THE AUTHORSHIP OF THE *HISTORIA AUGUSTA*: TWO NEW COMPUTATIONAL STUDIES

The *Historia Augusta* is a collection of biographies of Roman emperors stretching from Hadrian (117-138) to Carus (282-83) and his son Carinus (283-285). The lives purport to be written by six different authors, Aelius Spartianus, Julius Capitolinus, Vulcacius Gallicanus, Aelius Lampridius, Trebellius Pollio, and Flavius Vopiscus, working under the Emperors Diocletian (284-305) and Constantine (306-337). For much of the period it covers, the *HA* represents the only extended narrative source, and the testimony it offers is invaluable. Unfortunately, the *HA* is also famous for its bizarre details and puzzling omissions, its lurid focus on emperors' peccadilloes and personal habits to the detriment of their political accomplishments. It also notoriously includes documents – speeches, letters, laws – which are almost certainly fabricated, and cites a whole host of authors nowhere else attested and which are probably invented. But the problem of the *HA* is not only its unreliability as an historical text: it also includes throughout troubling anachronisms, mentions of office and titles that only came into being in the middle of the fourth century, decades after the supposed date of its composition. In 1889, Hermann Dessau put forth the provocative thesis that the *HA* was in fact the work of a single author working under the reign of Theodosius (379-395), and that division of the lives between six authors and their dedications to Diocletian and Constantine were merely a literary ploy.[1] Ronald Syme – the most influential exponent of the Dessau thesis – would famously term the author 'a rogue *grammaticus*.'[2]

[1] H. Dessau, "Über Zeit und Persönlichkeit der Scriptores Historiae Augustae," *Hermes* 24 (1889), 337-92.
[2] Syme, *Ammianus and the Historia Augusta* (Oxford 1968), 207.

I A COMPUTATIONAL SOLUTION?

As early as the late 1970s it was realized that this question of single or multiple authorship in a corpus offered a perfect test case for statistical methods of authorship attribution. Ian Marriott conducted a ground-breaking analysis, published in *Journal of Roman Studies* in 1979, which suggested that computational analysis indicated single authorship of the corpus.[3] This was the first application of forensic stylometrics, as developed by Mosteller and Wallace, to a Latin text.[4] Unfortunately, his analysis was marred by methodological errors, particularly the use of sentence length as a criterion of authorship, which is no longer considered an effective stylometric feature even for modern texts, and should definitely not be used for ancient texts, where the punctuation is due to the modern editor.[5]

Subsequent analyses, foremost among them by the trio of Emily K. Tse, Fiona J. Tweedie, and Bernard Frischer in the September 1998 issue of *Literary and Linguistic Computing*, which was substantially devoted to the authorship of the *Historia Augusta*, strongly supported the opposite contention, that the *Historia Augusta* "is not more variable than a corpus constructed to mimic the authorial structure as outlined in the manuscript tradition … [T]he variability of usage of function words may be used as a measure of multiple authorship, and that based on the use of these function words, the *SHA* appears to be of multiple authorship."[6] Most historians (though by no means all) accept some version of the Dessau theory of single authorship.[7] This disjunct between the evidence from historiography and traditional philology

---

[3] I. Marriott, " The Authorship of the *Historia Augusta*: Two Computer Studies," *JRS* 69 (1979), 65-77.
[4] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist* (Cambridge, MA, 1964).
[5] D. Sansone, "The Computer and the *Historia Augusta*: A Note on Marriott," *JRS* 80 (1990), 174-77.
[6] E. K. Tse, F. J. Tweedie, and B. D. Frischer, "Unravelling the Purple Thread: Function Word Variability and the *Scriptores Historiae Augustae*," *LLC* 13 (1998): 141-49 at 145-6. The same issue contains three articles by P. J. and L. W. Gurney, and a cautionary note by J. Rudman (see below).
[7] See most recently, D. Rohrbacher, *The Play of Allusion in the Historia Augusta* (Madison 2016), 4-6. In the twentieth century, the most prominent voice calling the Dessau thesis into question was that of A. Momigliano; see for example his "An Unsolved Problem of Historical Forgery: The Scriptores Historiae Augustae" *Journal of the Warburg and Courtauld Institutes* 17 (1954): 22-46. D. den Hengst is one scholar who felt the need to revisit the

and computational analysis has probably lead to a devaluation of computational methods in classical scholarship, and made computational linguists reluctant to work on *Echtheitskritik* of Latin texts.

