	*Bacchis lampade nos vocat* *Euboe, Oggigie, adveni!*
542-543	*Mundus serta decentia* *munus, Bacche, tuum tulit.*
751-752	*Disce ex marito denique insigne facinus* *audere, Progne!*

Table S2. Specific instances of misidentified enjambments in Correr's *Procne***.**

	TP	FP	FN	Precision	Recall
Octavia	77	0	14	1.0	0.85
Phoenissae	43	0	13	1.0	0.77
Livy 22.1-15	67	2	9	0.97	0.88

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.

	pronouns
1	frequency of personal pronouns
2	frequency of demonstrative pronouns
3	frequency of *quidam*
4	frequency of third-person reflexive pronouns
5	frequency of *iste*
	non-content adjectives
6	frequency of *alius*
7	frequency of *ipse*
8	frequency of *idem*
	conjunctions
9	aggregate frequency of conjunctions
10	frequency of *atque* followed by a consonant
	subordinate clauses
11	frequency of conditional clauses
12	frequency of *cum* clauses
13	frequency of *quin* clauses
14	frequency of *antequam* clauses
15	frequency of *priusquam* clauses
16	frequency of *dum* clauses
17	fraction of sentences containing a relative clause
18	mean length of relative clauses
19	number of relative clauses per sentence
	miscellaneous
20	frequency of interrogative sentences
21	frequency of superlatives
22	frequency of ut clauses
23	frequency of selected gerunds and gerundives
24	mean sentence length
25	variance of sentence length

Table S4. Full feature set for stylometric analysis of Livian citation. The 25 features are divided into five broad grammatical and syntactical categories.

