

Motivation

The study of intertextuality, the shaping of a text's meaning by other texts, remains a laborious process for the literary critic. Kristeva¹ suggests that "Any text is constructed as a mosaic of quotations; any text is the absorption and transformation of another." The nature of these mosaics is widely varied, from direct quotations representing a simple and overt intertextuality, to more complex transformations that are intentionally or subconsciously absorbed into a text.

Since, in many cases, the problem is one of pattern recognition, it is a good candidate for automated assistance by computers.

As a case study for our computational analysis of intertextuality, we turn to Paul the Deacon's 8th century poem *Angustae Vitae*, which we suggest has a strong connection to the poetry of Catullus.

The Clues: *Angustae Vitae* and Catullus

Although Paul the Deacon posits the classical and monastic worlds as opposites, the use of Catullan diction and models of poetic exchange recalls the paradigm of the Neoteric, proto-elegiac lover, his beloved, and his poetological concerns:

Catullus 2

PASSER, **deliciae** meae puellae,
quicum **ludere**, quem in sinu tenere,
cui primum digitum dare appetenti
et acris solet incitare morsus
cum desiderio meo nitenti
carum nescio quid lubet iocari,
credo ut, cum gravis acquiescet ardor,
sit solacium sui dolaris,
tecum **ludere** sicut ipsa possem
et tristis animi levare curas!

Angustae Vitae, lines 1-4:

Angustae vitae fugiunt consortia Musae,
Claustrorum septis nec habitare volunt,
Per rosulenta magis **cupiunt sed ludere prata**,
Pauperiem fugiunt, **deliciasque colunt**:



Angustae Vitae is peppered with classical intertexts, but it remains an open question (the indications of influence to this point have been "rare, faint, and probably indirect") as to whether Paul the Deacon had read Catullus. Thus, to build out case, we ask:

What similarities exist in sound, word frequency, and metrical form?

The Functional n-gram Analysis

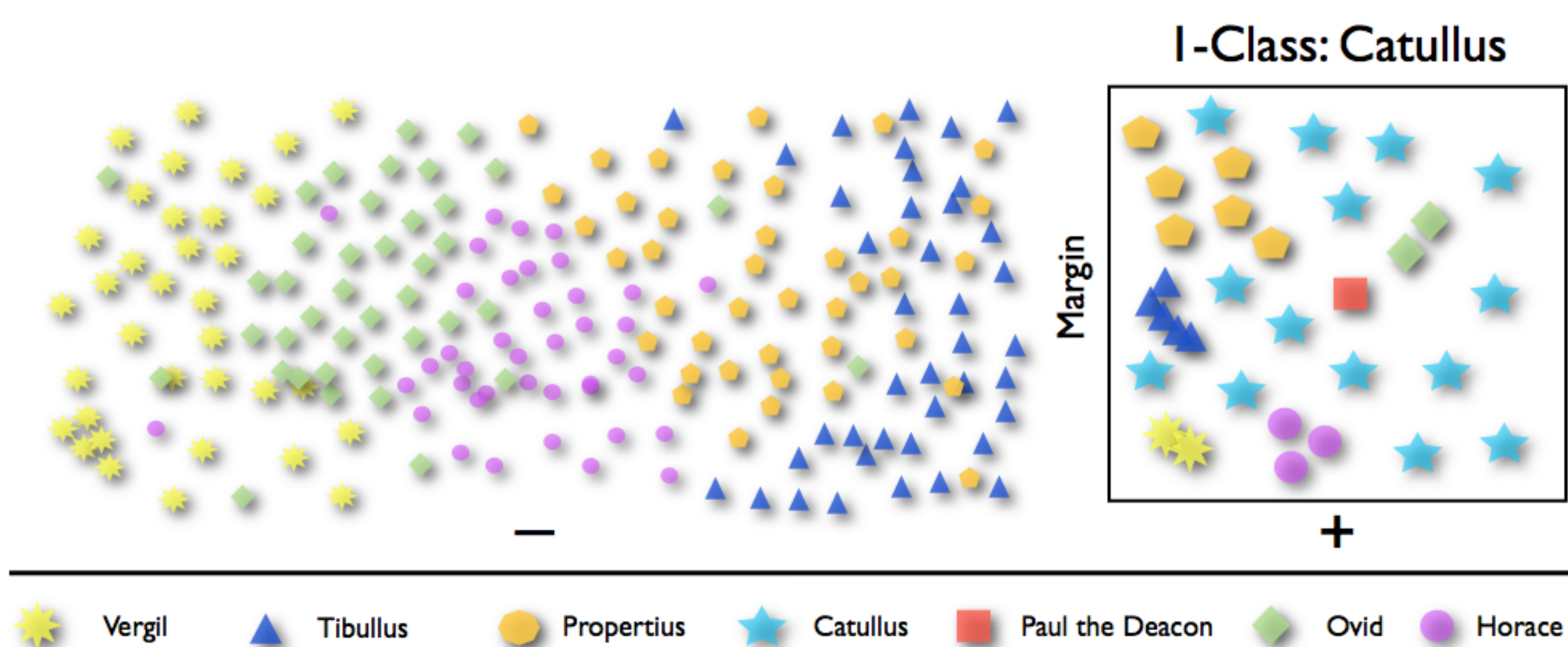
Observation: Sound plays a fundamental role in an author's style, particularly for poets.