Additionally, Joseph Rudman published a damning critique of the state of the art in computational *HA* studies in the same issue of *LLC* in 1998 and few studies have dared to take up the case study afterwards.[8] Rudman's critique is -- sometimes unreasonably -- harsh on previous scholarship and addresses issues which are nowadays considered much less problematic than he did in 1998.[9] The problem of homonymy in word counting or minor reading errors in the transmitted manuscripts, to name but two examples, are no longer considered a major impediment in automated authorship studies anymore.[10] Scholars generally have also obtained a much better understanding of the effect of genre signals or the use of background corpora.[11] Most importantly, however, the widely available computational tools available today are exponentially more powerful than what was available a decade ago and stylometric analysis has seen a tremendous growth and development.[12] While we should not overestimate the performance of modern techniques, the *HA* is too interesting a case study in stylometry to be abandoned altogether. One interesting development is that previous studies sometimes adopted a fairly static conception of the phenomenon of authorship, in the traditional sense of an *auctor intellectualis*. A wealth of studies in more recent stylometry have problematized this concept,

---

question of single authorship subsequent to the 1998 papers, suggesting that a naive sense of single authorship was no longer tenable; see "The Discussion of Authorship," in the *Emperors and Historiography* (Leiden 2010), 177-185, originally published in G. Bonamente and F. Paschoud, eds. *Historiae Augustae Colloquium Perusinum* (Bari 2002), 187-195. R. Baker has recently upheld a multi-authorial view of the text, in his 2014 Oxford D.phil. thesis, "A study of a late antique corpus of biographies [*Historia Augusta*]."

[8] J. Rudman, "Non-Traditional Authorship Attribution Studies in the *Historia Augusta*: Some Caveats," *LLC* 13 (1998): 151-57.

[9] Cf den Hengst (2010), 184.

[10] M. Eder, "Mind your corpus: systematic errors in authorship attribution", *LLC* 28 (2013): 603-614.

[11] P. Juola, "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions", *DSH* [2016].

[12] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods". *JASIST* 60 (2009): 538-556.

also from a theoretical perspective, shedding light on more complex forms of collaborative authorship and translatorship, or even cases where layers of 'editorial' authorship should be discerned.[13] As such, more subtle forms of authorship, including the phenomenon of *auctores manuales*, have entered the stylometric debate.

In this paper, we report the results of new computational experiments in the corpus of the *Historia Augusta*, and argue that they indicate that the problem of the authorship of the corpus too complex to be reduced to the bare alternative between single or multiple authorship. In the past, the *HA* has been primarily studied as a problem in authorship *attribution*, which as we will argue below, is not necessarily the optimal framework to assess the authorship of the *HA* in. One important novelty is therefore that we also approach the *HA* specifically as a problem in authorship *verification*, an innovative setup which was introduced only recently.[14] Next, we also apply the more conventional technique of PCA to obtain more insight into the stylistic structure of the work from an interpretative point of view, based on the results obtained from the verification analysis. Both analyses allow us to argue that (at least) two distinct authorial layers can be detected in the text, and that the computational data indicates a solution complementary to the solutions developed from traditional philological methods.

II DIVIDING THE CORPUS

One striking feature of the transmitted text of the *HA* is the lacuna, or gap, covering the emperors between Gordian III and Valerian, that is the eventful years between 244 and 253. This gap seems to correspond with a natural division in the text, since all the lives before are assigned

[13] See e.g. N.B. Reynolds, G.B. Schaalje & J.L. Hilton, "Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works" *DSH* 27 (2012): 409-425 or M. Kestemont, S. Moens & J. Deploige, "Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux" *DSH* 30 (2015): 199–224.