SI Appendix, Figures

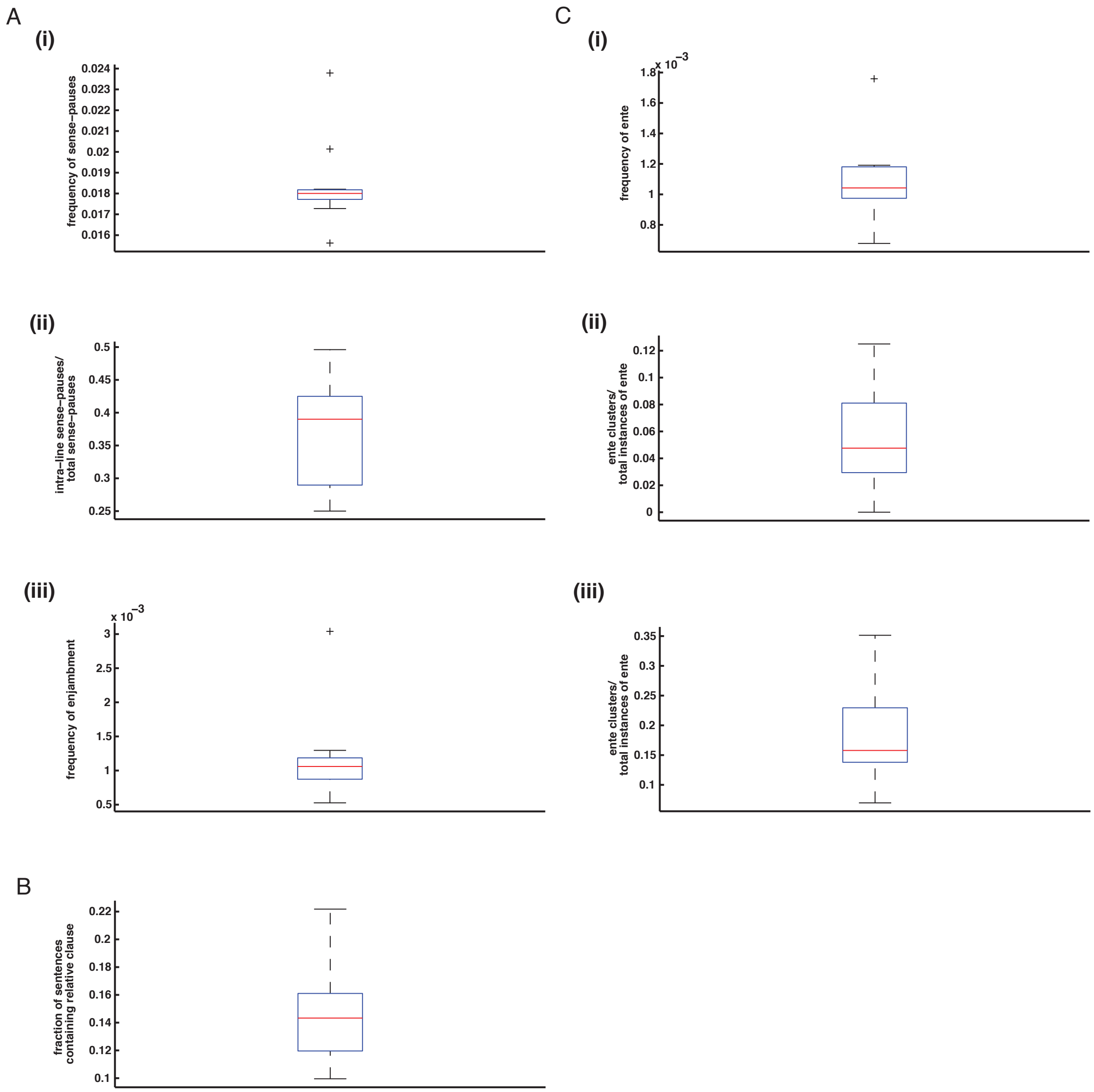

Fig. S1. Outliers in Senecan stylometric data. Box plots of the data presented in Fig. 2. *(A, i-iii)* correspond to Fig. 2*A*, *i-iii*, *(B)* corresponds to Fig. 2*B*, and *(C, i-iii)* correspond to Fig. 2*D*, *i* and *iv*. *C, ii* is for clusters within one line, *C, iii* for clusters within five lines. 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 $>$ Q3 + 1.5IQR or $<$ Q1 − 1.5IQR, where Q is the quartile and IQR is the interquartile range.

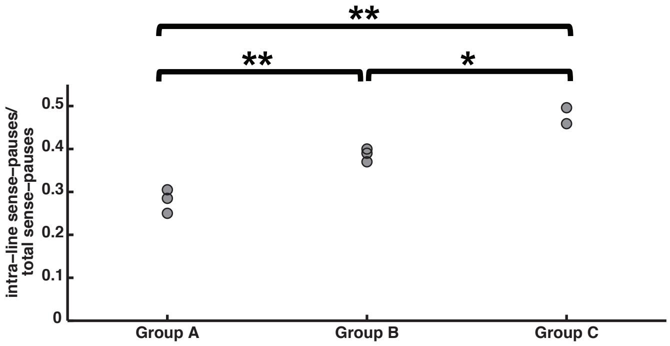

Fig. S2. Statistical analysis of Fitch's proposed groupings. Ratio of intra-line to total sense pauses for putatively early (group A), middle (group B), and late (group C) tragedies. Groupings follow Fitch 1981; sense-pauses were tabulated computationally using Peiper and Richter's text. At least one group is significantly different; $p < 0.001$ by a one-way ANOVA. Pairwise comparisons were made using a post-hoc Tukey HSD test; * $p < 0.05$, ** $p < 0.01$.

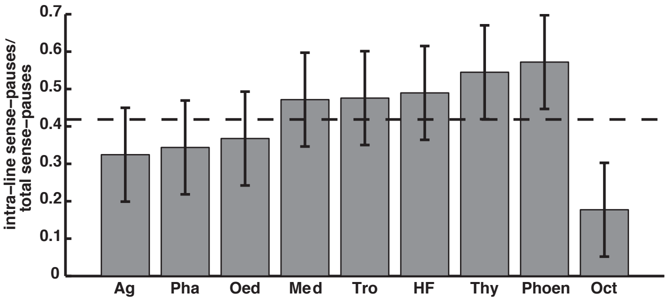

Fig. S3. Sense-pauses in Giardina’s Seneca. Ratio of intra-line to total sense-pauses. Statistics for the eight authentic tragedies are reprinted from Fitch 1981. The ratio in the *Octavia* was determined by manual tabulation using Giardina’s text. The dotted line denotes the mean; error bars denote one SD.

