Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae*

Christopher W. Forstall

State University of New York at Buffalo

Sarah L. Jacobson

State University of New York at Buffalo

Walter J. Scheirer

University of Colorado at Colorado Springs

Correspondence:

Christopher W. Forstall

338 MFAC

University at Buffalo

Buffalo NY 14621

USA

email: forstall@buffalo.edu

Abstract

In this study we use computational methods to evaluate and quantify philological evidence that an eighth century CE Latin poem by Paul the Deacon was influenced by the works of the classical Roman poet Catullus. We employ a hybrid feature set composed of n-gram frequencies for linguistic structures of three different kinds—words, characters, and metrical quantities. This feature set is evaluated using a one-class support vector machine approach. While all three classes of features prove to have something to say about poetic style, the character-based features prove most reliable in validating and quantifying the subjective judgments of the practicing Latin philologist. Word-based features were most useful as a secondary refining tool, while metrical data were not yet able to improve classification. As these features are developed in ongoing work, they are simultaneously being incorporated into an existing online tool for allusion detection in Latin poetry.

1 Introduction

The study of *intertextuality*, the shaping of a text's meaning by other texts, remains a labor-intensive process for the literary critic. Julia Kristeva, who coined the term *intertext*, suggested, 'Any text is constructed as a mosaic of quotations; any text is the absorption and transformation of another' (Kristeva, 1986, 37). Such transformations range from direct quotations, representing a simple and overt intertextuality, to more complex references that are intentionally or subconsciously absorbed into a text. In the years since Kristeva first drew attention to the phenomenon, the field of its study has become increasingly—in some cases debilitatingly—complex, as a recent study in this journal showed (Trillini *et. al.*, 2010). As this theoretical complexity grows, so does the burden upon the practicing philologist to verify suspected instances of intertextuality. The critic must command a large corpus of possible contributing works; meanwhile, objective criteria by which intertext may be measured are lacking. Since, in many cases, the problem is one of pattern recognition, the task of locating new relationships between texts and validating suspected ones is a good candidate for automated assistance by computers.

In this work, we propose the use of machine learning and related statistical methods to improve the process by which intertextuality is studied. Specifically, we bring to bear computational techniques from the field of authorship attribution in order to examine instances where an author who is familiar with a particular corpus deliberately or subconsciously reflects this in discrete passages within his own work. We have defined three different classes of style markers to quantify intertext: phonetic, metric, and dictional.

To evaluate the proposed style markers and classification methods, we have chosen an intriguing case study in Latin poetry. Traditional philological analysis suggests that Paul the Deacon's eighth century CE poem *Angustae Vitae* has a strong connection to classical Roman poetry of the first centuries BCE and CE, and specifically to the works of Catullus. Using the digital methods outlined above, we attempt to replicate and quantify this intertextual reading of Paul, with a view to developing more sensitive methods for detecting intertext with the computer.

2 Paul the Deacon's *Angustae Vitae* and a Hypothesized Link to Catullus

Paul the Deacon was an eighth century monk and an intellectual, a court poet and historian of the Lombards and briefly of Charlemagne (Harrington *et. al.*, 1997, 197). Although today he is better known for his historical and antiquarian work, his poetic output shows that he took an intellectual interest in matters of poetics: he engaged in friendly competition with contemporary men of letters, such as Peter of Pisa, he discussed classical authors in his verses, and he incorporated into his poems virtuoso tricks popular in his time, such as acrostic messages. He wrote verse both in the classical literary tradition and in the style of contemporary spoken Latin. This paper considers in particular the poem beginning *Angustae vitae* . . . (Poem 5 in the edition of Dümmler, 1881), an epistolary poem in which Paul juxtaposes poetic inspiration and production in the classical world with the writing of poetry in the Christian monastic context.

Although Paul posits the classical and monastic worlds as opposites, he uses diction and thematic models which recall the so-called Neoteric poets of the late Roman republic and their successors of the early empire: the poet/lover, his beloved, and his poetological concerns. Paul re-contextualizes this model to reflect monastic love and poetic exchange. We know from his other writings that he was well-versed in the poetry of Horace, Virgil and Ovid, and others have noted possible classical intertexts in this poem (e.g. Dümmler, 1881 on line 5.19). Here we argue that in several places his style is particularly reminiscent of Catullus. Although it is generally believed that Paul did not have knowledge of Catullus' poetry, in this work we make the assumption that he had, based on as-yet unpublished evidence brought forward by one of our authors.[1] For the present, we use digital analysis to support and quantify a traditional, subjective reading. As we refine the digital metrics introduced here, our eventual goal is to augment the capabilities of the web-based allusion-detection tool *Tesserae* (Coffee, 2010).

We begin with a brief survey of Paul’s twenty-line poem, identifying the places in which traditional philology suggests connections to Catullus. To furnish a control, we also identify places where connections to other classical poets are generally accepted. The principal question we address is: can the tools of authorship attribution be employed to validate or at least quantify subjective, literary claims about intertext?

3 An Intertextual Reading of *Angustae Vitae*

We begin with the text of Paul's poem:

Angustae vitae fugiunt consortia Musae,
Claustrorum septis nec habitare volunt,
Per rosulenta magis cupiunt sed ludere prata,
Pauperiem fugiunt, deliciasque colunt:
Quapropter nobis aversae terga dederunt,
Et comitem spernunt me vocitare suum.
Inde est quod vobis inculta poemata mitto,
Suscipe sed libens qualiacumque tamen.
Inmodico flagrat de vestro pectus amore,
Crede pater, nostrum, semper amande mihi,
Et peream, si non tecum captare per aevum
Per domini munus regna beata volo.
Hoc mihi est votum, hoc fido pectore spero,
Hoc licet indignus nocte dieque precor.
Tu quoque, si felix vigeas de munere Christi,—
Namque potes—misero redde, beate, vicem.
Ante potest flavos Hrenus repedare Suavos,
Ad fontem et versis pergere Tibris aquis,
Quam tuus e nostro labatur pectore vultus,

Ore colende mihi tempus in omne pater!

