SOME STYLOMETRIC REMARKS ON OVID'S *Heroides* AND THE *Epistula Sapphus*

PREPRINT, COMPILED SEPTEMBER 4, 2025

Ben Nagy

Institute of Polish Language, Polish Academy of Sciences (IJP PAN)

ABSTRACT

This article aims to contribute to two well-worn areas of debate in classical Latin philology, relating to Ovid's *Heroides*. The first is the question of the authenticity (and, to a lesser extent the correct position) of the letter placed fifteenth by almost every editor—the so-called *Epistula Sapphus* (henceforth *ES*). The secondary question, although perhaps now less fervently debated, is the authenticity of the 'Double *Heroides*', placed by those who accept them as letters 16–21. I employ a variety of methods drawn from the domain of computational stylometry to consider the poetics and the lexico-grammatical features of these elegiac poems in the broader context of a corpus of 'shorter' (from 20 to 546 lines) elegiac works from five authors (266 poems in all) comprising more or less all of the non-fragmentary classical corpus. Based on a variety of techniques, every measure gives clear indication that the poetic style of the *Heroides* is Ovidian, but distinctive; they can be accurately isolated from Ovid more broadly. The Single and Double *Heroides* split into two clear groups, with the *ES* grouped consistently with the single letters. Furthermore, by comparing the style of the letters with the 'early' (although there are complications in this label) works of the *Amores* and the late works of the *Ex Ponto*, the evidence supports sequential composition—meaning that the *ES* is correctly placed—and, further, supports the growing consensus that the double letters were composed significantly later, in exile.

Keywords: stylometry, Latin, poetics, computational linguistics, authorship attribution

1 INTRODUCTION

In form, Ovid's *Heroides* are a collection of twenty-one elegies of medium length.[1] They are written in epistolary style; the poet ventriloquises fifteen women from mythology, before finishing with three pairs comprising a letter from a 'hero' and then a response (the so-called 'Double' *Heroides*). The sole exception to this pattern is a letter from Sappho (who is not mythological) to Phaon (who probably is). By those who doubt its authenticity, this is known as the *Epistula Sapphus*, but it appears in most modern editions as *Heroides* 15. As works of literature, the *Heroides* are exceptionally important for their presentation of the events of typically male-centred genres, primarily epic and tragedy, from a woman's perspective. While still classified as love elegy (they are about love and are written in elegiac couplets), they reverse the male gaze and allow readers to assume a female point of view. In terms of characterisation, Ovid in the *Heroides* dons distinct foreign personalities, yielding an obvious and useful point of comparison to the *praeceptor amoris* of the *Ars Amatoria*.[2]

From a philological perspective, which is the one taken in this article, the poems raise a number of questions, the most important ones being the composition dates (are they early or late?) and the authenticity of several of them—for not only the *ES*, but all of the Double *Heroides* have been athetised by various scholars.

While these questions have been discussed at length by many excellent literary scholars (a summary appears in the following section), a modern stylometric treatment is overdue, and that is what this article will attempt to provide. Following a reprise of the key issues, and the standard arguments for and against, I present the results of a computational stylometric analysis that examines two broad domains: lexico-grammatical style (comprising word choice, phonetic preferences, and grammatical style) and 'poetic' style (which ignores lexis completely, considering only issues of prosody, rhyme and assonance). Based on a variety of techniques, every measure gives clear indication that the poetic style of the *Heroides* is Ovidian, but distinctive; they can be accurately isolated from Ovid more broadly. The single and double *Heroides* split into two clear groups, with the *ES* grouped consistently with the single letters. Furthermore, by comparing the style of the letters with the 'early' (although there are complications in this label) works of the *Amores* and the late works of the *Ex Ponto*, the evidence supports sequential composition—meaning that the *ES* is correctly placed—and, further, supports the con-

[1] About 190 lines, on average, significantly longer than the average length of the *Amores* (50), *Tristia* (71) or *Ex Ponto* (69), but not 'book length' like the *Fasti* or *Ars Amatoria*.

[2] Cf Catullus 16: *nam castum esse decet pium poetam | ipsum, versiculos nihil necesse est*, 'an upstanding poet should be virtuous themselves, but their verses need not be'.

correspondence: benjamin.nagy@ijp.pan.pl

tention that the double letters were composed significantly later, in exile.

2 THE BONES OF THE EXISTING DEBATE

2.1 The Authorship of the ES

As space is limited, here I present only the bones of the existing debate, aiming mostly to establish some context and offer a few opinions on certain points. For a complete history that is fairly up to date I can recommend no better source than Thorsen (2014, 97–122), which includes a conveniently detailed literary response to Tarrant (1981), whose article is the linchpin of the modern case against the *ES*.

The first thing that must be said is that the transmission of the *ES* is troubled, even in the context of the relatively poor transmission of the *Heroides* in general (Ovid more broadly survives in generally excellent condition). Of the oldest manuscripts, the full text of the *ES* appears only in the thirteenth century *Francofurtanus* (MS Barth. 110), although the letter must have been widely known in the middle ages, since excerpts survive in at least six Medieval *florilegia*. However after 1420 we have a profusion of manuscripts descended from a lost second source. For a detailed conspectus and discussion, see Dörrie (1971, 287ff.) (in Latin) or prefer Tarrant (1983) for concision.

The *ES* was, in 1629, placed fifteenth in the collection by Daniel Heinsius, following the notes of Scaliger.[3] Scaliger believed it genuine (but doubted the Double *Heroides*), as did Heinsius. It was first doubted in 1816 by Francke for mistaken reasons, then by Schneidewin who proposed, but later retracted, the idea that it was a Renaissance forgery. Really, though, the seminal doubts were planted by Lachmann, a towering Latinist, in 1848. His arguments convinced Palmer who, in his 1874 edition, omitted the poem since it was "condemned by Lachmann, and *every scholar possessed of common sense*" (my emphasis). Years of heated debate followed[4] and by 1898 Palmer had reassessed the arguments, so that when concluding the second edition he instructed his successor to "defend as far as possible [its] Ovidian authorship". Although I have glossed over the details, suffice it to say that the points of debate up to this point were both minor and few, and could easily be put down to problems with transmission.

The time has now come to address the elephant in the room. Those disputing the authenticity of the *ES* need to deal with one very good reason to believe it to be genuine—the fact that *Ovid himself tells us he wrote it*. The relevant passage[5] is from *Amores* 2.18:

> aut, quod Penelopes verbis reddatur Ulixi,
> scribimus[...
> ...]
> quodque tenens strictum Dido miserabilis ensem
> dicat et †Aoniae Lesbis amata lyrae†.[6]
>
> Ov. *Am.* 2.18.21–6

> I write what Penelope's words conveyed to Ulysses [... several references to other *Heroides*...] and what poor Dido, holding a drawn sword, might say, or the lover from Lesvos with the Aeonian lyre.

Mostly this problem is handled in one of two ways—either by claiming that Ovid did write a letter from Sappho, but that it was lost, so the present *ES* is a replacement by an interpolator,[7] or alternatively by arguing that the *ES* was never genuine and the lines in *Amores* 2.18 have been tampered with to authenticate it. Tarrant (1981) chooses the latter path, but Courtney's (1997, 163) rebuttal seems unassailable (emphasis mine):

> Suppose we want to declare the Letter of Sappho spurious; we then run up against the difficulty that Ovid himself twice refers to it in *Am.* 2.18 ... We then have to presume that the author of the Letter of Sappho validated his forgery by rewriting two lines of *Am.* 2.18 to introduce mention of it; *how did he then impose his will on the whole textual tradition?*

As for the rest of Tarrant's arguments, I doubt that I could add much to Thorsen's (2014, 105 ff.) point-by-point response and in any case that is not the aim of this article. Tarrant, as is common in Classical scholarship, focuses his attention on *rare events*—a metrical oddity here, an unusual collocation of words there, a turn of phrase that is not attested until Neronian times. In stylometric (and statistical) terms—and this will be discussed in more depth in §3Methodologysection.3—I am more concerned with analysis that considers *common* events, like the prosodic features of every line, or the relative frequencies of thousands of different words. In fact Tarrant (1983, 272) himself seems to acknowledge that further analysis was needed: "... a careful stylistic analysis of the collection has not yet been undertaken and the question therefore remains open" (although here he refers more to the authorship of the Dou-

[3] In **F** it appears before the rest, but the letters in that MS are jumbled, see Dörrie (1971, 288). For the full story, Thorsen (2014, 21–2).

[4] Thorsen (2014, 102 n. 28) finds at least nineteen articles in the intervening two decades!

[5] In fact there are two references to Sappho in *Am.* 2.18; I do not focus on the second because (although I reject this argument) it might be thought to refer one of the works of Sabinus, who was evidently inspired by Ovid's letters to compose some male replies.