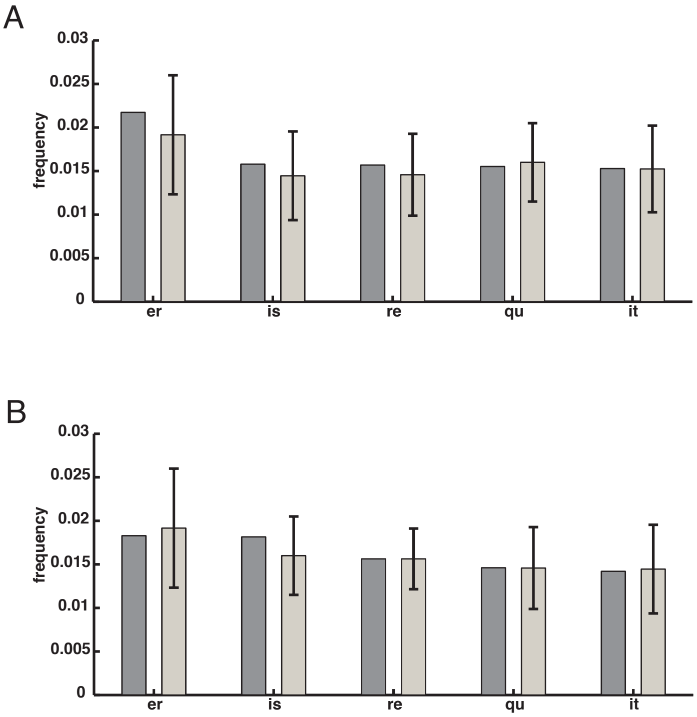

Fig. S4. Bigram analysis of the *Octavia* and *Hercules Oetaeus*. Per-character frequencies of the five most common bigrams in (*A*) *Octavia* and (*B*) *Hercules Oetaeus* (gray bars). Beige bars show the mean frequency of each n-gram across the 10 tragedies; error bars denote one SD.

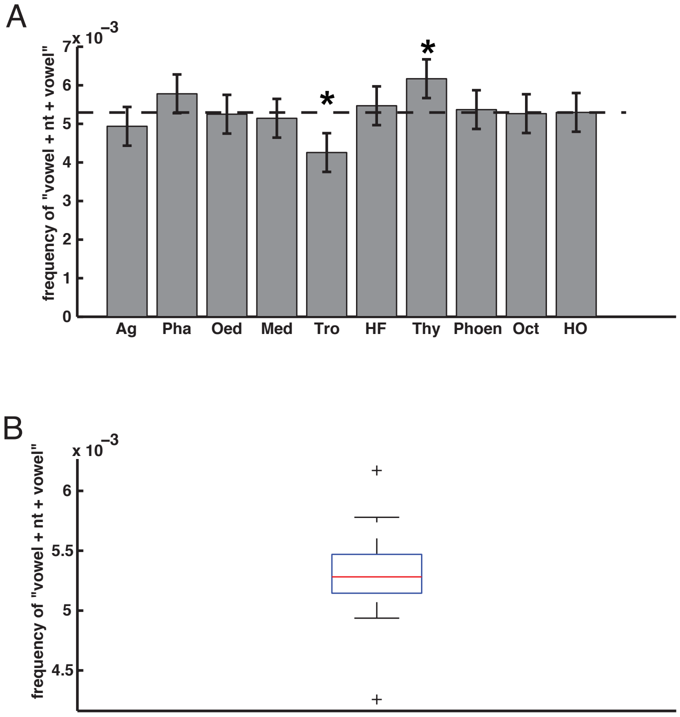

Fig. S5. Co-occurrences of “nt” with vowels. (*A*) Per-character frequency of the 20 combinations of the form “vowel + nt + vowel.” Error bars indicate one SD across the 10 tragedies. (*B*) Box plot of the data in *A*.

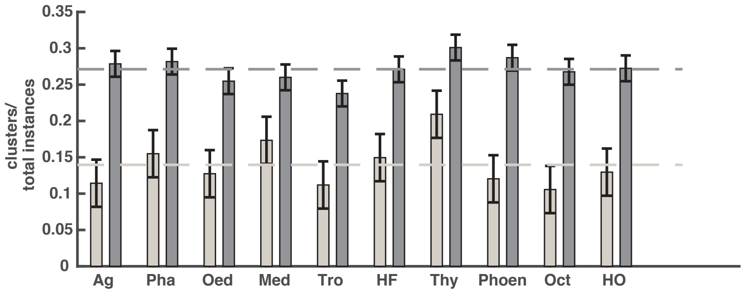

Fig. S6. Clusters of “vowel + nt + vowel” four-grams. Fraction of instances of “vowel + nt + vowel” four-grams that occur in clusters within each tragedy. The beige bars indicate instances within one line of each other, the gray bars within three.

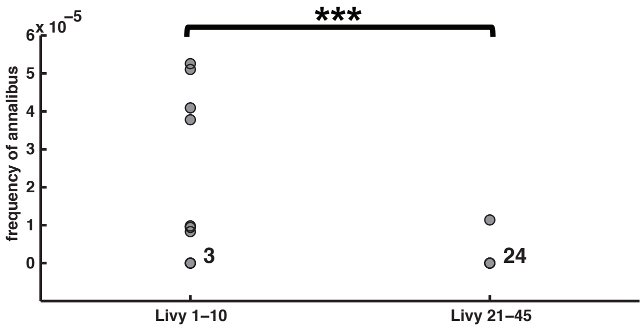

Fig. S7. Distribution of *annalibus* in Livy. Frequency of *annalibus* between the first decade (left) and subsequent (right) books of Livy. In multiple books of Livy the frequency of *annalibus* was 0 (indicated by the superscripts). *** $p < 0.001$ by a two-tailed unpaired t-test.

