

Promises from an Inferential Approach in Classical Latin Authorship Attribution

Giulio Tani Raffaelli¹

¹*Institute of Computer Science, Czech Academy of Sciences, Czech Republic*

Abstract

Applying stylometry to Authorship Attribution requires distilling the elements of an author's style sufficient to recognise their mark in anonymous documents. Often, this is accomplished by contrasting the frequency of selected features in the authors' works. A recent approach, CP2D, uses innovation processes to infer the author's identity, accounting for their propensity to introduce new elements. In this paper, we apply CP2D to a corpus of Classical Latin texts to test its effectiveness in a new context and explore the additional insight it can offer the scholar. We show its effectiveness on a corpus of classical Latin texts and how—moving beyond maximum likelihood—we can visualise the stylistic relationships and gather additional information on the relationships among documents.

Keywords

authorship attribution, inference, classical Latin, visualisation

1. Introduction

Despite the development of AI tools [2], the current state of their interpretability [16] and the need for transparent, versatile approaches to stylometry sustain the continued development and use of tool based on supervised feature selection [5, 1]. These can be general purpose [10, 7] or specific for the Latin language [5]. In philology, where the ground truth on the authorship is out of reach, knowing why the style of a document is close to an author's may be more interesting than the author's name.


A common practice [1, 12, 11] to gain an understanding of the relationships among documents is to project them on a 2-dimensional space. This allows to visualise the relative positions of documents or corpora. For example, one such common approach is Correspondence Analysis [3], which is now included in multiple R packages and tools as Hyperbase.¹ This allows to visualise at the same time the relationships among documents and how different elements (e.g., words or lemmas) contribute to the positioning.

The recently proposed CP2D approach [18] applies information theory and innovation processes to propose authorship. Each author is modelled as an information source emitting tokens and characterised by the token frequency in their samples and their tendency to innovate. The representation as a Poisson-Dirichlet process allows to estimate the likelihood that a given author produced the anonymous document. The attribution then follows a Maximum Likelihood

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

✉ tani@cs.cas.cz (G. Tani Raffaelli)

ORCID 0000-0003-0866-5210 (G. Tani Raffaelli)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://hyperbase.unice.fr/>

approach. If the anonymous text is split into fragments, the researchers either compute the text likelihood by assuming the fragments are independent or let each fragment cast a vote (Majority Rule). The authors test the approach on literary prose in three languages and informal English texts. This approach is interesting as it is transparent and can be applied without relying on language tools—e.g., lemmatisers or large pre-trained models—whose quality can vary dramatically from language to language.

While this approach is proven effective, its basic formulation does not fully exploit its capabilities. Although the model has few hyperparameters, in the case of small corpora, optimising the hyperparameters based on the best performance on known texts risks overfitting. In the same paper, the performance on the test set for the smallest corpus is considerably lower than on the training set, while on large corpora, it tends to be stable or increase [18, Table 1]. Also, while using lemmatisers is not necessary, this could still help overcome data sparsity when the corpus is small. On a different note, even in the case of dubious attribution, the likelihoods produced by CP2D can offer further insight. The actual distribution of the likelihood values can help assess the relative position of the document of disputed attribution. This paper aims threefold: testing the application of the CP2D to Classical Latin poetry, dealing with the risks of overfitting, and propose a projection to examine the model output directly.

2. Results

The first promising result is the correct attribution of at least 34 out of the 36 documents in the corpus when following the method used in [18]. We say “at least” because—in the method described by its proponents—no procedure is suggested to choose among different sets of hyperparameters, all sharing the same micro-averaged recall on the training data. For one fourth of the documents, the same maximum is obtained with at least 15 different hyperparameter sets. Lacking a way to select a single one based on the training corpus increases the risk of overfitting (selecting a parameter that is effective on the training set but not on the test).

Considering all sets of hyperparameters that offer performances comparable to the maximum (see Methods section) and choosing the most common author, the correctly attributed documents are 35. This requires accounting for a relevant fraction of all hyperparameters tested. The only document not assigned to its canonical author is the *Halieutica* from Ovid, whose authorship is indeed is debated [8, Chap. 12]. In most cases, the first author is selected with more than 70% of the hyperparameter sets, replacing the potential instability of the simple attribution with a clear rule (see Fig. 1, panel C). Moreover, these results are comparable to the baseline imposters method [14]. The size and relative simplicity of the corpus do not allow to claim a significant difference.