[6] The reading here is intractable, but no proposed version does away with the fact that it is a reference to Sappho.

[7] Thus eg McKeown (1998, 398) while acknowledging that it "puts an agonising strain on credibility".

ble *Heroides*, which are discussed next). I do not at all dismiss the approach taken in Tarrant's article (although I might debate the specifics); the point I am making is that computational assistance opens new and different avenues of investigation by allowing us to simultaneously consider thousands of stylistic markers in a way that is simply impossible for humans.

2.2 *The Double* Heroides

Despite his claim to the contrary, Tarrant, in questioning the authenticity of the *ES*, is influenced to some extent by his perception of its quality—"[i]t is my private opinion that the ES is a tedious production containing hardly a moment of wit, elegance, or truth to nature, and that its ascription to Ovid ought never to have been taken seriously" (Tarrant, 1981, 134–5). Indeed readers have often (and perhaps understandably) thought of the Single *Heroides* that "[i]t is difficult to rescue them, especially if read sequentially from the charge of monotony" (Kenney, 1996, 1).[8] For the Double *Heroides*, however, the case is quite different. Few modern critics will deny that their style is quintessentially Ovidian[9] nor that they are, in short, very good.[10] By varying the format, the poet profited from new opportunities for drama, development and characterisation.

So what, then, is the sticking point? There are a few questions of diction and style, and some problems with the transmission of large sections of letters 16 and 21,[11] but the most urgent concern, which may surprise non-specialists, relates to word length. In Augustan elegy (which is composed in dactylic couplets, one hexameter then one pentameter) it is conventional to end the pentameter with a disyllabic word. Ovid in particular is absolutely punctilious about this in his early verse (the other poets less so), but he varies his style in the exile poetry, permitting more and greater exceptions over time.[12] The Double *Heroides*, once assumed (if genuine) to be composed alongside the single letters, contain three polysyllabic endings. The modern debate has run something like this:

> [W]hen we take into account the development towards polysyllabic pentameter-endings indicated above… it will follow that these pairs of letters cannot be genuine, for it is inconceivable, and conceived by no-one, that Ovid could have written poetry of this type at Tomis.
>
> (Courtney, 1965, 64)

> Courtney… points out that parallels can be found only in the poems from exile; but why should Ovid not have composed *Epp*. 16–21 in exile?
>
> (Reeve, 1973, 330 n. 1)

> If we assume that they are late… the anomalies stop worrying us. The only price to pay would be to assume that Ovid is not very honest when he claims to have given up writing light elegiacs.
>
> (Barchiesi, 1996)

> [I]n the last twenty years scholars have started to read these poems as Ovid's erotica in exile… This is surely right. Stylistically the double epistles belong among the exile poetry, and with the later poems rather than the early *Tristia*.
>
> (Heyworth, 2016, 144–5)

This simple assumption also obviates another objection—that the letters are not mentioned in *Amores* 2.18 or indeed anywhere else in Ovid's work (unusual for a poet who loved to advertise). Hence the authorship of these poems is inextricably entwined with their dating. If they are shown to be Ovidian then they can only be late works, and if they are shown to be late it will be much easier for critics to accept them as Ovidian. I believe that the stylometric analysis demonstrates both.

3 METHODOLOGY

To begin, there are one or two important points that should be made regarding methodology. As mentioned above, the standard approach in authorship questions has been to proceed negatively, by identifying stylistic anomalies. These often take the form of 'X never uses this verb' or 'the word Y does not appear in poetry'—in other words the focus is on *rare events*. To some extent this is a valuable function of the human brain, which is extraordinarily good at noticing contextual anomalies in complex systems. As a stylometric approach, however, it is dangerous. The other methodological disease is the intrusion of subjective judgement, of which two common symptoms are abuses of 'Axelson's Law' in questions of priority and the conflation of 'quality' with 'genuineness'. Although these kinds of approaches are far from a thing of the past, the latter half of the last century began a slow movement away from them. Wilkinson (1955, 84) observes that

[8] Wilkinson (1955, 106) famously compared them to a glut of plum pudding: "The first slice is appetising enough, but each further slice becomes colder and less digestible until the only incentive for going in is the prospect of coming across an occasional ring or sixpence."

[9] Kenney (1996, 20) cites Rand (1925) with approval: "If [the double letters] are not from Ovid's pen, an *ignotus* has beaten him at his own game."

[10] Eg Reeve (1973, 330 n. 1): "Lachmann's observations seem to me much too weak to establish the existence of a second poet as talented as Ovid, or more talented, it might be thought, than the Ovid of *Epp*. 1–15"; even Courtney (1997, 157) (contra) is forced to concur: "[l]et me make no bones about assenting to the general verdict".

[11] See Heyworth (2016, esp. 142–5) with references for an efficient and up to date précis.

[12] Courtney (1997, 160–1) summarises Ovid's development. Interestingly, although not strictly relevant, Propertius went in the opposite direction, permitting all sorts of endings in his early works but becoming more disyllabic over time.

> the nineteenth century laid arrogant hands on the works of many authors. Large portions of the traditional texts were bracketed as interpolations simply because a particular editor felt them to be unworthy of the author, subsidiary 'arguments' being easily discovered *a posteriori*. The *Heroides* have received much attention of this kind.

Kenney (1999), too, begins an article on the Double *Heroides* with an admonition to "beware of hypercriticism", noting in particular that "[s]ingularity is not in itself a ground for suspicion. It makes no sense to require that a writer shall never do anything unless he does it at least twice." From a statistical point of view, the problem with focusing on rare events is *variance*. If I knew the average number of passengers on a London bus, I would find many 'anomalies' as I wandered the city—buses completely empty, or packed full. However, if I knew the average number of commuters for the entire city on Mondays, that number would be much more reliable as a city-wide benchmark for unusual activity. This, then, is the idea behind multivariate analysis, and the reason to focus not on rare events and exceptions but on stylistic events that happen in every line.

It will sound, at this point, as if I am about to proclaim the death of human criticism and the ascendency of the machine, but nothing could be further from the truth. In the first place, computational methods are most useful to test hypotheses, and those hypotheses come from the careful literary judgement of human scholars, once again using that marvellous ability for contextual analysis. Computational stylometry, like anything else, is capable of misapplication, misinterpretation or mistake. The great benefit of multivariate analysis is not to replace scholarly intuition, but simply to offer an objective tool that can work with thousands, or tens of thousands of stylistic factors to provide new kinds of evidence to existing areas of enquiry; instead of looking for anomalies (a negative approach) this is a *positive* approach, assessing overall conformance to a wide range of stylistic preferences.

On, then, to technical matters. In this article I perform two kinds of analysis, poetic and lexico-grammatical. In each case, there are two phases. The first thing that must be done is to establish that the features are stylistically discriminatory, in other words that they can be used to tell authors and works apart. This is evaluated using cross-validated classification accuracy under several different classification algorithms (to ensure that the features are generally useful, not just suited to one kind of model). Note that the goal is not to optimise hyperparameters and find the most accurate classifier possible, it is simply to understand how well works and authors separate in these feature universes. Once the validity of the feature sets has been established, the works are treated as points in high dimensional space and the relationships between them are interrogated using cluster analysis, in the context of a broad elegiac corpus. Here, again, multiple methods are used to ensure that the results are not a fluke of a single algorithm. I use two different methods to project high dimensional data to two dimensional figures, UMAP (McInnes, Healy, & Melville, 2018) and t-SNE (Van der Maaten & Hinton, 2008). Then, following the general techniques introduced by Eder (2017, 57–9), I create a 'bootstrap consensus tree' by aggregating the k-nearest neighbours data for many random feature subsets. This yields a weighted adjacency graph, which is laid out using Fruchterman-Reingold (Fruchterman & Reingold, 1991). In other words two cluster analyses infer relationships based on high-dimensional *position*, and one has the positions determined by the strengths of the *relationships* between works, something like working in the formal mathematical dual.

An important methodological concern is that the units of analysis (i.e. the poems) are small. Practitioners used to working with modern prose might be uncomfortable with samples that are smaller than a few thousand words—in our case, the samples are as small as twenty lines (less than two hundred words). Based on the classification accuracy, however, there is very good reason to believe that we can still draw sensible conclusions. Latin poetry is remarkably stylistically *dense*. A single line of verse is a complex balance between word choice, metre, ictus, the position of the caesurae, the effect of dactylic versus spondaic feet, the interplay of sound within the line and between lines …the list goes on; and with each conscious or unconscious choice the poet leaves stylistic fingerprints. In the lexico-grammatical analysis, I focus on text *n*-grams, since this too allows us to extract more information from small samples. For a morphologically inflected language, a 'standard' BOW (bag of words) analysis, as one might do for English prose, would require that words be lemmatized, which entails a loss of information. *N*-grams, by contrast, break the words into smaller chunks (e.g. 'habet' produces the 3-grams 'hab', 'abe' and 'bet') which captures information both on the lemma as well as on the grammatical inflection, allowing authorial preferences to be extracted in both domains. Once again, the classification accuracy shows that this process extracts enough stylistic signal to reliably separate both authors and works.