[14] M. Koppel, J. Schler, S. Argamon and Y. Winter, "The "Fundamental Problem" of Authorship Attribution", *English Studies* 93 (2012), 284-291.

indiscriminately to the four authors, Capitolinus, Spartianus, Lambridius and Vulcacius, while those following the lacuna are written successively by Trebellius and Vopiscus. Vopiscus is the only one who alludes to the other authors of the compilation. Of course this could mean that there was a real division into two parts and that the preface of the second part is simply lost in the lacuna due to the vagaries of transmission. But the coincidence of accidental loss of text occurring precisely at the point where the nature of the corpus and its authorship changes has led many to suspect that the lacuna is a deliberate feature of the corpus.[15]

But there is another significant piece of evidence: the lives of Trebellius and Vopiscus are almost entirely fictional, and are stuffed with forged documents, invented sources, and implausible anecdotes. The earlier lives, however, are much more varied in their reliability: some seem accurate and are relatively sober, even dull, such as those of Hadrian and Marcus Aurelius; others are fantastic and improbable such as that of the Caesar Lucius Aelius; and others still seem to have a mixture of reliable information and invented details, such as the life of Alexander Severus. Shortly after Dessau, it was noticed that there seemed to be a series of ten generally reliable lives of emperors at the beginning of the corpus – from Hadrian to Caracalla – into which were interspersed fanciful and unreliable lives of junior emperors or Caesars. This led to the division of the first part of the corpus into 'primary lives' (*Hauptviten*) and 'secondary lives' (*Nebenviten*), first proposed by Mommsen in 1890, presented here with Syme's promotion of the life of Verus into the first category and expulsion of that of Macrinus[16]:

Hauptviten	**Nebenviten**

[15] See Rohrbacher (2016), 9-10.
[16] T. Mommsen, "Die Scriptores Historiae Augustae," *Hermes* 25 (1890): 228-292; Syme's modifications were proposed in his essay on the secondary vitae from 1968-69, collected in *Emperors and Biography*: *Studies in the Historia Augusta* (Oxford 1971), 56-8.

Hadrian	Aelius
Antoninus	Avidius Cassius
Marcus Aurelius	Pescennius Niger
Verus	Clodius Albinus
Commodus	Geta
Pertinax	
Didius Iulianus	
Septimius Severus	
Caracalla	

After Geta, all the way up to lacuna, come the so-called intermediate lives, which display many of the same features as the secondary lives and those after the lacuna, but also seem to transmit some genuine information among their fancies.[17]

In addition to the significant features, we can also deduce from cross references within the *HA* that the collection began with Nerva. Hence, two lives seem to have been lost at the beginning, those of Nerva and Trajan. Assuming those lives belonged to the category of *Hauptviten*, the whole series would become a continuation of the work of the biographer Suetonius, who composed the lives of the twelve emperors from Augustus to Domitian. It has been further noted that the *HA* refers in most of those lives, as well as in the life of Heliogabalus (Elgabalus), to lives of the emperors written by Marius Maximus. Hence, one popular theory is that Marius Maximus composed the lives of twelve emperors as a sequel to Suetonius, and that the *Hauptviten* substantially represent a reworking of the earlier text. Once Marius Maximus gave out, the authors of the *HA* resorted to scrappier and less reliable sources, and gave freer rein to invention and fancy. Others, following Syme, reject the theory that Marius Maximus was the source text, and instead posit some other, unknown source (*Ignotus*).

[17] This fourfold division comes from Syme (1971). Modifications to this scheme have been proposed by Rohrbacher (2016): 8-9; he argues that the lives of Heliogabalus and Alexander Severus belong in their own category. See also D. W. Burgersdijk, "Style and structure of the *Historia Augusta*," PhD diss. (Amsterdam 2010); he argues that the division is reflected in the presence of interpolated materials, such as document, letters, *etc*.