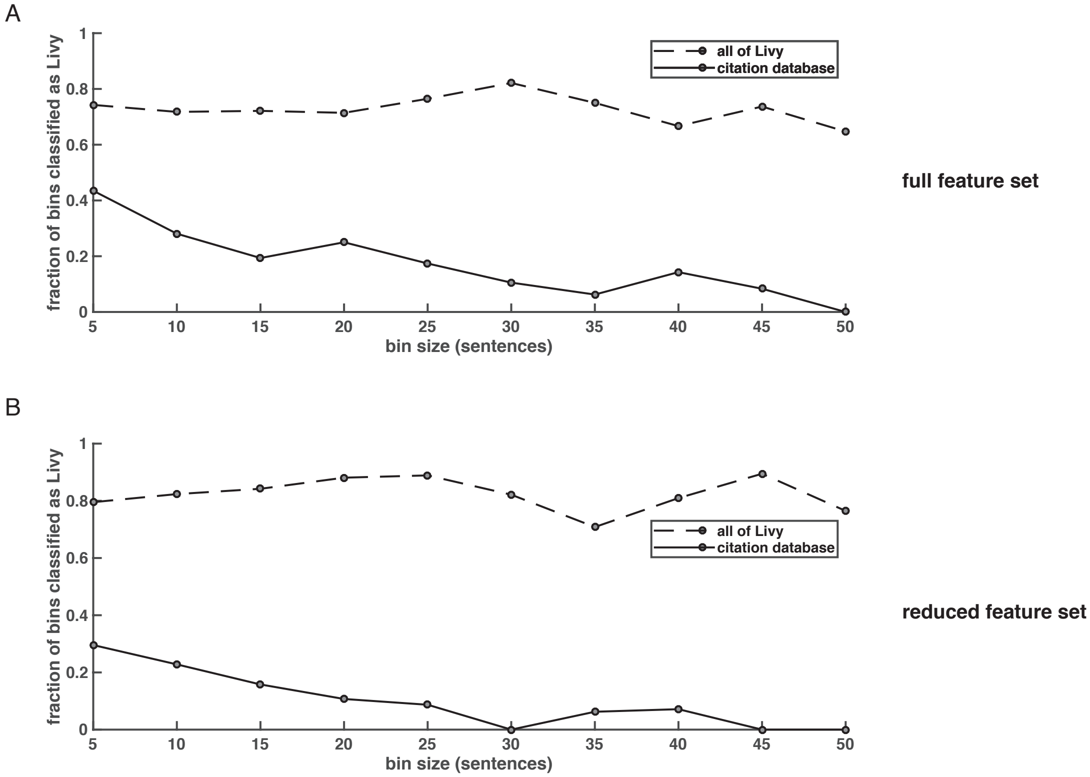

Fig. S8. Bin size and classifier performance. Fraction of bins from bulk Livy (dotted line) and the citation database (solid line) classified as Livian for bins of five sentences to 50 sentences using (*A*) the full set of 25 features and (*B*) the reduced set of five features.

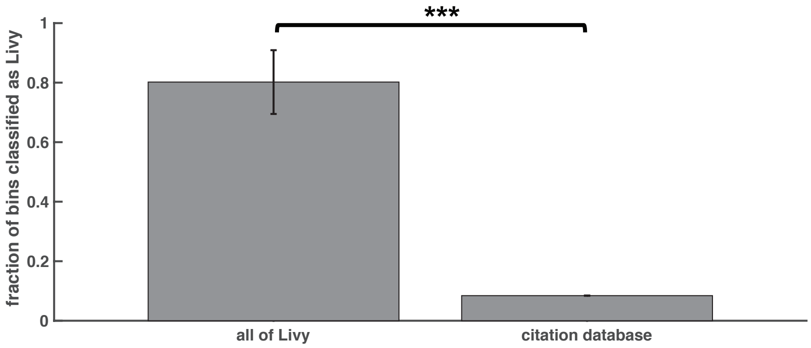

Fig. S9. Analysis of Livian citations using reduced feature set. Fraction of bins (random aggregates of 20 sentences) classified as Livian from bulk Livian material (left) and from the citation database (right) by a one-class SVM. Results are the mean ± one SD of 35 leave-one-out cross-validation experiments. *** $p < 0.001$ by a two-tailed unpaired t-test.