Finally, I note that both the complete corpus along with full implementations of the various analyses are freely available at the associated code repository (Nagy, 2022a).[13]

4 THE CORPUS: 'SHORTER' ELEGY

Although this analysis is focused on Ovid's *Heroides*, the validation phase requires them to be in the broader context of classical elegy, to confirm that there is a stylistic distinction between both authors and collections. The test corpus is only moderate in size. It includes 278 poems

[13] The reproduction data is presented in a series of code notebooks, using Python scikit-learn (Pedregosa et al., 2011) with visualisation via R and ggplot2 (Wickham, 2016).

Table 1: A summary of the test corpus.

Author	Work	Poems	Min. Length	Max. Length
Catullus	*Carmina*	5	24	158
Ovid	*Amores*	49	18	114
Ovid	*Heroides*	21	116	376
Ovid	*Ex Ponto*	46	22	166
Ovid	*Tristia*	50	26	578
Propertius	*Carmina*	91	6	150
Tibullus	*Carmina*	16	22	122

Figure 1: The start of the *ES*, before and after phonetic transformation.

ecquid ut inspecta est studiosae litera dextrae
protinus est oculis cognita nostra tuis
an nisi legisses auctoris nomina sapphus
hoc breue nescires unde ueniret opus

ekkwid ut inspekta_st studiosae litera dekstrae
protinus est okulis konjita nostra tuis
an nisi legisses auktoris nomina sappus
hok brewe neskires unde weniret opus

by four authors, drawn from seven 'works' (which here refers to a collection of poems, possibly including several 'books'). The total size of the corpus is 18,726 lines. I included a few works by Catullus (who wrote little elegy) since he seemed a useful point of comparison to the Augustan poets, but it did not seem sensible to venture as far forward as Martial, whose epigrams are a completely different genre, and mostly just a few lines long. Also excluded are Ovid's 'long' works, *Ars Amatoria*, *Fasti*, *Remedia Amoris* etc.—not only do we have enough Ovid already but there are also genre concerns. The final selection is summarised in Table 1A summary of the test corpus.table.caption.1. Finally, the text of every poem is phonetically transformed. This is needed in any case to perform the rhyme analysis, but it also enables the *n*-gram statistics to more accurately isolate sonic preferences—for example to detect an authorial preference for a /k/ sound beginning a word we would otherwise need to consider orthographic 'c' and 'qu' (Fig. 1The start of the *ES*, before and after phonetic transformation.figure.caption.2).

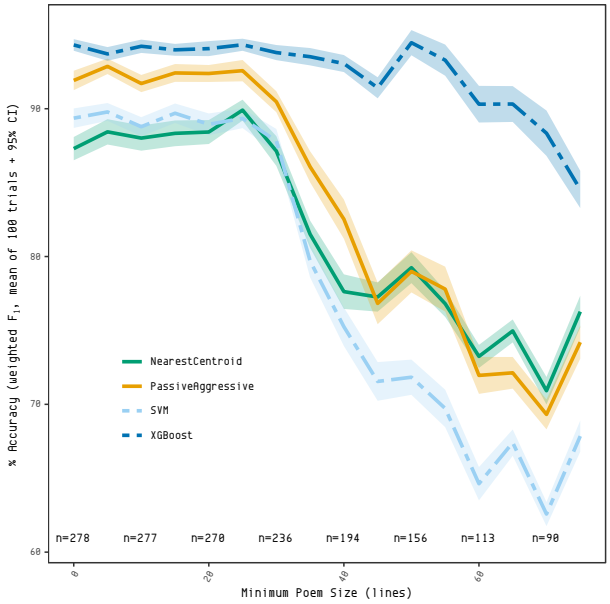
5 RESULTS: LEXICO-GRAMMATICAL ANALYSIS

5.1 Limitations and General Accuracy

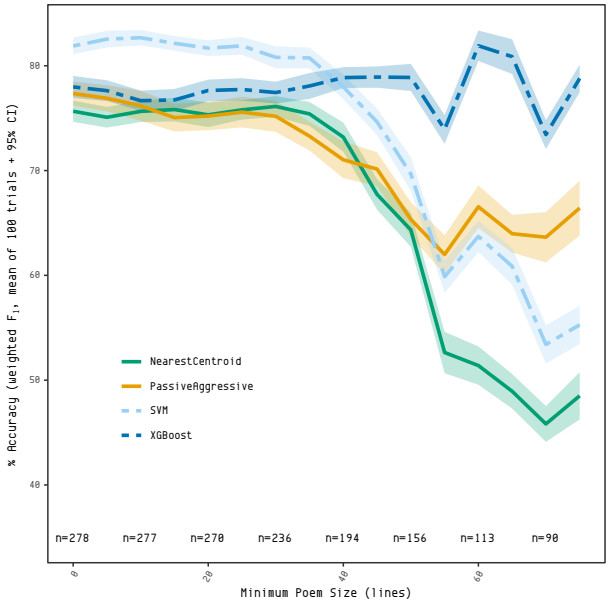
As outlined above, the lexico-grammatical analysis is conducted using TF-IDF scaled frequency counts for character 2-, 3-, and 4-grams. This results in a high dimensional feature space (about 30,000 features); in such environments some care needs to be taken. For hugely overspecified systems, the 'curse of dimensionality' may manifest in a variety of different ways (Zimek, Schubert, & Kriegel, 2012)—in this case, some classification models perform poorly, and dimension-reduction methods like t-SNE are unreliable unless the data is first reduced by some other means. To avoid this, the *n*-gram data is first reduced to 50 dimensions using Singular Value Decomposition. This technique, when applied to TF-IDF data, is fairly standard and is called Latent Semantic Analysis or LSA. Another potential source of error is that the results are affected by topic—for example poems about farming will employ distinctive agricultural lexicon, which can mask or overpower lexical choices imposed by authorial style. Nevertheless, character *n*-grams are a powerful and well-attested technique, shown here to effectively distinguish both authors and works.

First we consider the general accuracy of the supervised classification algorithms. As mentioned, this is presented as a proxy for the general utility of the feature universe; a way of estimating how well authorial style is reflected in these features. Since the size of the poems is a potential cause for concern, Fig. 2How does the classification accuracy change as the minimum poem size in the corpus increases? A comparison using four different algorithms.figure.caption.3 examines the classification accuracy both by work and by author, when considering poems longer than a certain size—the hypothesis being that accuracy should improve if the analysis is restricted to larger poems. As can be seen, this is not the case. In this instance it appears that increasing the size threshold lowers the number of poems available for training, damaging the overall accuracy. Thus, a conservative minimum size of 20 lines was chosen for the cluster analysis, eliminating only eight poems. Next, Fig. 3Confusion matrix, mean of 100 trials. Entries show the percentage of times that a y-axis work was classified as the given x-axis work. Classifier is scikit-learn `NearestCentroid()` using an 80/20 test/train split. Training data is TF-IDF transformed 2-, 3- and 4-gram frequencies for each work, reduced to 50 dimensions with SVD. figure.caption.4 shows the confusion matrix for the `NearestCentroid` classifier (by Work). This shows which works are consistently well-classified over many test-train splits, implying that they are fairly distinctive. As can be seen, the *Heroides* (Ep.) are perfectly classified; they can be reliably be distinguished from Ovid's other works as well as from the works of other authors. It is also apparent from the confusion matrix that the main source of inaccuracy in the model is the stylistic similarity between Ovid's *Tristia* (Tr.) and *Ex Ponto* (Pont.) (this is to be expected—both works are exilic and share many themes) and so the 'true' accuracy is somewhat better than the numbers in Fig. 2How does the classification accuracy change as the minimum poem size

Figure 2: How does the classification accuracy change as the minimum poem size in the corpus increases? A comparison using four different algorithms.