(The Muses flee the societies of the constrained life,
Nor wish to inhabit the gardens of cloisters;
Rather they desire to play in rosy meadows.
They flee from poverty; they cultivate delights;
So they turn their backs, averse to me,
And disdain to call me their companion.
Thus it is I send my uncultivated poems to you;
But receive them, such as they are, with good will.
My breast burns with unrestrained love for you—
Father ever beloved to me, trust that
I shall perish unless I wish at last to reach with you
The blessed kingdom, by the gift of the Lord.
This is my vow; for this I hope with a faithful heart;
Though undeserving, I pray for this night and day.
And you, if you should be well and flourishing by the gift of Christ,
Send a poor wretch a reply, fortunate one—for certainly you are able.
The Rhine will be able to turn from the blond Swabians
And Tiber, with his waters reversed, proceed toward his source
Before your image will slip from my heart,
Father worthy to be remembered on my lips for all time.
—Paul the Deacon, Poem 5 in Dümmler, 1881, with our own translation.)

As the poem opens, we learn that the Muses do not wish to live in the fenced-in gardens of monasteries, but rather they desire to play in rosy meadows. Here, *septa,* the cultivated, enclosed garden of a monastery, is contrasted with the *rosulenta prata*, a wild, open meadow. These lines may reference *Eclogue* 1 of Virgil (Mynors, 1969) where *ludere* (‘to play’), *fugere* (‘to flee’), and *pauper* (‘poor’) all play important roles; lines 17–18 also allude to Virgil’s *Eclogue* 1—indeed to one of the most famous adynata in Latin literature (*Ec.* 1.59-63; noted by Dümmler, 1881, *ad loc.*). Thus, knowledge of Virgil colors the reader’s understanding of the opening.

However, reading Paul's opening with Catullus provides a richer understanding of the themes of the poem. What the Muses desire in *Angustae Vitae* are cornerstones of Catullan diction. In particular, Paul's lines 2–3 recall lines 1–2 of Catullus' Poem 2 (Kroll, 1960) in which a woman plays with a pet sparrow: *Passer, deliciae meae puellae / quicum ludere . . . [solet]* ('O sparrow, the pet of my mistress, with whom [she is accustomed] to play . . .'). Paul's Muses desire to play in fields (*cupiunt sed ludere prata*) and to tend to (*colunt*) their delights (*delicias*), just as Catullus' lover is accustomed to play (*ludere solet*) with her pet (*deliciae*). *Ludere* ('to play'), furthermore, is a by-word in Catullus for both sexual activity and the production of poetry (see, for example, Poem 50.2).

The classical Muses, connected with elegiac, erotic love, are juxtaposed with the object of Paul's chaste, Christian love, a Benedictine Father. In line 9 he says that his heart blazes with love (*flagrat . . . amore*) for his friend. This expression, though recontextualized here, has classical sources, which color and give weight to the idea of love as the source of poetry. Virgil's Dido famously 'burned with love,' *flammavit amore* (*Aeneid* 4.54). Virgil does not, however, directly pair *flagrare* with *amor*. It is Catullus who provides the most telling consequences of 'blazing with love.' In Catullus' Poem 67, a man is described using just these words: *mens caeco flagrabat amore* ('his mind was blazing with a blind love,' 67.25). In Catullus' Poem 68, Laodamia enters her husband's house *flagrans amore* ('blazing with love,' 68.73). This immoderate love is highly problematic for Catullus: the poems cited imply that it leads, in the first case, to sexual violence and, in the second, to disastrous neglect for religious rites. For Paul, the opposite is the case. His love for the addressee is what allows him to attain the heavenly kingdom (11–12). Paul has taken a negatively charged trope in classical literature and made its ramifications positive by transposing it to a Christian context.

Angustae Vitae considers not just the creation of poetry, but specifically poetic exchange. In line 16, Paul asks the addressee to send him a poem in return: *redde . . . vicem*. This line has a parallel in Ovid, who says at *Amores* 1.6.23 (Kenney, 1961), *redde vicem meritis* ('give back in turn to those who deserve'). The phrasing characterizes literary creativity as a form of exchange. Examination of Catullus' Poem 50, also an epistolary poem on the subject of literary creativity, illuminates this practice. In this poem, addressed to his friend and fellow poet, Calvus, Catullus emphasizes the reciprocal character of poetic writing, also using the verb *reddo* ('return, give back, pay back'). Paul, like Catullus, is beseeching his addressee to participate in poetic exchange. Without a reply, Paul is *misero* ('wretched'). In Poem 50, Catullus describes himself as *me miserum* ('wretched me'). For both poets, being away from literary inspiration and without a reciprocal exchange of a poem leads to misery.

Paul offers his poems to his friend with disarming modesty: *suscipe . . . qualiacumque* (8, 'take them, such as they are'). The attitude and diction here recall Catullus Poem 1, in which he offers a small book to his friend Cornelius: *habe . . . qualecumque* (1.8–9, 'have [it], such as it is'). Yet in both cases this modesty is immediately undercut by a claim to a literary legacy. For Catullus, permanence depends on the Muse. In his final line he asks her, *o patrona virgo / plus uno maneat perenne saeclo* ('O patron maid, may [my book] remain enduring more than an age' 1.9–10). Paul, too, wishes to gain the heavenly kingdom *per aevum* (11, 'for an age'), and claims he will continue to praise his friend *tempus in omne* (20, 'for all time'). For Paul, however, the guarantor of permanence is not the Muse, but the *domini munus* (12, 'the gift of the Lord').