A second observation is that, while CP2D does not require the use of lemmatisers, we find that—in this corpus—the use of sequences of lemmas instead of words increases the number of sets of hyperparameters that offer performances comparable to the maximum up to one-fifth of all parameters tried (see Fig. 1, panel A). At the same time, simply relying on any set of parameters that gives the best attribution gives 33 correct attributions. In this case, one fourth of the documents has at least 22 best-performing hyperparameter sets. These changes, possibly due to reduced sparsity, require additional care in identifying which parameters to

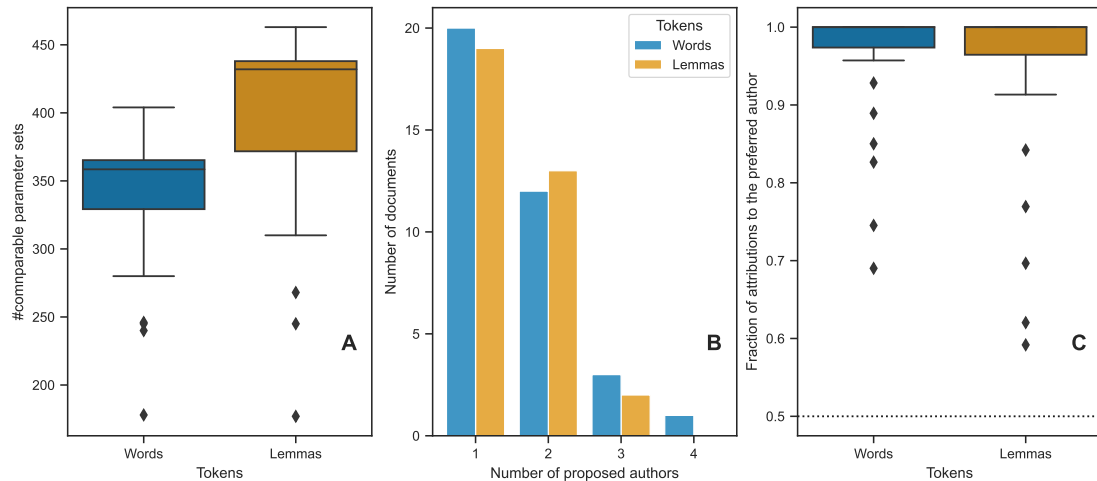


Figure 1: Statistics of the attribution accounting for multiple equivalent sets of hyperparameters in CP2D. **A:** Distribution of the number of equivalent sets of hyperparameters changing the definition of the tokens. **B:** Distribution of the number of authors selected by at least one set of hyperparameters for each document. **C:** Stability of the attribution of each document to its preferred author while changing sets of hyperparameters. The boxes span from the first to the third quartile of the distribution, and the whiskers extend to the last point that is no more than 1.5 IQR from the nearest box edge. The bar across the boxes marks the median.

trust. However, for most documents with both kinds of tokens, a single author is identified as the most likely with all sets of hyperparameters, Fig. 1, panels B. Also, even when more than one author is proposed, most hyperparameter sets usually select the same one, Fig. 1, panel C. Considering lemmas, the baseline method has two misclassified documents, the *Heroides* and the *Consolatio ad Liviam* for which Seneca is preferred.

The method so far has two ways to offer better insight into the position of each document relative to the candidate authors: how often each author is selected with different sets of hyperparameters or—for each set of hyperparameters in the model—the relative likelihood of the authors or the number of fragments assigned to each one. However, these methods do not allow for the easy accounting of more documents at once. Here, we try to overcome this issue by projecting the documents on a hyper-sphere where the relationship between texts and authors and among texts are encoded as angles.

In Fig. 2, we show the positioning of the documents and texts of uncertain attribution in our corpus and a sample document. In every plot, we show the position of all documents of the three closest authors.