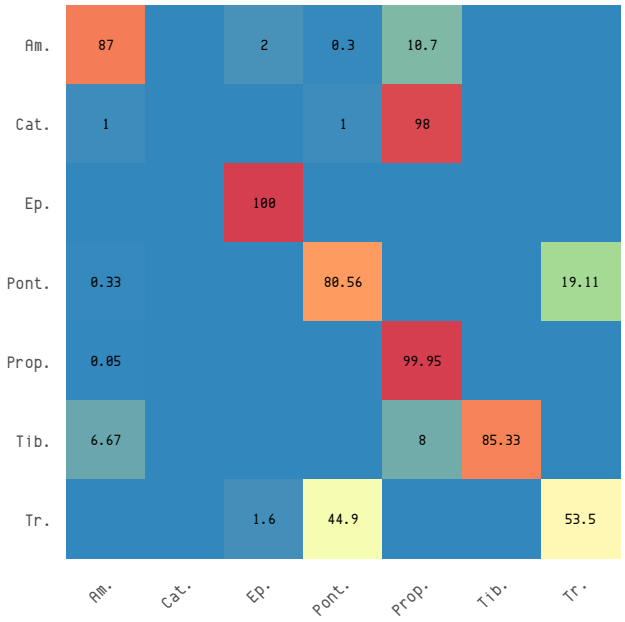


(a) Accuracy by Author



(b) Accuracy by Work

Figure 3: Confusion matrix, mean of 100 trials. Entries show the percentage of times that a y-axis work was classified as the given x-axis work. Classifier is scikit-learn `NearestCentroid()` using an 80/20 test/train split. Training data is TF-IDF transformed 2-, 3- and 4-gram frequencies for each work, reduced to 50 dimensions with SVD.



in the corpus increases? A comparison using four different algorithms.figure.caption.3 would suggest.[14]

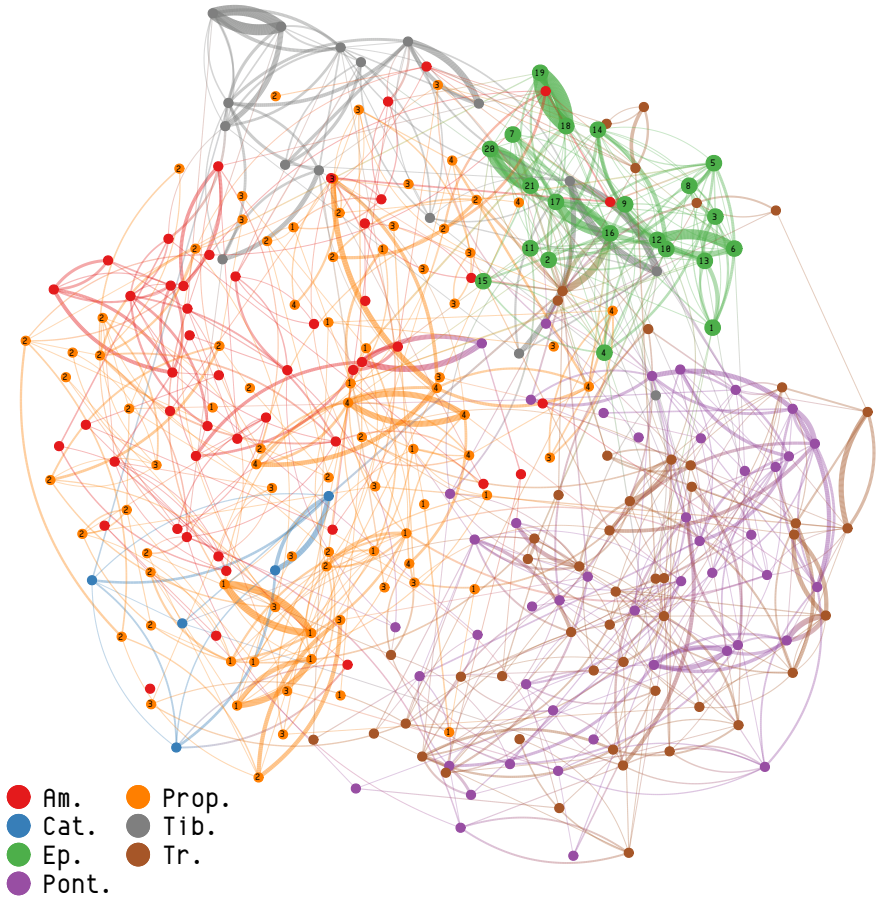
5.2 *Cluster Analysis*

The previous section established that *n*-gram frequencies for the poems contain enough stylistic signal to discriminate meaningfully between both authors and works. Given that this is the case, it is now reasonable to examine the results of unsupervised cluster analysis. The general assumption for all clustering algorithms is that points which are 'close' are somehow related, however the underlying mathematics (as well as precisely what is meant by 'close') can vary greatly. Here, three different clustering algorithms are used, and it is reassuring that the general picture is the same in all three cases. In Fig. 4Cluster analysis of the LSA data (dimension=50), showing lexico-grammatical style.figure.caption.5 we are looking at a latent semantic analysis, which includes a great deal of information from the specific lexicon being employed (from which, if the goals were different, the semantic field could be inferred). Unsurprisingly, in Fig. 4Cluster analysis of the LSA data (dimension=50), showing lexico-grammatical style.figure.caption.5(a) we see a great deal of semantic overlap between Ovid's *Amores* and the works of Propertius ('canonical' love elegy)[15] although

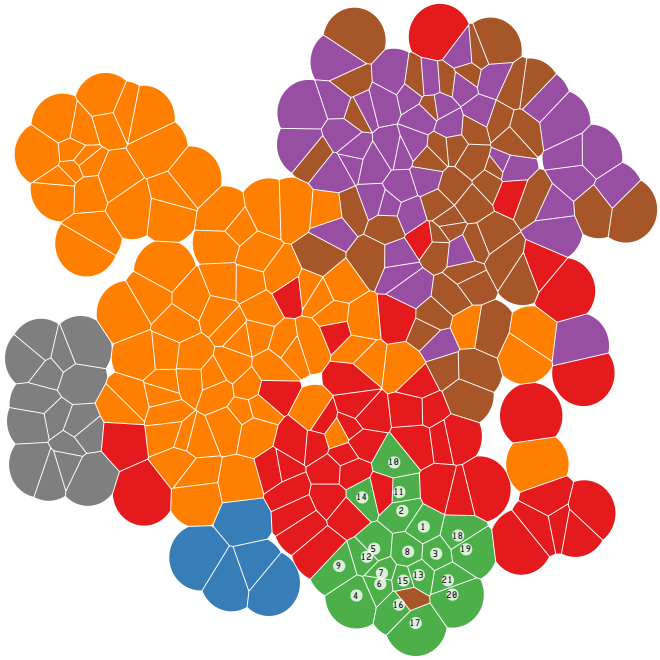
[14] Eliminating just *Ex Ponto* from the corpus increases per-work accuracy by about 10% for both the LSA and poetic models.

[15] The *Amores* are also thought by many to contain deliberate imitation of Propertius; as a starting point see Morgan (1977).

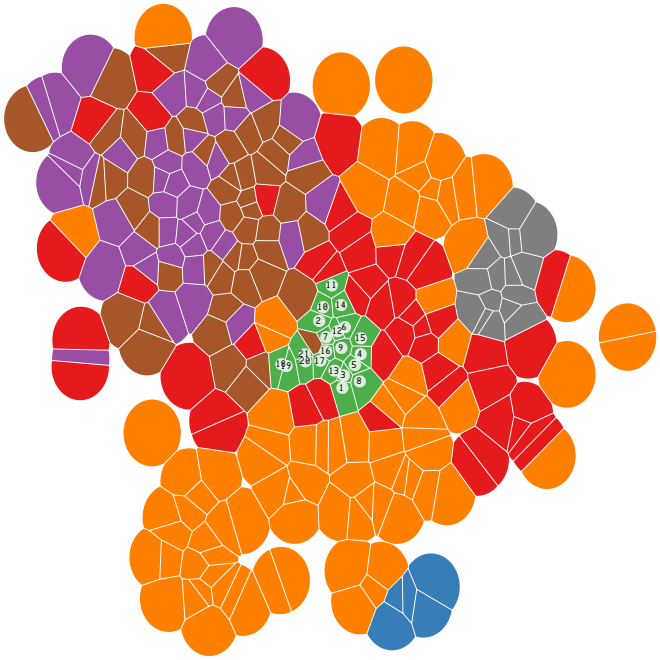
Figure 4: Cluster analysis of the LSA data (dimension=50), showing lexico-grammatical style.



(a) Bootstrap Consensus Tree. Aggregated kNN (k=3, metric=cosine) data from 500 subsets of 15 features. Layout via Fruchterman-Reingold, thicker edges are stronger links.



(b) UMAP Projection (metric=euclidean)



(c) t-SNE Projection (metric=euclidean, perplexity=10)

Tibullus maintains some stylistic separation. Ovid's *Tristia* and *Ex Ponto* mostly form their own cluster, and we can see that the *Heroides* are stylistically distinct, with very strong internal links, most particularly in the paired Double *Heroides*; 16/17, 18/19 and 20/21. Letter 15, the *ES*, clusters neatly with the rest of the *Heroides*, giving no cause to question its authenticity. A more or less identical picture arises from the two other algorithms, Figs 4Cluster analysis of the LSA data (dimension=50), showing lexico-grammatical style.figure.caption.5(b) and 4Cluster analysis of the LSA data (dimension=50), showing lexico-grammatical style.figure.caption.5(c)—The *Amores* mix with Propertius, Tibullus is a little more distinctive, and the *Heroides* form a tight group.