4 Methodology and Feature Sets for Intertextual Analysis

In order to assess the statistical evidence of Catullan influence in *Angustae Vitae*, we prepared a variety of texts for processing. For Paul's poems, our texts were transcribed from Dümmler (1881). All other Latin texts were retrieved from the *Tesserae* web site (Coffee, 2010). In addition to Paul and Catullus, we also processed texts from five other Latin poets. The *Elegies* of Tibullus and Propertius were chosen for their elegiac style, which is often reminiscent of Catullus. Ovid's *Amores* and Book Four of Virgil's *Aeneid* were chosen for their intertextual relationship to *Angustae Vitae*, described above. Horace's *Epistles* were chosen as a negative test, with no overt stylistic similarity to Catullus, nor any identified intertextual relationship to *Angustae Vitae*. Each feature set was generated using a series of Perl scripts written specifically for this work, with a supporting machine-learning package applied where necessary.

4.1 *Open Set Attribution*

In traditional authorship attribution (Diederich *et al*., 2003**)**, classification experiments have most often taken the form of *closed set* attribution, where all works considered are hypothesized to belong to some finite set of authors for which classifiers exist. An alternative *open set* approach is to presume that any work from any author can be considered for classification using a classifier for a known author. Unlike the closed set scenario, open set attribution is not constrained by any *a priori* assumptions made on the testing data. Open set attribution has been applied successfully to document classification (Manevit and Yousef, 2001), and attribution for English language novelists and Rabbinical Hebrew writing (Koppel and Schler, 2004; Koppel *et al.*, 2007). We can extend this idea to the study of stylistic similarity; our goal, however, is not to attribute authorship, but to attribute influence.

For this work, we want to test the stylistic similarity of any poet to Catullus. Thus, we require only a single influence classifier trained on representative samples from Catullus. Ideally, as depicted in Fig. 1, samples from poets with any stylistic similarity to Catullus (such as Tibullus and Propertius, fellow elegists of the late republic and early empire) will receive a positive score, while all others will receive a negative score. A one-class support vector machine trained on textual samples from Catullus that are closest in our feature space to his Poems 1 and 2, where much of the suspected intertext exists, provides us with the appropriate tool.

4.2 *The Functional n-gram Feature*

Our work in authorship and stylistic analysis has considered the importance of phonetic style markers, with the observation that sound plays a fundamental role in an author's style, particularly for poets. To capture sound information, we have developed a feature that we call a functional n-gram (Forstall and Scheirer, 2010), whereby the power of the Zipfian distribution (Zipf, 1949) is realized by selecting the n-grams that occur most frequently as features, while preserving their relative probabilities as the actual feature element. By using more primitive, sound-oriented features—namely, character-level n-grams—we are able to build accurate classifiers with the functional n-gram approach.

$$P(e_n|e_{n-N+1}^{n-1}) = \frac{C(e_{n-N+1}^{n-1}e_n)}{C(e_{n-N+1}^{n-1})} \iff freq(e_{n-N+1}^{n-1}e_n) > \phi$$

The formula above expresses the functional n-gram feature for any length *n*. For some character e_n, we would like to know the probability of its occurrence given a sequence of characters directly preceding it. We calculate this probability by counting all instances of the character string of length *n*, and dividing by the count of the character string of length *n*-1 preceding e_n. Most importantly, we consider the resulting probability a feature if and only if the frequency of the character string of length *n* exceeds some threshold φ. This threshold limits our selection of n-grams to those at the farthest left of a plotted Zipfian distribution (assuming the *x* axis is organized from most frequent to least frequent). This process is illustrated in Fig. 2.

4.3 *The Low-Probability Feature*

At the word level, we use a style marker for diction that is somewhat the opposite of the functional n-gram approach. Considering the Zipfian distribution once again, we turn to elements that occur with low probabilities. The power of functional n-grams relies on the amount of information carried by the elements at the left side of the Zipfian distribution. Now we consider the right—features that occur infrequently, though not necessarily *hapax legomena*.

We fix a desired probability range for words that occur infrequently, and look for n-gram sequences composed of only those words in a particular passage, ignoring all others:

$$(P_{low} < \mathrm{Pr}(\mathrm{word}_1) < P_{high}) \ldots (P_{low} < \mathrm{Pr}(\mathrm{word}_2) < P_{high}) \ldots (P_{low} < \mathrm{Pr}(\mathrm{word}_n) < P_{high})$$

where $n \geq 3$ in the above example. The probability of the resulting n-gram is compared to pre-computed probabilities of the same n-gram (should it exist) for specific authors, or literary groups. This type of style marker is very well suited to our case study, where certain word sequences are common to a particular group (Catullus and his elegiac successors), but uncommon or non-existent in the work of other groups.

To compute the actual feature, a count C_1 of the chosen n-gram sequence must be calculated. Following this, a second count C_2 must be calculated from all other n-gram sequences starting with word$_1$ and bounded by P_{low} and P_{high}. The final probability feature is given by: C_1/C_2. If the sequence exists, it will have a very low probability, otherwise, it receives a probability of 0; in most texts, the target n-gram should not exist. This feature can also be computed in a 'commutative' fashion (for a bi-gram: 'word$_1$ word$_2$' or 'word$_2$ word$_1$'). The resulting probability features can be used to augment the existing functional n-gram features to train a more accurate one-class SVM.

4.4 *The Metrical Feature*

A third feature type, derived from syllabic quantity, attempted to quantify larger scale patterns in the poets' use of meter. Paul wrote *Angustae Vitae* in elegiac couplets, in imitation of his classical predecessors, despite the fact that in the intervening centuries the prosody of spoken Latin had changed greatly. A couplet is composed of two lines of slightly different prescriptions; in each the number and quantities ('lengths') of syllables must conform to one of a very limited set of patterns (see Fig. 3). Barring rare exceptions, there are sixteen possible forms for the first line and four for the second. The quantity of a syllable depends upon the length of the vowel and the disposition of following consonants (including those in a following word). Elided syllables, pronounced weakly or not at all, are not counted as having any metrical quantity. Poems written in elegiac couplets may be as short as a single couplet or may comprise hundreds.