We notice how we can have different scenarios with a shared message. The attribution of *Halieutica* is either barely to Ovid or to Horace (panels A and B), but other documents from all authors are always distant from it. The attribution can remain correct over the full range of variation of the Macro Recall (panels C and D). The difference in recall is driven by documents of other authors crossing attribution boundaries but does not affect stable attributions. A document can change attribution even with the same Macro Recall on the training set (panels G and H). However, even in panels F and G, when the anonymous document crosses the boundary

Table 1

Correctly attributed documents varying the definition of the tokens and the attribution procedure. For the Naïve CP2D we report the minimum number of correct attributions (see text). The baseline imposters method is applied to the 2000 most common features using the Wurzburg Delta.

	Nearest Author	Author of nearest doc.	Naïve CP2D	Imposters
Words	35	35	34	35
Lemmas	35	35	33	34

towards a different author, it remains close to the books of the actual author more than to any other. This suggests that assigning the documents to the author of the nearest document could give better results if cases like panels F and G were common. However, on this corpus, there is no noticeable difference.

Figure 2 shows that the *Halieutica* seems far from all the authors in our corpus, while the *Heroidum Epistulae* and the *Consolatio ad Liviam* are well integrated into the Ovidian production. We also observe how different books from the same collection tend to be grouped. On a similar note, we can read that while the fourth book of Propertius seems close in style to its reported author, it may be the least typical in its author’s production. Perhaps unsurprisingly [6], even removing the subdivision in verses for poetical documents, the works in prose from Seneca form a well-isolated group, and none of the other documents is ever attributed to Seneca with any choice of parameters. Less obvious is that using lemmas and ignoring the *Heroides*, no author is proposed outside Ovid, Propertius and Tibullus for all documents in elegiac distich.

These observations suggest that closeness between documents in this space is a good proxy for stylistic similarity. At the same time, being closer to the bottom or one of the top corners of the graphs indicates similarity to that author.

3. Discussion

We showed that the CP2D approach is also effective on Classical Latin texts, even on a small and imbalanced corpus. We showed that it is possible to increase the stability of the results by accounting for the many equivalent sets of hyperparameters and that using lemmas instead of words expands the subspace of hyperparameters where CP2D has high accuracy. We also showed how a suitable projection of the documents gives a meaningful representation of the relationships among documents. This representation can offer insight into the stylistic properties of the documents. Lastly, we proposed different approaches to attribution leveraging the distances among documents.

However, this corpus proved to be simple, and it is not possible to judge if some of the proposed alternatives (use of lemmas, authorship of the nearest document) would have a positive effect on corpora where the simple majority vote over the equivalent set of hyperparameters is not satisfying. A future challenge will be attributing not entire books but individual poems. This tougher challenge—some poems are only a few tens of words long—is of greater interest as often—e.g., is the case of the *Heroidum Epistulae*—the uncertainty in attribution is mainly

Table 2

Documents included in the corpus. All but one lemmatised documents were downloaded from Hyperbase as part of the LASLA collection [15]. The lemmatisation of the *Consolatio ad Liviam* was provided by Noemi Daria Zaccagnino as personal communication.

Author	Title	Short Title
Catullus	<i>Carmina</i>	Catul1-3
Horace	<i>Carmen Saeculare</i>	HorSaecu
	<i>Carmina</i> 1-4	HorCarm1-4
	<i>Amores</i> 1-3	OviAmor1-3
	<i>Ars Amatoria</i> 1-3	OviArsA1-3
	<i>Fasti</i> 1-6	OvFasti1-6
Ovid	<i>Halieutica</i>	OviHalie
	<i>Heroidum Epistulae</i>	OviEpist
	<i>In Ibin</i>	Ovilbin
	<i>Medicamina Faciei Feminae</i>	OviMedic
	<i>Remedia Amoris</i>	OviRemed
Propertius	<i>Consolatio ad Liviam</i>	OviConsLiv
	<i>Elegiae</i> 1-4	Propert1-4
	<i>Ad Helviam Matrem De Consolatione</i>	SenHelvi
Seneca	<i>Ad Marciam De Consolatione</i>	SenMarci
	<i>Ad Polybium De Consolatione</i>	SenPolyb
Tibullus	<i>Elegiae</i> 1-3	TibEleg1-3

on selected poems [17].