III AUTHORSHIP VERIFICATION

Computational authorship studies are an increasingly popular research topic, both in Computer and Information Sciences, as well as in the Digital Humanities. It can be considered a form of style-based document authentication (*Echtheitskritik*), which has valuable applications, which extend well beyond the domain of literary analysis, to, for instance, for the domain of forensic sciences. Quoting Stamatatos's 2009 survey of the field, it is clear that "[t]he main idea behind statistically or computationally-supported authorship attribution is that by measuring some textual features we can distinguish between texts written by different authors."[18] This basic assumption implies that it should be possible to assess, for any new unseen document, whether or not it was written by other authors for which we have texts available. Nowadays computational authorship studies are often considered a subfield of 'stylometry' in the Digital Humanities, the broader computational study of the writing style of texts.[19]

While stylometry has a rich history, dating back to at least the 19th century, it is clear that it received its most important impetus only in the past two or three decades, stimulated by the rise of (personal) computing and the increased availability of large bodies of text in electronic form. Apart from the influential, yet more conventional statistical analyses carried by pioneers such as Mosteller & Wallace or John Burrows well before the 1990s, an influential approach in authorship studies has been to approach the attribution of anonymous texts as a 'text categorization' problem.[20] Heavily influenced by parallel research in Computer Science, the idea was to optimize a statistical classifier on example texts by a number of available candidate authors, much like a spam filter nowadays is still trained on manually annotated emails to learn

---

[18] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods". *JASIST* 60 (2009): 538.
[19] D. Holmes, "The Evolution of Stylometry in Humanities scholarship", *LLC* 13 (1998): 111-17.
[20] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist* (Cambridge, MA, 1964) and J. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels* (Oxford, 1987).

how to distinguish between 'junk' email and normal messages.[21] After training such a classifier on this example data, the classifier could then be used to categorize or classify anonymous text as belonging to one of the training authors' oeuvres.

Nowadays, this text categorization setup is commonly known as 'authorship *attribution*'.[22] It resembles a police lineup, in which the correct author of an anonymous text has to be singled out from the available candidates. For a number of years, practitioners of stylometry have come to acknowledge the limitations of authorship attribution, because it necessarily assumes that the correct target author is indeed included in the set of available candidates. In many real-world cases, however, this problematic assumption cannot possibly be made, because the set of candidates cannot be known beforehand. Because of this, the setup of authorship verification recently has been designed as an experimental framework: here, the task is to verify whether *or not* an anonymous document was written by one or several of a series of candidate authors. In some sense, authorship verification redefines the text categorization problem by adding an additional category label: 'None of the above'.

Verification is hence an increasingly common experimental setup in authorship studies, and is the topic of a dedicated track in the yearly PAN competition.[23] Generally speaking, authorship verification is a more generic problem than authorship attribution - i.e. every attribution problem could in principle be cast as a verification problem - but it has also proven to be more challenging. In the present context, it should be emphasized that the problem posed by *HA* is a 'vanilla' example of a problem in authorship verification: while the corpus indeed contains a number of (auto-)attributions, the veracity of all of these has been questioned in

[21] F. Sebastiani, "Machine Learning in Automated Text Categorisation", *ACM Computer Surveys* 34 (2002): 1-47.
[22] The following paragraphs heavily draw on M. Koppel & Y. Winter, "Determining if two documents are written by the same author", *JASIST* 65 (2014): 178–187.
[23] The competition's website is pan.webis.de. The most recent survey of an authorship verification track is: E. Stamatatos et al., "Overview of the Author Identification Task at PAN 2015". In L. Cappellato, et al. (eds.), *Working Notes Papers of the CLEF 2015 Evaluation Labs* (2015).

previous scholarship. In our experiments, we have therefore attempted to radically minimize any assumptions on our part as to the authorial provenance of the texts in the *HA*. For each piece of text analyzed below, we propose to independently assess the probability that it was written by one the (alleged) individual authors identified in the corpus.

Authorship verification is commonly based on a form of thresholding: only if an anonymous document is close enough to a given target author's oeuvre, it will be attributed to that author. In all other cases, a verification system refrains from an attribution. The primary question then is how to calculate the similarity between the unknown document and a given oeuvre. In this paper, we make use of the General Imposters (GI) framework to this end, a highly successful approach to authorship verification. Apart from a seminal application to historical Latin, variants of this system have consistently ranked very highly in recent editions of the PAN competition.[24] The GI starts out from the assumption that it is dangerous to base a verification system on the direct (or 'first order') comparison between two documents: if two documents, by different authors, happen to address the same topic, this would for instance artificially increase their stylistic similarity.[25] Additionally, the GI puts a lot of weight on the idea that two documents cannot be compared in a stylistic vacuum: determining whether two documents were authored by the same individual, should make use of relevant comparands. If two documents are written by the same author, they should be consistently more similar to each other than to other texts written by different authors.