Although the number of possible combinations of long and short syllables in a couplet is limited, poets nevertheless do tend to show large-scale habits, the study of which has been a productive branch of scholarship in Latin poetry since antiquity itself (for example Platnauer, 1971). However, previous work in this line has concentrated on describing the corpus-scale differences between authors (thousands of lines), whereas here we hope to leverage meter in order to make some sense of similarities, and to consider patterns at a more local level. In this first attempt, the metrical feature was a simple one: having scanned a poem, we calculated functional n-gram frequencies for sequences of quantities just as described for characters in Section 4.2.

5 The Statistical Evidence of Catullan Influence in *Angustae Vitae*

5.1 *The Functional n-gram Analysis*

Using the functional n-gram feature, we processed Catullus' Poems 1–64, as this is where most of the suspected intertextual connections were identified. First, we examined the most frequently occurring character-level bi-grams in the Catullan poems (Table 1). The character-level bi-gram was chosen because it represents a sound primitive that also captures some word information, making it a highly descriptive feature. An examination of the ten most frequently occurring bi-grams revealed that while 'er' was the top bi-gram, 're' produced the highest probability (excluding the bi-gram 'qu,' which always produces 1), indicating a tighter coupling between characters. Further, 're' appears as an important sound element in several of the words from phrases common to both Catullus and *Angustae Vitae* (*ludere* and *redde*).

Considering just this ‘re’ bi-gram, we generated probabilities for all of Catullus 1–64, split into individual 20 line samples (with overlap between poems). This sample size was chosen in order to give us an ample amount of training data for our machine learning while still preserving some of the boundaries between poems. With much suspected similarity between Catullus 2 and *Angustae Vitae*, we chose the probability feature generated from Catullus 1 and 2 (each poem is 10 lines long) as our representative sample, and selected all other samples most resembling it (within +/- 0.06 from the representative sample; this range was defined empirically based on the samples available for Catullus) to form our training data. The complete training data for our classifier is shown in Table 2. Using the *libsvm* package (Chang and Lin, 2001) a one-class SVM was trained using a radial basis function kernel with $\gamma = 1$ and $\nu = 0.1$.

Results for the functional n-gram analysis confirm that there is a striking similarity between *Angustae Vitae* and the poetry of Catullus for one of the most common sound patterns found in each. Based on raw probability features alone, the 20 line sample composed of Catullus 1 and 2 yields: 0.458, while the 20 lines of *Angustae Vitae* yield: 0.486. Hence by using the ‘Catullan influence’ SVM, we achieved a positive classification for *Angustae Vitae*. The results for the other poets also satisfied our hypotheses. Each corpus from the remaining poets was also split into 20 line samples, with 40 random samples chosen for each (because of its smaller size, book 4 of the *Aeneid* only produced 35 samples in total) and processed accordingly. Samples from Tibullus and Propertius showed the most similarity to Catullus, while all other poets showed little similarity. A summary of all classification results is given in Table 3.

5.2 *The Low-Probability Analysis*

Using the low-probability analysis, we processed all of our texts once again, looking for the key word-level n-grams that represent the intertexts introduced in Section 2. Here, we express these n-grams using the following generalized templates: delic_ lu(de|si)_ (representing the notion of ‘delights and play’), fla(gr|mm)_ amor_ (representing the notion of ‘burning with love’), and redde_ miser_ poema_ (representing to the notion of ‘misery induced by the absence of poetic exchange’). Results for our search for these templates over our 20 line samples of text are shown in Table 4. New possible intertexts were found in both Propertius and Tibullus. We also note that we were able to find additional possible intertexts in Propertius 2.15, 4.4, and 4.7, even though these were not aligned with our sequentially generated 20 line samples. Our samples from Horace and Ovid did not contain any of the target sequence templates.

With the word-level n-grams from Table 4 that are common to most poets, we created two SVMs for delic_ lu(de|si)_ and fla(gr|mm)_ amor_. These SVMs use the training data of Table 2, augmented with the probability feature 0.043 from the sample containing Catullus 1 and 2 for delic_ lu(de|si)_, and the feature 0.024 from a sample containing the relevant portion of Catullus 67 for fla(gr|mm)_ amor_. The testing data from the functional n-gram experiments was once again used for this new set of experiments, with the addition of low-probability features computed over each sample. After applying our SVMs, we noted a reduction in positive classifications compared to the results in Table 3 for some poets. For the delic_ lu(de|si)_ SVM, Ovid, Horace, and Virgil each contributed one positive classification, while Propertius contributed five. For the fla(gr|mm)_ amor_ SVM, Virgil again contributed one positive classification. Table 5 summarizes these results. Based on these results, it appears likely that this feature does not ordinarily contribute to further positive classifications, but does help refine the results.

5.3 *The Metrical Analysis*

Metrical features were used only in comparing poems written in elegiac couplets, and so a slightly different selection of poems was used. Paul’s entire elegiac corpus was compared to similar-sized elegiac corpora by Catullus, Tibullus, Propertius, and Ovid (see Table 6). Samples were twenty lines each.

Frequencies were calculated for bi-grams and tri-grams composed of long and short syllable values. Because the features considered were composed of only two quantities, in cases a pairs of frequencies necessarily added up to 100%, for example, long-long-short and long-long-long. One member of such pairs was discarded. Furthermore, several patterns were not possible given the constraints of elegiac couplets, for example, long-short-long and short-short-short. All remaining features were considered individually to identify those which produced the best separation between authors. Principal components analysis was performed on the complete feature set.