4. Methods

We selected a corpus of 34 documents from six different authors writing in Classical Latin for this work. Poems in elegiac distich form the main part of the corpus (works by Ovid, Propertius and Tibullus), followed by other works in various metres (works by Horace and Catullus) and three examples of prose from Seneca (*Consolationes: Ad Marciam, Ad Helviam matrem, Ad Polybium*). We designed the corpus to be imbalanced (the works of Ovid comprise half of the documents) and divided into literary genres that we expect to challenge the attribution to different extents. The *Consolationes* in prose from Seneca might show similarities with the Ovidian text of similar topic. Moreover, the corpus contains four documents considered entirely or partially from a different author. These are the third book of the *Elegiae* from Tibullus [8, Chapters 8-11] and Ovid’s *Halieutica* [ibid., Chapter 12-13], the *Consolatio ad Liviam* [ibid., Chapter 14] and *Heroidum Epistulae* [ibid., Chapter 15]. The lemmatised sequences are publicly available in the LASLA collection from the University of Liège [15]; with the exception of the *Consolatio as Liviam*; see Table 2 for the complete list of the documents included. Despite the known relevance of morphosyntactic annotations [9], for this work, we took into consideration only the lemmas. We executed the entire analysis in Python, using standard packages (numpy, scipy) and the cp2d module from [19]. The code is available at <https://github.com/GiulioTani/CHR24>.

We prepared the texts, removing the separation in verses. The distinction between ‘u’ and

‘v’ is already removed in the documents, and we removed the distinction between upper- and lower-case letters and all the non-alphabetic characters (i.e., punctuation). We considered the sequences of words and lemmas and of N -grams with $N \in [3, 6]$ for both. Note that the built-in definition of N -grams in CP2D is derived from [13] and allows a space to appear only at the beginning or at the end of the N -gram. This definition excludes words or lemmas shorter than $N - 2$ (accounting for spaces at both ends).

As a baseline method, we used the imposters [14] approach built-in in Stylo [7], using the top 2000 character 4-grams and the Wurzburg delta.

We followed a nested leave-one-out paradigm to evaluate the CP2D’s performance. This is because it works better maximising the size of the training corpus and most authors in the corpus have only 3-5 documents. All the results are obtained by excluding one document at a time and treating it as anonymous. Then, we optimise the model hyperparameters, maximising the attribution in a new leave-one-out experiment. Finally, we evaluate the attribution of the left-out document. This procedure requires each author to have at least three documents. To this end, we split the book of Catullus into three parts containing 39 carmina each, in order of appearance. The final corpus contains 36 documents.

The simplest approach to attribution requires searching—for every document—the set of hyperparameters that maximises the attribution on the remaining corpus. We followed the authors in [18] and used a grid search considering two normalisations of P_0 (constant and author dependent), five lengths of fragments (full documents, 50, 100, 150 and 300 tokens as the shortest document contains 339 words), five token definitions (full words and four lengths of N -grams), two options for the attribution (Maximum Likelihood and Majority Rule) and 21 values of delta logarithmically spaced between 0.01 and 100. The left-out document is attributed using (one of) the sets of hyperparameters that give the best accuracy out of the 2100 taken into consideration. The search over the entire space of parameters for the attribution of one document (including the use of lemma and word sequences) takes about two hours on a regular laptop computer (8×2.4 GHz CPU, 16 GiB RAM).

The first step forward is not to limit the analysis to the set of parameters that offers the best attribution on the training set but to consider all other sets that provide comparable results. To determine which results are comparable, we will assume that for every set of parameters, a “true” probability of correct attribution exists. We sample this probability in a leave-one-out experiment, but the number of correctly attributed texts can be higher or lower than expected due to chance. To limit the effect of the class imbalance, we will consider—instead of the simple fraction of correctly attributed books—the macro averaged recall. Taking the best-performing set of parameters as a reference, we consider all the sets for which the fraction of correctly attributed texts is at least at the 2.5th percentile in the confidence interval of the best result, assuming a Bernoulli distribution. This choice will allow us to distinguish cases where the attribution is unanimous and where different authors compete. In this case, every set of parameters will vote for the final attribution.