6 RESULTS: POETIC ANALYSIS

In this section I shift from well-established methods in computational stylometry to newer techniques that examine non-lexical features that are indigenous to poetry. This combines and extends research that I began on hexameter metre (Nagy, 2021) as well as work on the stylistic signal contained in deliberate sonic correspondence (Nagy, 2022b). Thus, while the results are emphatic and the techniques appear powerful, it should be made clear that this is an area of emerging research. The features being examined here are purely poetic—there is no consideration of the words being used, only the way in which the verse is constructed. The features are separated into broad domains. First, foot patterns and pauses, which includes the type of foot at each position (dactyl or spondee) as well as the position of all caesurae and diaereses; second the interplay between ictus (the start of the metrical foot) and accent (syllables bearing stress); and third, the sonic style, including the number and strength of 'sonic correspondences' (rhyme, roughly speaking) of various kinds. Finally, since it is crucial to the traditional debate, I measure the degree to which the length of the final word in the pentameter is allowed to vary, and also consider the overall length of the poems (this was included because some of the *Heroides*, 16 in particular, are considerably longer than the general trend; a point which has been raised during debate over its authenticity). These features are summarised in Table 2A summary of the poetic featurestable.caption.6. Not every possible feature is measured—for example the final two feet of the pentameter are mandatory dactyls, and so the features `P3SP` and `P4SP` would contain no stylistic information. The final universe contains 43 features; a reasonable number, but hopefully few enough to escape the label of 'high dimensionality'. The hope (which seems to be borne out by the results) is that poetic features are less obscured by things like topic and genre, and should present a cleaner stylistic signal.

6.1 *Limitations and General Accuracy*

As before, the first task is to establish that the features are able to meaningfully distinguish authorial style. The results here are broadly similar to those in §5.1Limitations and General Accuracysubsection.5.1. In Fig. 5How does the classification accuracy change as the minimum poem size in the corpus increases? A comparison using four different algorithms.figure.caption.7 we see again that the accuracy decreases as the shorter poems are filtered out. The best performing classifier is now `NearestCentroid`, and the F_1 accuracy by Work (the most important figure in this context) is a little lower—although it should hardly be surprising that it becomes more difficult to identify the theme of a poem if one is forbidden from reading the actual words! It is important to note, however, that the accuracy by Author is slightly higher in the poetic models than the previous LSA models. At the risk of subjectivity, this seems to suggest that the underlying performance of the models that use poetic features is just as strong, but that the additional information derived from lexicon and topic in the LSA models gives the per-Work accuracy an 'unfair' boost. In the confusion matrix (Fig. 6Confusion matrix, mean of 100 trials. Entries show the percentage of times that a y-axis work was classified as the given x-axis work. Classifier is scikit-learn `NearestCentroid()` using an 80/20 test/train split.figure.caption.8) we see again that the *Heroides* are quite distinctive in their style, while Ovid's *Tristia* and *Ex Ponto* are easily confused. Overall, particularly considering the extremely good performance when identifying authors, the poetic features seem at least as useful and, based on the clustering results, are probably more so.

6.2 *Feature-Wise Analysis*

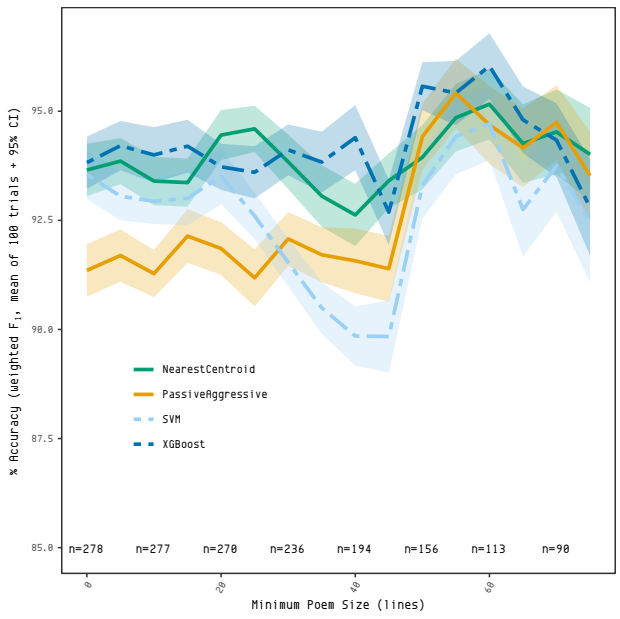
In Nagy (2021, 1007–9) I discussed a statistical technique that used the Mahalanobis distance to not only determine how consistent an observation was with a given comparison set,[16] but also to provide a per-feature analysis of the most unusual features. In this section, the same technique is applied to the elegiac corpus. The advantage of the Mahalanobis distance is that it corrects for feature correlation and covariance. In the case of prosodic features there are many correlations—as just one example, the practice of ending a pentameter with a disyllable (which primarily affects the feature `PFSD`) also forces a weak caesura in the fourth foot (`P4WC`). Taking the centroid of all the Ovidian works (164 poems), the Mahalanobis distance was calculated for every poem in the corpus, representing the degree to which the poems conform to Ovid's 'typical' style. From there, since the (squared) Mahalanobis

[16] This is a form of outlier detection, sometimes called a 'one-class problem' in authorship attribution circles, and it has its own branch of the literature. Koppel, Schler, Argamon, and Winter (2012) is a reasonable place to start.

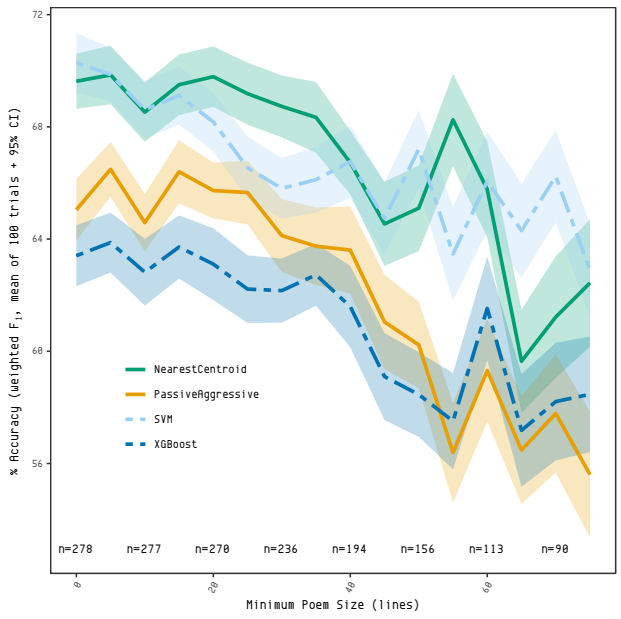
Table 2: A summary of the poetic features

Feature	Description
`HnSP, PnSP`	Proportion of couplets where foot *n* is a Spondee in either the (H)exameter or (P)entameter.
`HnCF, PnCF`	Proportion of couplets with Ictus/Accent Conflict in foot *n*.
`HnDI, PnDI`	...Diaeresis in foot *n* (— ⏑ ⏑ ‖ or — — ‖).
`HnSC, PnSC`	...Strong Caesura (— ‖).
`HnWC, PnWC`	...Weak Caesura (May only exist in dactylic feet: — ⏑ ‖).
`ELC`	Average number of elisions per line. Does not include prodelision.
`LEN`	The length, in lines, of the entire poem.
`RS`	The average rhyme strength. Measures both the number and strength of rhymes (or other sonic correspondences) whether vertical or horizontal.
`LEO`	The average number of 'leonine' rhymes (rhymes between the words at the central caesura and the end-of-line).
`PFSD`	The standard deviation of the length (in syllables) of the final word in the pentameter. Early Ovidian practice (ending every pentameter with a disyllable) would show a `PFSD` of zero.

Figure 5: How does the classification accuracy change as the minimum poem size in the corpus increases? A comparison using four different algorithms.