Results from the metrical feature produced no useful refinement of the values from either of the other two features, and were largely incompatible due to the restricted data set. Nevertheless, considered separately, the results from this feature show some promise. Overall, the metrical n-gram features produced poor separation of the five poets considered (see Figs. 4 and 5). Inasmuch as authors could be distinguished, Paul was most different from Catullus, and most similar to Ovid. Paul varied internally more than the others.

Several of the variables seemed to vary roughly chronologically across the material sampled. Over the generation between Catullus (c. 84–c. 54 BCE) and Ovid (43 BCE–17 CE) we tend to see, for example, a trend downwards in the frequency of long-long-long. Indeed, the more refined critics of the Roman empire considered their republican forbears' extended sequences of long syllables ponderous. Paul (d. 799 CE) is later than rest by far longer than the period which they span, yet a proportional difference was not seen with respect to the metrical features examined here. The trend from Catullus to Ovid represents a continuous, living tradition of poetry performed in a spoken language, but by Paul's time the meter of poetry in contemporary, spoken Latin followed entirely different rules. Paul had to learn the tradition of classical elegiac poetry entirely from books. His access to works from different periods without the bias of one contemporary literary clique may contribute to the greater variance seen in his works.

6 Conclusions and Further Work

These results show how new computational methods can have value for the study of intertextuality: specifically, that many aspects of subjective philological analysis can be quantified for better comparison between studies, and that machine learning has the potential to draw out new aspects of stylistic influence and reference.

In our phonetic oriented analysis, the functional n-gram feature provided a tool for quantifying a notable stylistic similarity between Catullus' Poems 1 and 2 and *Angustae Vitae.* Our technique for computing word-level n-grams over disconnected sequences of low-probability words was able to refine our classification results. Our metrical analysis was not able to improve results further, and did not show a strong level of similarity between Catullus and *Angustae Vitae.* It did, however, identify a widely accepted trend in Latin poetics. We suspect that the inclusion of word boundaries as well as syllable quantities would greatly increase the power of this feature set. We also introduced the notion of open set recognition for influence recognition, which we believe has great potential value for this field.

This work is part of an ongoing effort to understand the relationship between the frequency of linguistic elements and the creation of literature. The metrics introduced here will be integrated into the University at Buffalo's *Tesserae* (Coffee, 2010), a new project which provides web-based users with intertextual search tools for Latin poetry.

Notes

1 It is generally believed that at the time of Paul's life, Catullus' works were unknown. The text of his poems seems to have survived antiquity in only a single manuscript, which was dramatically rediscovered in the thirteenth century after centuries of dormancy. While various scholars have since suggested that particular medieval texts contain 'echoes' of Catullan language (a host of examples are surveyed by Ullman, 1960), none of these potential links has been accepted widely as secure (Butrica 2007, 24).

Yet in Paul's epitaph for Arichis (Poem 33 in Dümmler, 1881), he writes (ll. 29–30), *Heu mihi, quam subito perierunt omnia tecum / Gaudia, prosperitas, paxque quiesque simul* ('Alas, how suddenly all my joys have perished with you: happiness, peace and tranquility together'). His words seem an incontrovertible parallel to Catullus' mourning for his dead brother (68.23): *Omnia tecum una perierunt gaudia nostra* ('All my joys have perished together with you'). This was noted recently by Sarah Jacobson using entirely non-digital means and is therefore not the focus of the present paper; pending her publication of a more detailed analysis, we take it for granted here that Catullus is a plausible influence on Paul's other poems.

Acknowledgments

We thank Joseph Pucci of Brown University and Marilyn Skinner of the University of Arizona for advice on Paul and Catullus. We thank Neil Coffee of the University at Buffalo for the use of *Tesserae*. Continued development of the methods described here is to be funded in part by the Digital Humanities Initiative at Buffalo under the aegis of its Textual Analysis Working Group.

References

Chang, C. and Lin, C. (2001). *LIBSVM: a Library for Support Vector Machines*, http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed 18 September 2010).

Coffee, N. (2010) (ed). *Tesserae: Intertextual Phrase Matching.* http://tesserae.caset.buffalo.edu (accessed 18 September 2010).

Butrica, J. (2007). History and Transmission of the Text. In Skinner, M. B. (ed), *A Companion to Catullus*. Malden, MA: Blackwell.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence*. **19**(1–2): 109–123.

Dümmler, E. (1881) (ed). *Poetae Latini Aevi Carolini*: *Tomus I.* Berlin: Weidmann.

Forstall, C. W. and Scheirer, W. J. (2010). Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, **1**(2). Chicago: The Division of the Humanities at the University of Chicago.

Harrington, K., Pucci, J., and Elliott, A. G. (1997). *Medieval Latin* (2nd ed.). Chicago: University of Chicago Press.

Kenney, E. J. (1961) (ed). *P. Ovidi Nasonis: Amores; Medicamina Faciei Femineae; Ars Amatoria; Remedia Amoris.* Oxford: Clarendon Press.

Koppel, M. and Schler, J. (2004). Authorship Verification as a One-Class Classification Problem. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, pp. 489–495.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research* **8**: 1261–1276.

Kristeva, J. (1986). Word, Dialogue and Novel. In Moi, T. (ed.), *The Kristeva Reader*, New York: Columbia University Press, pp. 34–61.

Manevit, L. and Yousef, M. (2001). One-Class SVMs for Document Classification. *Journal of Machine Learning Research* **2**: 139–154.

Mynors, R. A. B. (1969) (ed). *Publi Virgili Maronis: Opera*. Oxford: Clarendon Press.

Platnauer, M. (1951). *Latin Elegiac Verse: A Study of the Metrical Usages of Tibullus, Propertius and Ovid*. Cambridge: Cambridge University Press.