The functional n-gram³ is a feature for stylistic analysis, whereby the power of the Zipfian distribution is realized by selecting the n-grams that occur most frequently as features, while preserving their relative probabilities as the actual feature element.

Feature: The Functional n-gram:

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

In this work, we consider primitive sound elements as functional character level bi-grams.



In authorship attribution, we typically consider a closed set problem. Here, we consider an open set problem, where we'd like to test the stylistic similarity of any author with Catullus. A 1-Class SVM⁴ trained on textual samples from Catullus that are closest in our feature space to poems 1 & 2 provides us with the appropriate tool.

Low-Probability Analysis

Observation: Words that occur infrequently are often intertexts.

By fixing a desired probability range for words that occur infrequently, and scanning for n-gram sequences composed of only those words in a particular passage, ignoring all others, we gain an additional feature to augment our functional n-gram vector.

$$(P_{low} < \Pr(\text{word}_1) < P_{high}) \dots (P_{low} < \Pr(\text{word}_2) < P_{high}) \dots (P_{low} < \Pr(\text{word}_n) < P_{high})$$

Key n-gram sequences common to *Angustae Vitae* and Catullus:

delic(ias)lae ludure flagra(t)ns(bat) amor_ redde miser poema

Metrical Analysis

Metrical considerations, including the alternation of heavy and light syllables within a line, are essential in examining poetic style⁵.

For example, *Angustae Vitae* is composed in elegiac couplets, which regularly take the following form:

— — — — — — — — — —
— — — — — — — — — —

Many of Catullus' poems, including 1 and 2, are in hendecasyllabic lines:

— — — — — — — — — —

Īnmōdīcō flāgrāt dē vēstrō pēctūs āmōrē,

Crēdē pātēr, nōstrūm, sēmpēr āmāndē mīhī. (9–10)

Cuī dōnō lēpīdūm nōvūm libēllūm (1.1)

Using TEI XML, we add the metrical values to the text. This provides flexibility in extracting hybrid feature sets. We calculate bi- and tri-gram frequencies for the metrical values as for characters in the text. A tiny set of high-frequency features accommodates smaller samples.

The Evidence

For our experiments, we considered source text from *Poetae Latini Aevi Carolini*⁶, *Medieval Latin*⁷ (2nd ed.), and *The Tesserae Project*⁸. Key character-level n-gram considered: "re".

Functional n-gram Analysis:

12 training samples (20 lines) most resembling Catullus 1 & 2 (-0.06, +0.06). RBF kernel. *Angustae Vitae* receives a positive classification, indicating stylistic similarity to Catullus. Raw probability features: *Angustae Vitae*: 0.486 Catullus 1 & 2: 0.458

For comparison, we also tested other poets, including those inspired by the Latin Neoterics. The table below reflects the number of samples classified as being stylistically similar to Catullus 1 & 2.

Ovid <i>Amores</i> 2/40	Horace <i>Epistles</i> 3/40	Vergil <i>Aeneid</i> (book 4) 2/35
Tibullus <i>Elegies</i> 5/40	Propertius <i>Elegies</i> 6/40	<i>Angustae Vitae</i> 1/1

Low-Probability Analysis:

Serves as a refinement for the base learning results, removing fringe candidates:

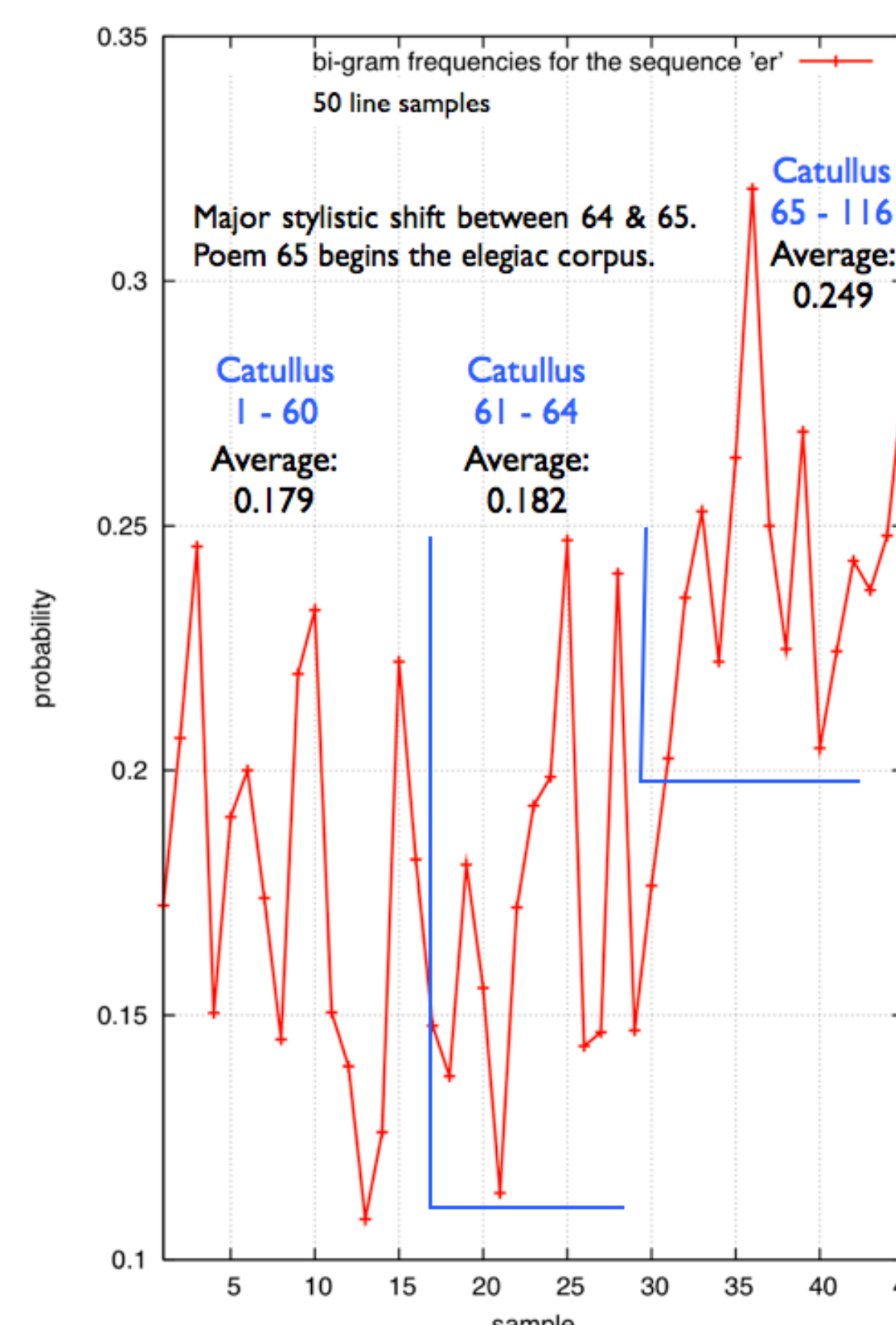
Ovid <i>Amores</i> 1/40	Horace <i>Epistles</i> 1/40	Vergil <i>Aeneid</i> (book 4) 1/35
Tibullus <i>Elegies</i> 5/40	Propertius <i>Elegies</i> 5/40	<i>Angustae Vitae</i> 1/1

Metrical Analysis:

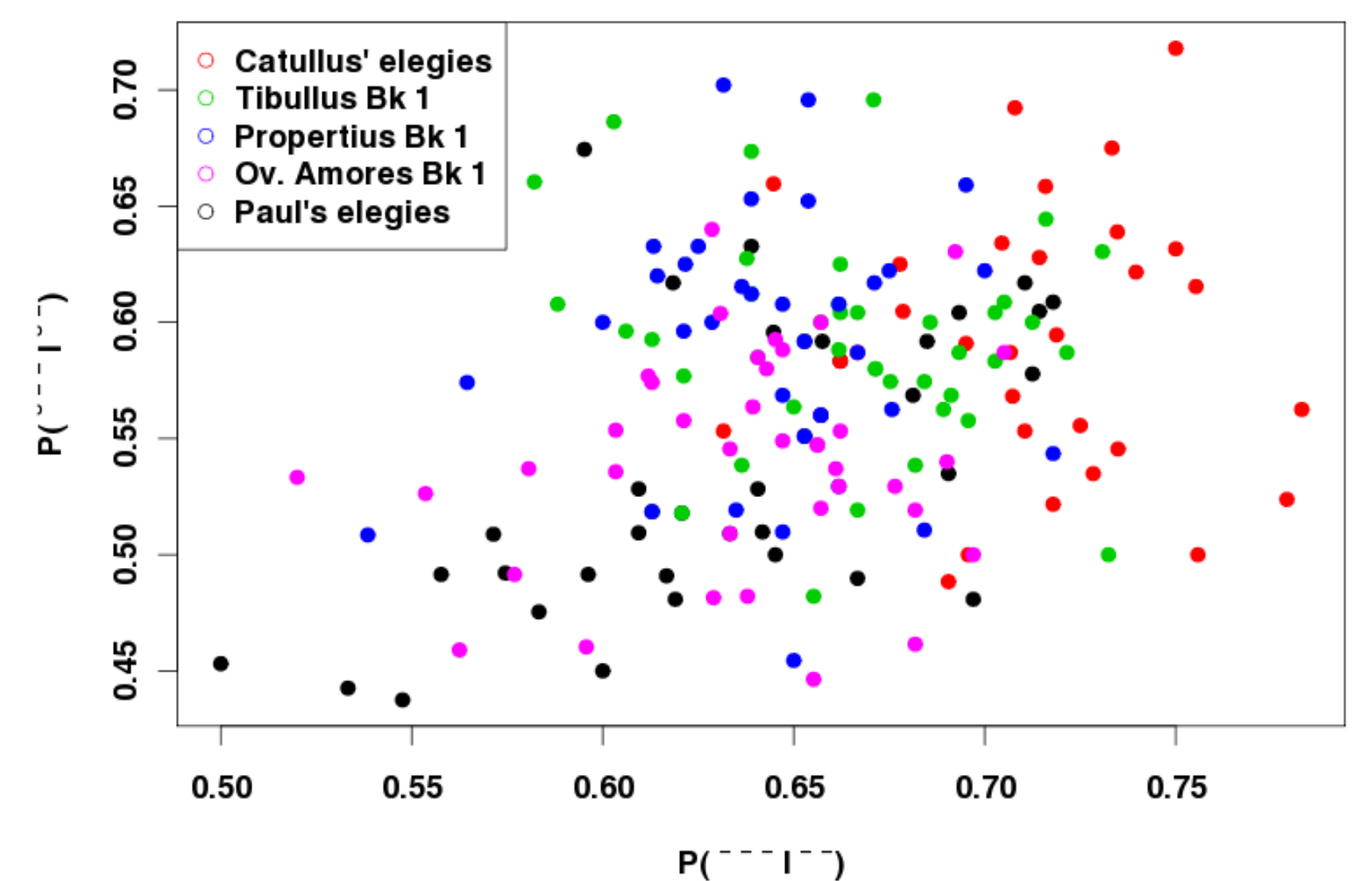
Bi-gram frequencies ("—" & "—") for Catullus proved to have too much variation for good training:

1:0.4583 2:0.4254 3:0.7564	1:0.5 2:0.4071 3:0.7500	1:0.48 2:0.3750 3:0.7467
1:0.4118 2:0.4191 3:0.7595	1:0.4063 2:0.4383 3:0.7018	1:0.4444 2:0.1768 3:0.5000
1:0.4545 2:0.5429 3:0.8636	1:0.4348 2:0.4259 3:0.9459	1:0.4634 2:0.1947 3:0.5000
1:0.5238 2:0.3896 3:0.8696	1:0.48 2:0.3758 3:0.9130	1:0.4643 2:0.1771 3:0.5000

Other Discoveries



While the metrical analysis did not show similarity between Catullus and *Angustae Vitae*, it does help to show the direction of literary development.



The X axis is the probability of "---", given "---"; The Y axis is the probability of "---", given "---".

References

- J. Kristeva, "Word, Dialog and Novel," in Toril Moi, ed., *The Kristeva Reader*, New York: Columbia University Press, 1986, p. 34-61.
- G. Conte, *Latin Literature: A History*, Transl. by J. Solodow, rev. by D. Fowler and G. Most, Johns Hopkins University Press, 1994.
- C. Forstall and W. Scheirer, "Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound," Chicago Colloquium on Digital Humanities and Computer Science, 2009.
- L. Manevit and M. Yousef, "One-Class SVMs for Document Classification," *Journal of Machine Learning Research* (2), pp. 139-154, 2001.
- M. Platnauer, *Latin Elegiac Verse: A Study of the Metrical Usages of Tibullus, Propertius & Ovid*. Cambridge University Press, 1951. See for example pp. 36-39.
- E. Dümmler, ed., *Poetae Latini Aevi Carolini: Tomus. I*. Berlin 1881, pp. 35-86.
- K. Harrington, J. Pucci, and A.G. Elliott, *Medieval Latin* (2nd ed.), University of Chicago Press, 1997.
- N. Coffee, J. Koenig, S. Poornima, and C. Forstall, *The Tesserae Project*, <http://tesserae.caset.buffalo.edu>