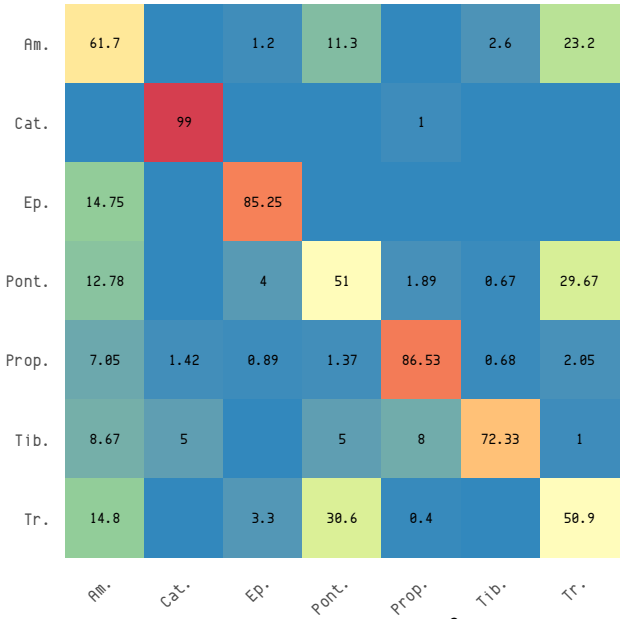


(a) Accuracy by Author



(b) Accuracy by Work

Figure 6: Confusion matrix, mean of 100 trials. Entries show the percentage of times that a y-axis work was classified as the given x-axis work. Classifier is scikit-learn `NearestCentroid()` using an 80/20 test/train split.



distance is distributed according to χ^2, a P-value can be calculated. The method is reasonably powerful. Of the (102) non-Ovidian works, only five would be accepted as Ovidian style at the 99% confidence level.[17] Of the (164) Ovidian works, there are fourteen outliers at that same confidence level,[18] but none of the *Heroides*, including, obviously, the *ES*, are outliers at any level of confidence. The obvious objection to this analysis is that none of the non-Ovidian poems were attempts at imitation. Could an interpolator have simultaneously imitated every one of Ovid's poetic tendencies (while also using Ovidian lexico-grammatical style)? In my view, no; but Late Antique and Humanist authors achieved levels of Latinity that few today could match. Or, could the statistical features (foot patterns, caesura positions, etc.) be imitated 'instinctively' by a gifted poet who had completely internalised Ovidian style? In the end, it cannot be ruled out. However, as previously pointed out, we would be dealing with someone with, essentially, Ovid's talent, as well as an extraordinary gift for imitation.

6.3 Cluster Analysis

Looking at the clustering results, the first observation is that there is clear support for the general hypothesis that poetic features provide a cleaner stylistic signal. Where Fig. 4Cluster analysis of the LSA data (dimension=50), showing lexico-grammatical style.figure.caption.5 suffered from significant overlap between authors that was driven by topic, in Fig. 7Cluster analysis of the z-scaled data (43 features), showing poetic style.figure.caption.9(a) we have clear separation by authorial style, as well as further discrimination within Author clusters that appears to be a temporal signal (discussed in §6.4Searching for a Temporal Signalsubsection.6.4). In all three analyses, the bulk of the Single *Heroides* form one cluster within Ovidian style, and the Double letters cluster tightly among themselves, showing, again, a stylistic 'break' between the two groups. The *ES* again falls in the middle of the Single letters—it is completely typical of the style of this group, as well as being characteristically Ovidian. In the UMAP and t-SNE projections (Figs 7Cluster analysis of the z-scaled data (43 features), showing poetic style.figure.caption.9(b) & 7Cluster analysis of the z-scaled data (43 features), showing poetic style.figure.caption.9(c)) the separation between the *Amores* and *Tristia*/*Ex Ponto* can be seen more clearly, with the Single *Heroides* clustering in with the former and the Double letters with the latter.

6.4 Searching for a Temporal Signal

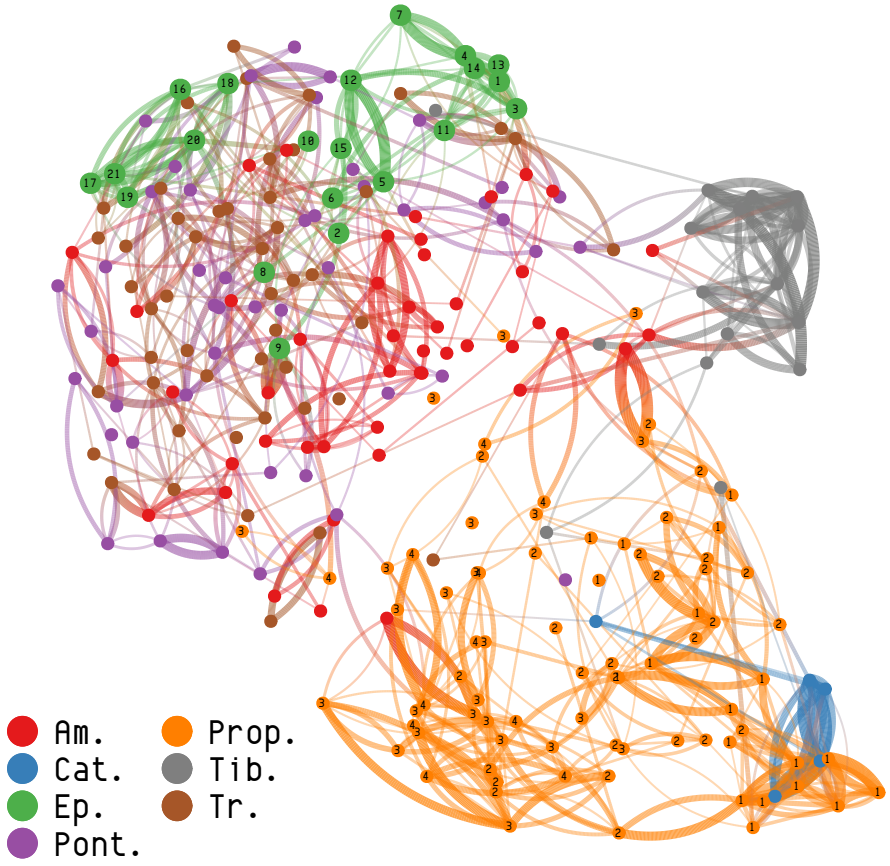
As a final experiment, here is a more specific attempt to detect a temporal signal in the poetic style (recall that by 'poetic' I mean specifically non-lexical style—prosody, metre, soundplay…) of the *Heroides*. The central assumption is that we have two points of comparison, the *Amores* for early style, and the *Epistulae Ex Ponto* for late style. There are complications, though, in the 'early' style of the *Amores*—not only was the collection heavily edited from five books to three, but modern scholarship tends towards the view that some of the poems were included or revised later in Ovid's career.[19] Nevertheless, these are the best witnesses we have to Ovid's earliest elegiac style. The general approach taken is to consider each letter and plot the relative similarity to either the *Amores* or to *Ex Ponto*, with the centre of the graph being stylistically equidistant. There is no attempt here to classify the *Heroides*, and no test/train partition of the data. Two models are trained on the full corpus, one `SVM` and one `NearestCentroid` (these models are the easiest from which to extract a relative similarity metric, and were also the most generally accurate in testing—see Fig. 5How does the classification accuracy change as the minimum poem size in the corpus increases? A comparison using four different algorithms.figure.caption.7). In Fig. 8Progression of the prosodic style of the *Heroides* showing the comparative similarity to either the *Amores* (early style) or *Ex Ponto*

[17] In order of decreasing similarity: Prop. 4.11, 4.6, 3.24, 4.4; Tib. 1.4.

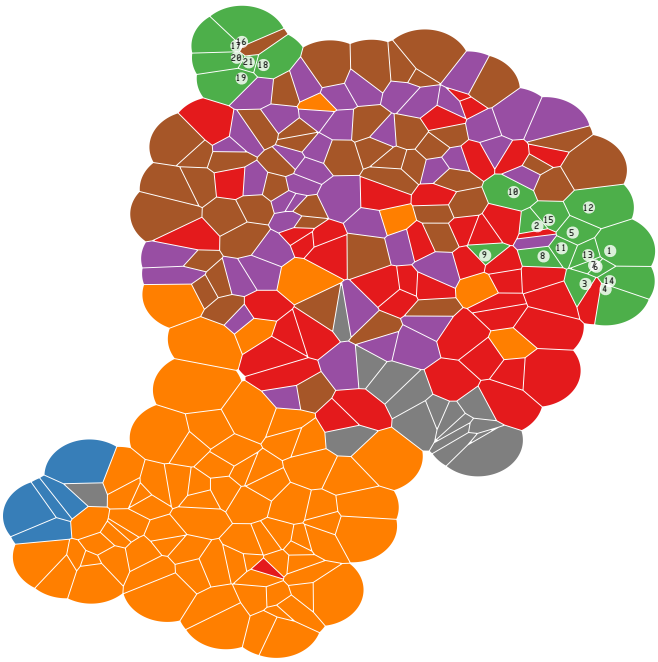
[18] Mostly from *Tristia* and *Ex Ponto*, but also *Amores* 1.11, 2.8, and 2.13.