Trillini, R. H. and Quassdorf, S. (2010). A 'Key to All Quotations'? A Corpus-Based Parameter Model of Intertextuality. *Literary and Linguistic Computing* **25**(3): 269–286.

Ullman, B. L. (1960). The Transmission of the Text of Catullus. *Studi in onore di Luigi Castiglione* **2**: 1027–1057. Florence:Sansoni.

von Kroll, W. (1960) (ed). *C. Valerius Catullus*. Stuttgart: Teubner.

Zipf, G. (1949). *Human Behavior and the Principle of Least-effort*. Cambridge: Addison-Wesley.

Tables

Table 1 The ten most frequently occurring character-level bi-grams in Catullus' Poems 1–64

Rank	bi-gram	Probability	Rank	bi-gram	Probability
1	er	0.180	6	te	0.251
2	qu	1.000	7	es	0.146
3	is	0.169	8	um	0.164
4	en	0.162	9	in	0.140
5	**re**	**0.275**	10	it	0.133

Table 2 Features used to train the 'Catullan influence' one-class SVM. Each numerical feature represents the character-level bi-gram 're' for a poem or pair of poems (depending on poem length, for sample consistency).

Feature	Poems	Feature	Poems
0.458	1 and 2	0.435	38 and 39
0.412	2b and 3	0.480	44 and 45
0.455	4 and 5	0.480	50 and 51
0.524	7 and 8	0.444	62
0.500	13	0.463	64
0.406	17 and 21	0.464	64

Table 3 Results for the functional n-gram analysis expressed as samples classified positively with Catullus out of all samples

More Like Catullus		**Less Like Catullus**	
Text	**Positive Class.**	**Text**	**Positive Class.**
Angustae Vitae	**1/1**	Ovid *Amores*	2/40
Propertius *Elegies*	6/40	Horace *Epistles*	3/40
Tibullus *Elegies*	5/40	Virgil *Aeneid*	2/35

Table 4 Computed low-probability features from the word-level n-grams made up of our three intertextual sequence templates. Entries in bold represent new possible intertexts discovered by our automated analysis.

Word-level n-gram	Probability	Source
delic_ lu(de|si)_	0.043	Catullus 2
delic_ lu(de|si)_	0.005	Catullus 50
delic_ lu(de|si)_	0.013	Catullus 50
fla(gr|mm)_ amor_	0.024	Catullus 67
fla(gr|mm)_ amor_	0.028	Catullus 68
redde_ miser_ poema_	0.015	Catullus 50
delic_ lu(de|si)_	0.014	*Angustae Vitae*
fla(gr|mm)_ amor_	0.020	*Angustae Vitae*
redde_ miser_ poema_	0.018	*Angustae Vitae*
delic_ lu(de|si)_	**0.019**	**Propertius 2.34b**
fla(gr|mm)_ amor_	**0.027**	**Propertius 3.8**
fla(gr|mm)_ amor_	**0.014**	**Propertius 3.19**
fla(gr|mm)_ amor_	**0.023**	**Propertius 2.34b**
fla(gr|mm)_ amor_	**0.125**	**Tibullus 1.9**
fla(gr|mm)_ amor_	**0.143**	**Tibullus 2.4**
fla(gr|mm)_ amor_	**0.080**	**Tibullus 3.12**
fla(gr|mm)_ amor_	0.033	Aeneid 4

Table 5 Results for the low-probability analysis expressed as samples classified positively with Catullus out of all samples. The results here show improvement over those in Table 3.

More Like Catullus		**Less Like Catullus**	
Text	**Positive Class.**	**Text**	**Positive Class.**
Angustae Vitae	**1/1**	Ovid *Amores*	1/40
Propertius *Elegies*	5/40	Horace *Epistles*	1/40
Tibullus *Elegies*	5/40	Virgil *Aeneid*	1/35

Table 6 Texts used for metrical analysis

Author	**Works**	**Total lines**
Paul the Deacon (as numbered in Dümmler, 1881)	Poems 2, 4, 5, 9, 10, 12 (lines 37–42), 14, 16, 19, 21–24, 26 (*ll.* 4–13), 27–33, 36, 37, 39, 40, 42, 43, 48, 50, 56 (*ll.* 38–41)	755
Catullus	Poems 65–116	637
Tibullus	*Elegies* Book 1	804
Propertius	*Elegies* Book 1	703
Ovid	*Amores* Book 1	767

Figures

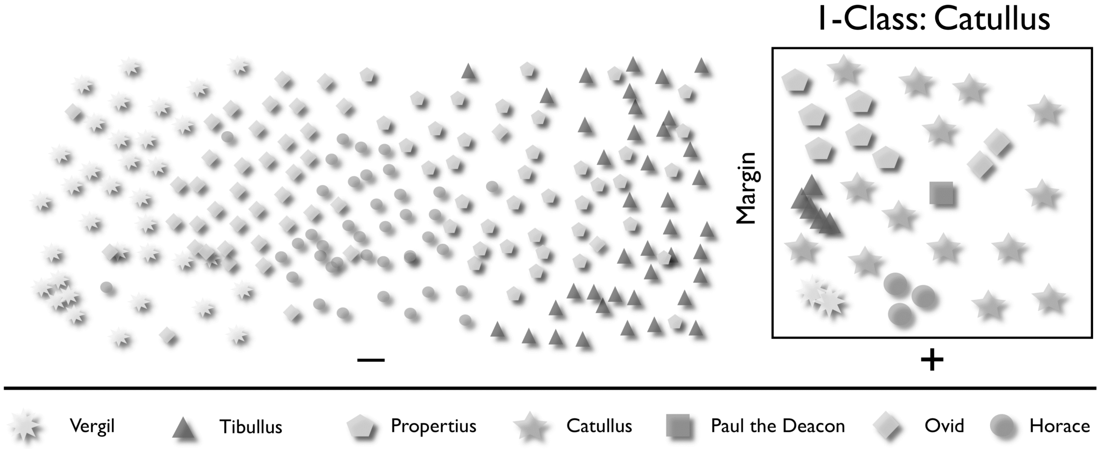

Fig. 1 In authorship attribution, we typically consider a closed set problem. Here, we consider an open set problem, where we want to test the stylistic similarity of any Latin poet to Catullus. A one-class SVM trained on textual samples from Catullus that are closest in our feature space to Poems 1 and 2 provides us with the appropriate tool. Actual results for our functional n-gram analysis are shown.