While this procedure allows attribution, it does not allow comparisons between documents. For every document t_j , the software returns the average log-likelihood per token $\frac{1}{N} \log \mathcal{L}(A_i | t^j) = \frac{\mathcal{L}_i^j}{N}$ of every author A_i , with N number of tokens. These likelihoods are not directly comparable across documents. Indeed, in the leave-one-out approach, each known

document of an author and the anonymous are compared against slightly different versions of the author’s corpus. For each of the M documents of A_i , the likelihood $\mathcal{L}(A_i | t_j)$ will be computed using a corpus of $M - 1$ documents. The reference corpus of A_i contains all M for the anonymous document.

To compare documents, we will ignore this aspect for two reasons: First, the reference corpus is meant to reflect the best available description of the author as a proxy for the author’s style. Each of the different versions represents the author with varying approximations. Second, from a more technical point of view, the effect on the likelihood of the changing reference corpus decreases with the size of the corpus itself.

We will now consider minus the inverse of the output of CP2D, i.e., $x_i^j = -N/\log(\mathcal{L}_i^j)$, and treat these as Cartesian coordinates. With this transformation, the most likely author is still associated with the maximum coordinate, and each author identifies with one of the axes in space. The smallest angle ϕ_i^j between the document and the axes in the n -dimensional space, with n the number of authors, identifies the most likely author. In the limit of $\mathcal{L}_i^j \rightarrow 1$ (increasing likelihood of the author), the associated coordinate $x_i^j \rightarrow \infty$ and the document moves towards the axis $\phi_i^j \rightarrow 0$.

The same attribution results would be achieved by projecting all the points on the surface of an n -ball, i.e., an $(n - 1)$ -sphere. Since the variability of the values \mathcal{L}_i^j is limited in practice, most documents are scattered around the n -dimensional bisector. Thus, the distance from the origin encodes general information on the typicality of the documents. In the following, we will disregard this information and work only with the ϕ_i^j computed as:

$$\phi_i^j = \arctan \frac{\sqrt{\sum_{m=i+1}^n x_m^j{}^2}}{x_i^j} \quad (1)$$

with N the number of candidate authors and $\phi_n^j = \pi/2 - \phi_{n-1}^j$. We apply this transformation only to the likelihood values computed with the sets of hyperparameters that include Maximum Likelihood attribution, setting aside the attribution with Majority Rule. When computing attribution based on the angle between documents, we use the cosine distance of the x_i^j to determine the nearest document.

This measure misses some characteristics of a proper metric. Most notably, the angle between two documents can be zero without being the same text. If the two texts differ only in the order of the words and in words that appear only in the individual documents (and with the same distribution of the frequencies), every author will have the same likelihood for both texts, which will have zero distance. The distance between texts should not be interpreted as a measure of their textual difference, as the position in space depends on the relationship with the authors. However, it can be viewed as a measure of the stylistic difference.

This projection allows us to visualise on 2D paper the relationship with up to three authors without dimensionality reduction (a 3D sphere has 2D surface). This natural representation allows visualising decision boundaries, defining regions associated with each author and corresponding to the ML attribution. Moreover, when interested in stylistic relationships and not in attribution, we can use just a single level of the leave-one-out procedure. This means looking at the documents of a group of authors when none of them is treated as anonymous. Here,

each document is compared against all others. In practice, in Fig. 2, this is the case of the works of Horace, Propertius and Tibullus in panels A–F.

Acknowledgments

The author thanks Noemi Daria Zaccagnino for providing the lemmatisation of the *Consolation ad Liviam* and other advice in the assembly of the corpus. Her contribution was essential to the completion of this work.