[19] On these and other chronology questions, the most up-to-date summary is Davis (202x, i–iii). The idea that parts of the *Amores* may have been revised over decades might account for the wide variation in style that can be seen in Fig. 7Cluster analysis of the z-scaled data (43 features), showing poetic style.figure.caption.9.

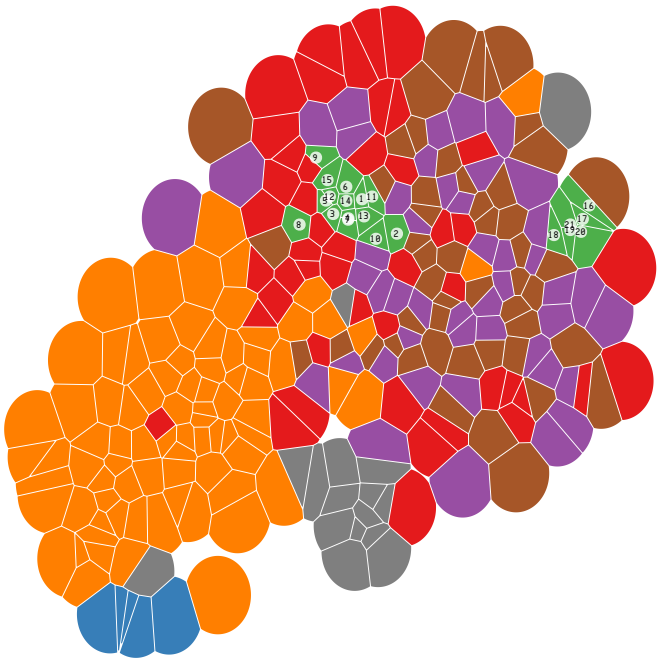
Figure 7: Cluster analysis of the z-scaled data (43 features), showing poetic style.



(a) Bootstrap Consensus Tree. Aggregated kNN (k=3, metric=cosine) data from 500 subsets of 15 features. Layout via Fruchterman-Reingold, thicker edges are stronger links. *Heroides* enlarged and numbered, works by Propertius numbered by Book.

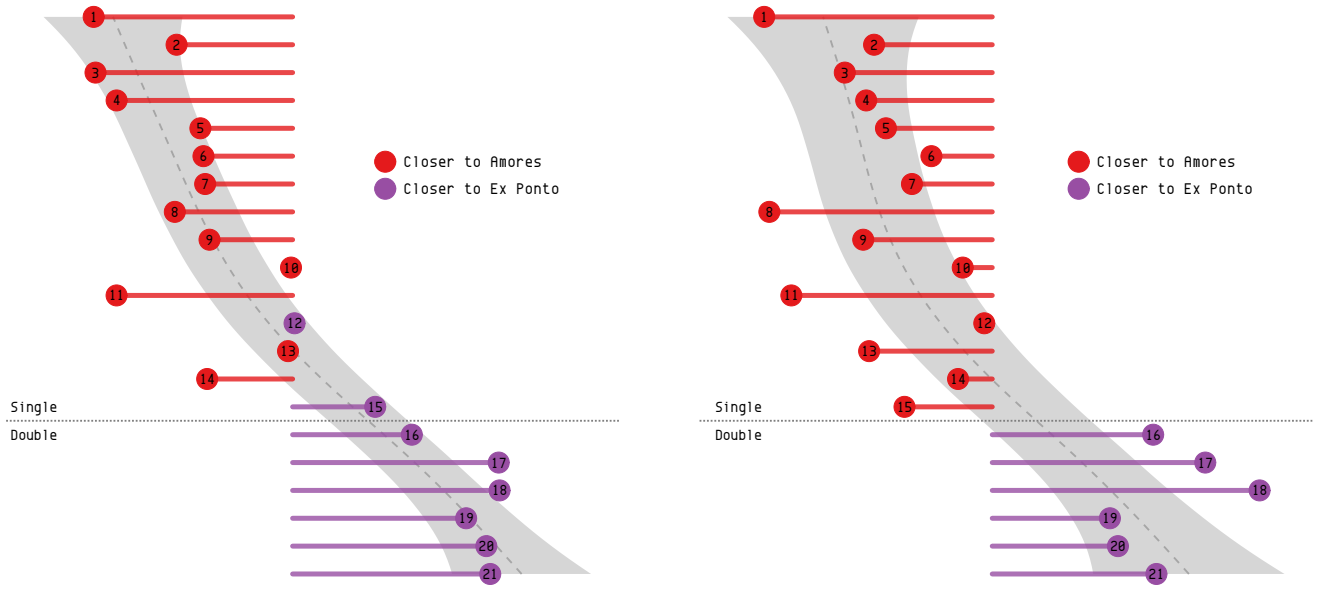


(b) UMAP Projection (metric=euclidean)



(c) t-SNE Projection (metric=euclidean, perplexity=12)

Figure 8: Progression of the prosodic style of the *Heroides* showing the comparative similarity to either the *Amores* (early style) or *Ex Ponto* (late style). In grey: a GAM trendline with 95% confidence interval.



(a) SVM (relative confidence margins, based on distance from respective hyperplanes)

(b) Nearest centroid (Euclidean distance)

(late style). In grey: a GAM trendline with 95% confidence interval.figure.caption.10 we see the results. Although the fine details differ according to the model, the results support two assertions: first that the style of the *Heroides* gradually evolves from 'early' to 'late', supporting the idea of sequential composition; second, that the Double *Heroides* seem to be markedly 'later' in style, supporting the view that they were composed in exile. In terms of the *ES*, the style seems 'fairly late', which is consistent with the other evidence placing it in its present position as the fifteenth letter, but in my view the results are too variable to make that case in the absence of other evidence, and they are certainly too weak to suggest any kind of re-ordering of the Single letters (for example the eleventh letter from Canace to Macareus). Since the temporal signal seems genuine, it is tempting to look at the tight cluster of the Double *Heroides* in Fig. 7Cluster analysis of the z-scaled data (43 features), showing poetic style.figure.caption.9(a) as evidence that they were composed within a fairly short space of time (since there is little stylistic drift, unlike the Single letters), but that is merely speculation.

Although it is not directly relevant to the *Heroides*, it is also interesting to consider the poetic development of Propertius, which can be seen strongly in Fig. 7Cluster analysis of the z-scaled data (43 features), showing poetic style.figure.caption.9(a). The first book (called the *monobiblos*) is strongly self-similar, and there is a fairly obvious split between the first two books (mainly Cynthia poems) and the third and fourth which some argue were subject to the influence of Maecenas (and thus indirectly of Augustus).[20] There also appears to be a fairly close link between the *monobiblos* and the poetics of Catullus, which has a bearing on the stylistic origins of Augustan elegy.[21] In summary, this technique seems sensitive, even at this early stage, to very fine gradations of poetic style; not only authorial, but even to the development of an author's stylistic preferences throughout their career. There is significant potential here for future work.

7 CONCLUSIONS

Kenney (1979, 395) observed that when attempting to prove authenticity (as opposed to disputing it) "[t]he most

[20] As well as the potential for political influence, which after all might be expected to affect genre and topic but not necessarily style, it should be borne in mind that the interaction between the various poets of Maecenas' circle was likely to produce some stylistic cross-pollination.

[21] Catullus' metrical technique was once considered 'careless and uncouth', but this view was challenged by D. A. West (1957) and revisited by Duhigg (1971). Nobody questions the status of Catullus as *some* kind of precursor to the Augustan elegists (although much ink has been spilled over the details), but given the closeness (based on the cluster analysis, in particular Fig. 7Cluster analysis of the z-scaled data (43 features), showing poetic style.figure.caption.9(a)) it is worth wondering: did Propertius deliberately model his earliest metrics on Catullus? Or perhaps does the stylistic link pass through the lost elegies of Gallus? Of course it could be mere coincidence—a much more rigorous analysis would be needed before pressing this case.

that it is usually realistic to expect is what I have called a verdict of *nihil obstat*: a conclusion that, on balance, a disputed work or passage contains nothing fundamentally inconsistent with its reputed authorship". In this article I have attempted to show that, with computational methods, it may be possible to expect more. More than simply showing that there is nothing inconsistent with Ovidian authorship, multivariate analysis demonstrates that the style of the *Heroides*, including the Double letters (and including, by the way, *Heroides* 12 whose authenticity has at times been questioned) conforms closely with dozens of features that reflect Ovid's typical practice. This is *positive* evidence—in every poetic feature that was measured, whether the distribution of dactyls and spondees in each foot position, whether the type and frequency of caesurae, the frequency of elision, even the frequency and position of ictus conflicts, in all of these measurements the *Heroides* are consistent with Ovidian practice. This does not 'prove' Ovidian authorship, but it demonstrates the incredible lengths, far beyond simply writing Ovidian words, to which the hypothetical interpolator must go to produce a false positive. Two kinds of analysis were performed, and in each case it was shown that there is enough stylistic signal available to reliably differentiate both authors and works. LSA (§5Results: Lexico-Grammatical Analysissection.5) reflects a wide range of authorial preferences, lexical, grammatical, and phonetic. While it can be clouded by external factors like genre and topic, it is a reliable and well-established approach. The poetic analysis (§6Results: Poetic Analysissection.6) examined technical minutiae of prosody and metre that are standard considerations in classical philology; the innovation lies in their automatic extraction and their treatment with modern multivariate analysis and statistics. The inclusion of some stylistic features related to rhyme is perhaps controversial, but my recent research (Nagy, 2022b) has established that there is authorial signal in this domain. The poetic models performed very strongly, particularly in terms of differentiating authors, and it is in this domain that the separation between the Single and Double letters is most obvious.