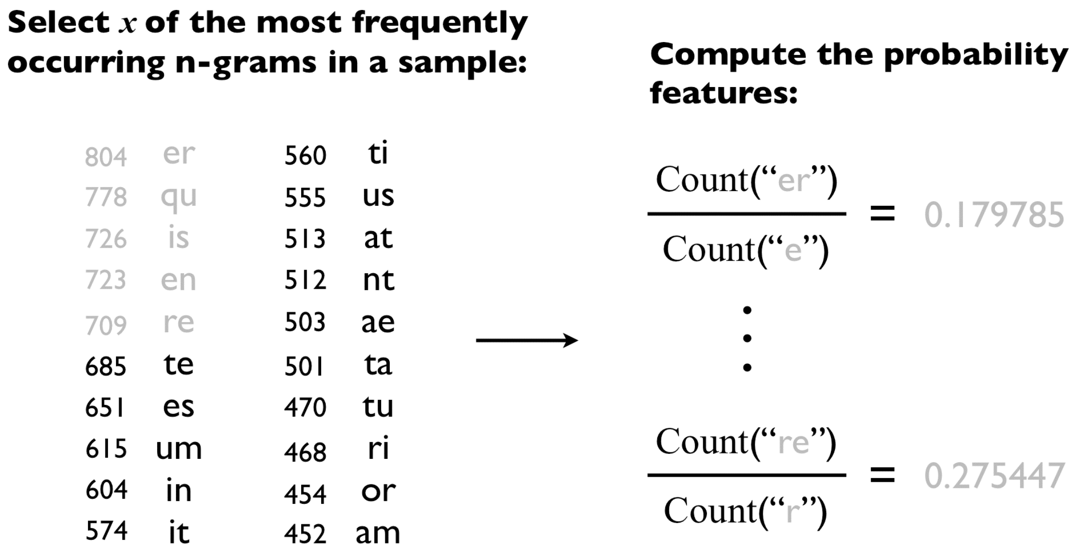
Select x of the most frequently occurring n-grams in a sample:
804 er
778 qu
726 is
723 en
709 re
685 te
651 es
615 um
604 in
574 it
560 ti
555 us
513 at
512 nt
503 ae
501 ta
470 tu
468 ri
454 or
452 am
Compute the probability features:
Count("er") / Count("e") = 0.179785
Count("re") / Count("r") = 0.275447

Fig. 2 An overview of the functional n-gram process, which is designed to capture and express the most frequent information in a sample of text. The above example is for characters, though the same process can be used for words and metrical syllables as well.

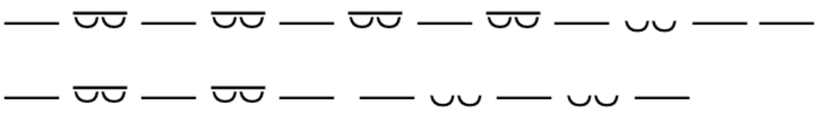

Fig. 3 The metrical quantities of syllables in an elegiac couplet. The couplet has two lines, one slightly longer than the other. In each, the quantity (length) of syllables is restricted to a few patterns. In this schematic diagram, – represents an obligatory long, ˘ an obligatory short, and ⏗ the poet's choice of a long or two shorts.