References

- [1] P. Agapitos and A. van Cranenburgh. *A Stylometric Analysis of Seneca’s Disputed Plays. Authorship Verification of Octavia and Hercules Oetaeus*. Tech. rep. 1. Darmstadt: TU Darmstadt, 2024, 31 Seiten. DOI: <https://doi.org/10.26083/tuprints-00027394>.
- [2] D. Bamman and P. J. Burns. *Latin BERT: A Contextual Language Model for Classical Philology*. 2020. arXiv: 2009.10053 [cs.CL].
- [3] J.-P. Benzécri. *L’Analyse des Correspondances*. Vol. 2. 2 vols. Paris, Bruxelles, Montreal: Dunod, 1973. 625 pp.
- [4] J.-P. Benzécri. *L’Analyse des Données*. 2 vols. Paris, Bruxelles, Montreal: Dunod, 1973.
- [5] T. J. Bolt, J. H. Flynt, P. Chaudhuri, and J. P. Dexter. “A Stylometry Toolkit for Latin Literature”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Ed. by S. Padó and R. Huang. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 205–210. DOI: 10.18653/v1/D19-3035.
- [6] P. Chaudhuri, T. Dasgupta, J. P. Dexter, and K. Iyer. “A small set of stylometric features differentiates Latin prose and verse”. In: *Digital Scholarship in the Humanities* 34.4 (2018), pp. 716–729. DOI: 10.1093/llc/fqy070.
- [7] M. Eder, J. Rybicki, and M. Kestemont. “Stylometry with R: A Package for Computational Text Analysis”. In: *The R Journal* 8.1 (2016), pp. 107–121. DOI: 10.32614/rj-2016-007.
- [8] T. E. Franklino and L. Fulkerson. *Constructing Authors and Readers in the Appendices Vergiliana, Tibulliana, and Ovidiana*. Oxford University Press, 2020. DOI: 10.1093/oso/9780198864417.001.0001.
- [9] R. Gorman. “Morphosyntactic Annotation in Literary Stylometry”. In: *Information* 15.4 (2024). DOI: 10.3390/info15040211.
- [10] P. Juola. “JGAAP: A system for comparative evaluation of authorship attribution”. In: *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*. 2009. DOI: 10.6082/m1n29v4z.
- [11] F. Karsdorp, M. Kestemont, and A. Riddell. *Humanities Data Analysis: Case Studies with Python*. Princeton University Press, 2021.

- [12] Kestemont, Mike and Moens, Sara and Deploige, Jeroen. “Collaborative authorship in the twelfth century: a stylometric study of Hildegard of Bingen and Guibert of Gembloux”. In: *Digital Scholarship In The Humanities* 30.2 (2015), 199–224.
- [13] M. Koppel, J. Schler, and S. Argamon. “Authorship attribution in the wild”. In: *Language Resources and Evaluation* 45.1 (2011), pp. 83–94.
- [14] M. Koppel and Y. Winter. “Determining if two documents are written by the same author”. In: *Journal of the Association for Information Science and Technology* 65 (1 2014), pp. 178–187. DOI: 10.1002/asi.22954.
- [15] D. Longree and M. Fantoli. *LASLAfiles_Latin_APNformat*. Version V1. 2023. DOI: 10.58119/ulg/qjj0sa.
- [16] B. Nagy. (Not) *Understanding Latin Poetic Style with Deep Learning*. 2024. DOI: 10.48550/arXiv.2404.06150.
- [17] B. Nagy. “Some stylometric remarks on Ovid’s *Heroides* and the *Epistula Sapphus*”. In: *Digital Scholarship in the Humanities* 38 (3 2023), pp. 1183–1199. DOI: 10.1093/llc/fqac098.
- [18] G. T. Raffaelli, M. Lalli, and F. Tria. “Inference through innovation processes tested in the authorship attribution task”. In: *Communications Physics* 2024 7:1 7 (1 2024), pp. 1–8. DOI: 10.1038/s42005-024-01714-6.
- [19] G. Tani Raffaelli, M. Lalli, and F. Tria. *GiulioTani/InnovationProcessesInference: Accepted*. Version v1.0.0. 2024. DOI: 10.5281/zenodo.12163218.