My interpretation of the results is quite simple: all twenty-one of the *Heroides* are genuinely Ovidian, the *ES* is in the correct position, and the Double *Heroides* were written in exile. The analysis worked surprisingly well, considering the very small samples—this demonstrates the stylistic density of Latin poetry, and is cause for optimism regarding other questions of authorship. In this case, particularly because of the stylistic distinctiveness of the *Heroides*—they form tight clusters—the results are quite clear. Other questions are less so. For example, the authenticity of *Amores* 3.5 is still questioned by some; so far I can say that nothing in this analysis suggests that *Am.* 3.5 is *not* Ovidian (*nihil obstat!*), but the *Amores* show much more stylistic diversity than Ovid's other works, so some more focused analysis would be prudent. At the end of the day, computational stylometry is an exercise in statistics and probability: the data will not always provide a clear answer to every question, but in the case of the *Heroides*, I suggest that it does.

8 FUTURE WORK

I conclude with some thoughts about work left undone. In the *Heroides*, the authorship of two lengthy sections (16.39–144 and 21.145–248) have been questioned.[22] These sections are transmitted via a single witness which is late (1477), and mechanically printed. Since the sections are long enough to be analysed as a unit with the techniques used here, they could be isolated and examined somehow. However, my view—which I have not formally tested—is that a hundred or so lines of foreign material (a little less than a third and a half of the poems respectively) would certainly have altered the stylistic fingerprint of those poems at least to the point where they appeared unusual, and so I accept them with only the slightest reservation. In other Ovidian matters, above I mentioned *Amores* 3.5, which seems worth investigating further, since the debate has persisted into this century.[23] The challenges will be the smaller size of the work (although 46 lines, based on these results, now seems fairly reasonable) but more so the huge stylistic variation encountered in the *Amores*—the more things vary, the more difficult it is to decide if something is atypical. Considering other authors, there would no doubt be great interest in the long-debated elegies of Sulpicia (which appear in the so-called *Corpus Tibullianum*)—the only surviving (ancient) Latin verse written by a woman. I am not, however, at all optimistic. The poems are tiny, and the undisputed Tibullan corpus is meagre in comparison to Ovid.[24] Finally, there remain many ways in which the analysis of non-lexical style might be improved, such as measuring 'complexity', register and intertextuality, all of which are challenging problems in machine learning. The methods used here are still evolving. Nonetheless, it is hoped that the present work might contribute to furthering our understanding of this debated and divisive Ovidian collection.

ACKNOWLEDGEMENTS

This work was supported by project 2020/39/O/HS2/02931, funded by Poland's National Science Centre.

Parts of this article were given as a paper in the seminar series of the department of Classics, Archaeology and Ancient History

[22] Kenney (1979) examines the passages in detail. He considers the literary arguments against to be insufficient.

[23] E.g. McKeown (2013), which also serves as a fine synthesis of the issues.

[24] Tibullus is not the only alternative candidate that has been proposed, however a fair amount of the historical debate is little more than thinly-veiled sexism. Mathilda Skoie's (2002) monograph is a good place to begin further investigation.

at the University of Adelaide in October 2021, and I am grateful to the participants for their useful comments and advice. Further thanks are due to Prof. Peter Davis for a variety of suggestions that have undoubtedly improved the article, and for piquing my interest in *Am* 3.5.

Dr J. Stover, editor of the forthcoming *Oxford Guide to the Transmission of the Latin Classics*, provided feedback on the draft, and was kind enough to send me Prof. J. B. Hall's entry on the *Heroides*. I note with satisfaction that Hall shares the growing consensus that all 21 letters are genuine. Dr Stover, also a computational stylometrist (e.g. Stover and Kestemont (2016)), further informs me (pers. comm.) that his own independent analysis had produced results consonant with those here.

REFERENCES

Barchiesi, A. (1996). Review of EJ Kenney (ed.), Ovid Heroides XVI–XXI (Cambridge). *Bryn Mawr Classical Review*, *12*.

Courtney, E. (1965). Ovidian and non-Ovidian Heroides. *BICS*(12), 63–66. doi: 10.2307/43646359

Courtney, E. (1997). Echtheitskritik: Ovidian and non-Ovidian Heroides again. *CJ*, *93*(2), 157–166.

Davis, P. J. (202x). *Ovid Amores book 3 edited with introduction, translation, and commentary*. OUP (forthcoming).

Duhigg, J. (1971). The elegiac metre of Catullus. *Antichthon*, *5*, 57–67.

Dörrie, H. (Ed.). (1971). *Epistulae heroidum*. de Grutyer.

Eder, M. (2017). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, *32*(1), 50–64.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164.

Heyworth, S. J. (2016). Authenticity and other textual problems in Heroides 16. In R. Hunter & S. P. Oakley (Eds.), *Latin literature and its transmission: Papers in honour of Michael Reeve* (pp. 142–170). CUP.

Kenney, E. J. (1979). Two disputed passages in the Heroides. *CQ*, *29*(2), 394–431.

Kenney, E. J. (Ed.). (1996). *Ovid Heroides XVI–XXI*. CUP.

Kenney, E. J. (1999). Anomaly, innovation and genre in Ovid, Heroides 16–21. *Aspects of the Language of Latin Poetry (Proceedings of the British Academy)*, *93*, 399–414.

Koppel, M., Schler, J., Argamon, S., & Winter, Y. (2012). The "Fundamental Problem" of Authorship Attribution. *English Studies*, *93*(3), 284–291. doi: 10.1080/0013838X.2012.668794

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*.

McKeown, J. C. (1998). *Ovid Amores: Text, Prolegomena and Commentary* (Vol. III). Francis Cairns.

McKeown, J. C. (2013). The authenticity of Amores 3.5. In *Vertis in usum* (pp. 114–128). BG Teubner.

Morgan, K. (1977). *Ovid's art of imitation: Propertius in the Amores* (Vol. 47). Brill Mnemosyne Supplements.

Nagy, B. (2021). Metre as a stylometric feature in Latin hexameter poetry. *Digital Scholarship in the Humanities*, *36*(4), 999–1012. doi: 10.1093/llc/fqaa043

Nagy, B. (2022a). Preprint: Some stylometric remarks on Ovid's *Heroides* and the *Epistula Sapphus*. https://github.com/bnagy/heroides-paper. doi: 10.5281/zenodo.6099889

Nagy, B. (2022b). Rhyme in classical Latin poetry: stylistic or stochastic? *Digital Scholarship in the Humanities*, (in press). doi: 10.1093/llc/fqab105

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reeve, M. D. (1973). Notes on Ovid's Heroides. *CQ*, *23*(2), 324–338. doi: 10.2307/638190

Skoie, M. (2002). *Reading Sulpicia: Commentaries, 1475-1990*. OUP.

Stover, J., & Kestemont, M. (2016). Reassessing the Apuleian corpus: A computational approach to authenticity. *CQ*, *66*(2), 645–672. doi: 10.1017/S0009838816000768

Tarrant, R. J. (1981). The authenticity of the letter of Sappho to Phaon (Heroides XV). *Harvard Studies in Classical Philology*, *85*, 133–153. doi: 10.2307/311169

Tarrant, R. J. (1983). Ovid, Heroides. In L. D. Reynolds (Ed.), *Texts and Transmission* (pp. 268–73). OUP.

Thorsen, T. (2014). *Ovid's early poetry: from his Single Heroides to his Remedia Amoris*. CUP. doi: 10.1017/CBO9781139628952

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

West, D. A. (1957). The metre of Catullus' elegiacs. *CQ*, *7*(1-2), 98–102.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wilkinson, L. P. (1955). *Ovid Recalled*. Cedric Chivers. (repr. 1974)

Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, *5*(5), 363–387. doi: https://doi.org/10.1002/sam.11161