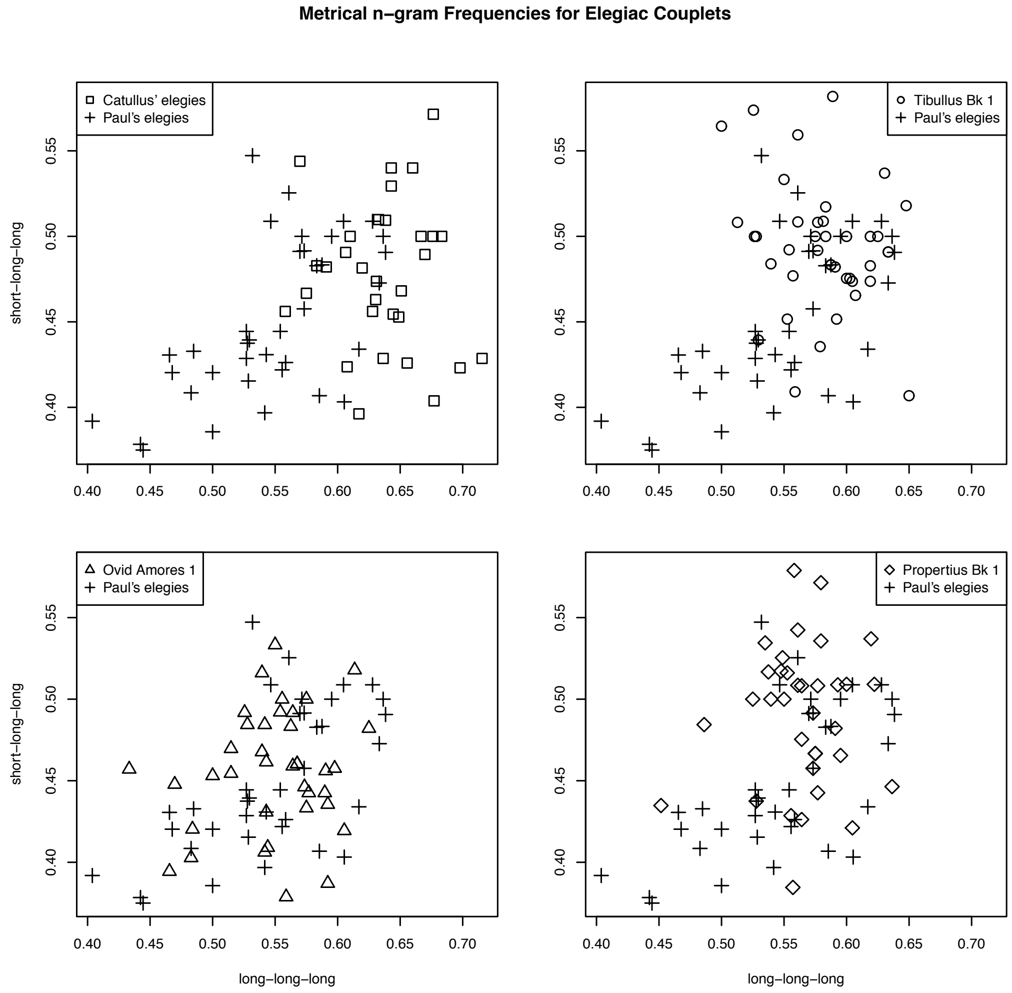

Fig. 4 Frequencies of two metrical tri-grams based on syllable weights in elegiac couplets. Paul the Deacon is compared with each of Catullus, Tibullus, Propertius, and Ovid. The y-axis gives the frequency with which the sequence short-long is followed by a long; the x-axis, the frequency with which long-long is followed by long. The scale and the data for Paul are the same in each.

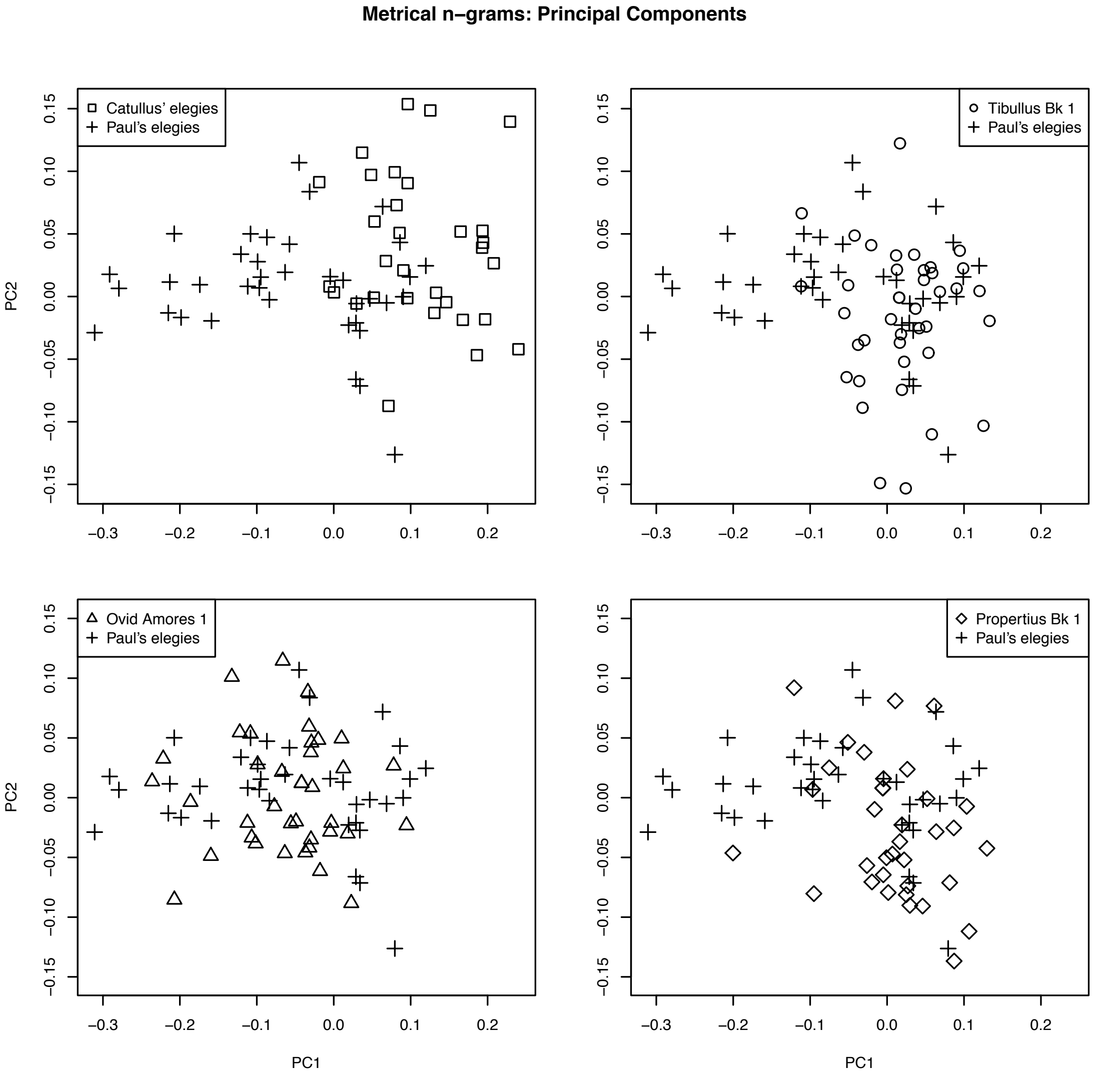

Fig. 5 First two principal components calculated from the complete set of metrical n-gram frequencies, as in Fig. 4