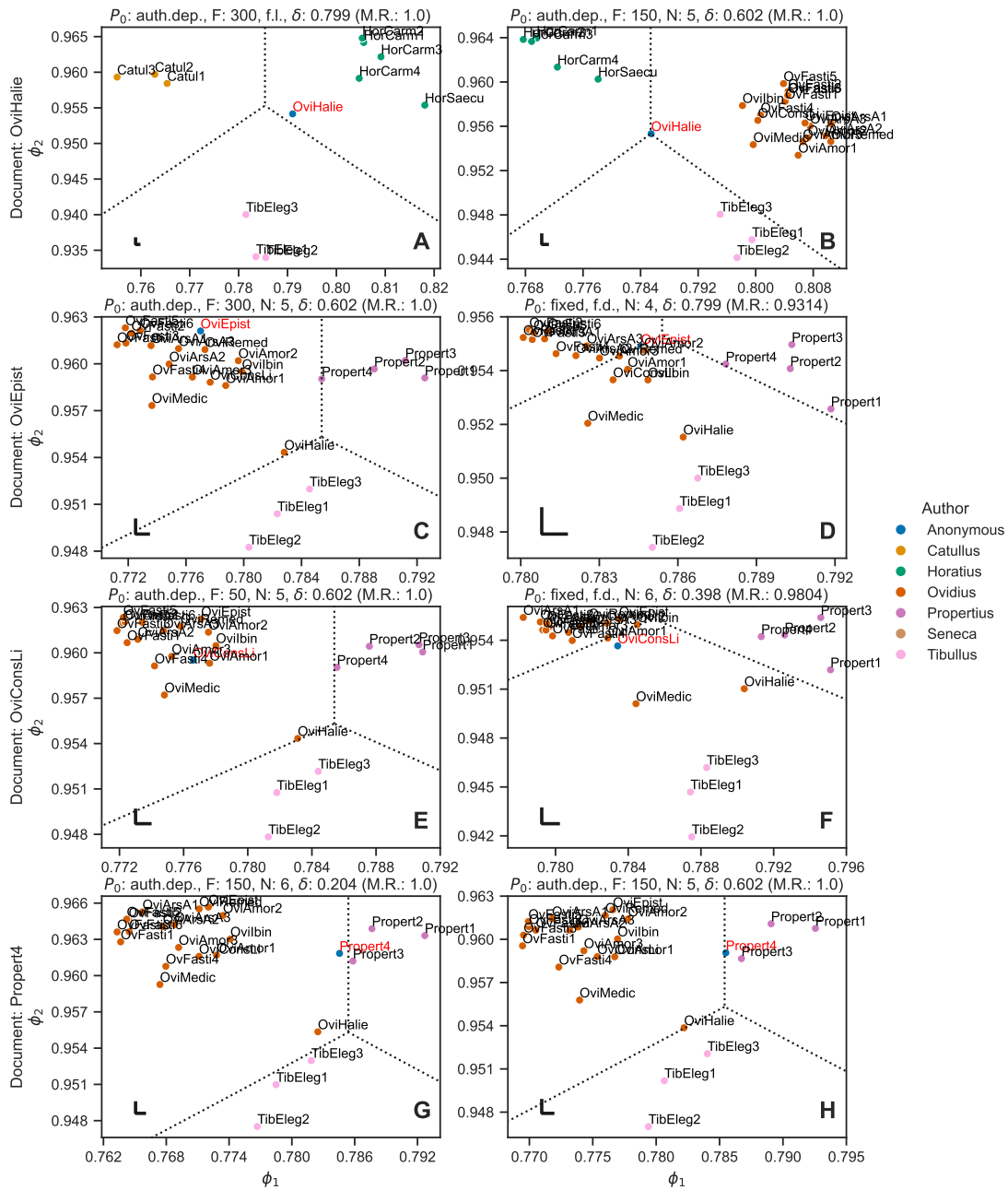


Figure 2: Positions in the space of three documents of debated authorship (*Halieutica*, *Consolatio ad Liviam* and *Heroidum Epistulae* from Ovid) and one of unstable attribution (the fourth book of the *Elegiae* from Propertius), using lemmas. The dotted lines correspond to equal likelihood of the authors on the two sides and mark the decision boundary when using direct attribution based on likelihood. In all panels, documents towards the bottom of the plots would be assigned to Tibullus. Documents in the top-left corner would go to Catullus (A), Horace (B) or Ovid (C–H). Documents in the top-right corner would go to Horace (A), Ovid (B) or Propertius (B–H). The title of every panel reports the set of parameters used for plotting and the macro-averaged recall (M.R.). The normalisation of P_0 is either fixed or author-dependent (auth.dep.), ‘F’ is the number of tokens per fragment (f.t. for full documents), ‘N’ is the size of N -grams (f.l. for full lemmas), and ‘ δ ’ is the correction to P_0 . We chose these sets of parameters for illustrative purposes among those offering results comparable to the maximum. The scale bars in the bottom left corner of each panel have a fixed size of 0.001 rad.