The GI therefore proposes the following iterative procedure, which can be likened to forms of 'bootstrapping'. Let *x* represent an unknown document and let *y* represent a random target author's stylistic profile. During 100 iterations, will randomly select (a) 50% of the

[24] Compare the setup in: J.A. Stover, Y. Winter, M. Koppel and M. Kestemont, "Computational authorship verification method attributes a new work to a major 2nd century African author", *JASIST* 67 (2016): 239–242.
[25] M. Koppel and S. Seidman, "Automatically Identifying Pseudepigraphic Texts", In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, 2013), 1449-1454.

available stylistic features available (e.g. word frequencies) and (b) 30 distractor authors, or 'impostors' from a pool of similar texts. In each iteration, the GI will compute whether *x* is closer to *y* than to any of profiles by the 30 impostors, given the random selection of stylistic features in that iteration. Instead of basing the verification of the direct ('first order') distance between *x* and *y*, the GI proposes to record the proportion of iterations in which *x* was indeed closer to *y* than to one of distractors sampled. This proportion can be considered a 'second' order metric and will automatically be a probability between 0 and 1, indicating the robustness of the identification of the authors of *x* and *y*. In the past, it has already been demonstrated that the GI system produces state-of-the-art verification results even for classical Latin prose.[26]

We have applied a generic implementation of the GI to the *HA* as follows. We split the individual lives into consecutive samples of 1000 tokens (i.e. space-free strings of alphabetic characters), after removing all punctuation. Each of these samples was analyzed individually by pairing it with the profile of one the *HA*'s six alleged authors, including the profile consisting of the rest of the samples from its own text. We represented the sample (the 'anonymous' document) by a vector comprising the relative frequencies of the 10,000 most frequent tokens in the entire *HA*. For each author's profile, we did the same, although the profile's vector comprise the *average* relative frequency of the 10,000 words. Thus the profiles would be the so-called 'mean centroid' of all individual document vectors for a particular author (excluding, of course, the current anonymous document).[27]

---

[26] Compare the setup in: J.A. Stover, Y. Winter, M. Koppel and M. Kestemont, "Computational authorship verification method attributes a new work to a major 2nd century African author", JASIST 67 (2016): 239–242. Our verification code is publicly available from the following repository: https://github.com/mikekestemont/ruzicka. This code will be described in a forthcoming paper: M. Kestemont, J.A. Stover, M. Koppel, F. Karsdorp, and W. Daelemans, "Authenticating the Writings of Julius Caesar".

[27] M. Koppel and S. Seidman, "Automatically Identifying Pseudepigraphic Texts", In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, 2013), 1449-1454.

Next, we ran the verification approach: during 100 iterations, we would randomly select 5,000 of the available word frequencies. We would also randomly sample 30 impostors from a large 'impostor pool' of documents by Latin authors, mostly historians such as Suetonius or Livy.[28] In each iteration, we would check whether the anonymous document was closer to the current author's profile than to any of the impostors sampled. In this study, we use the 'minmax' metric, which was recently introduced in the context of the GI framework.[29] For each combination of an anonymous text and one of the six target authors' profiles, we would record the proportion of iterations (i.e. a probability between 0 and 1) in which the anonymous document would indeed be attributed to the target author. The resulting probability table is given in full in the appendix to this paper. Although we present a more detailed discussion of this data below, we have added **Figure 1** below as an intuitive visualization of the overall results of this approach. This is a heatmap visualisation of the result of the GI algorithm for 1,000 word samples from the lives in the *HA*. Cell values (darker colors mean higher values) represent the probability of each sample being attributed to one of the alleged *HA* authors, rather than an imposter from a random selection of distractors. To the left, a clustering has been added on top of the rows, reflecting which groups of samples behave similarly.

[28] The pool of imposter authors can be inspected in the code repository for this paper.
[29] See once again Koppel & Winter (2014).

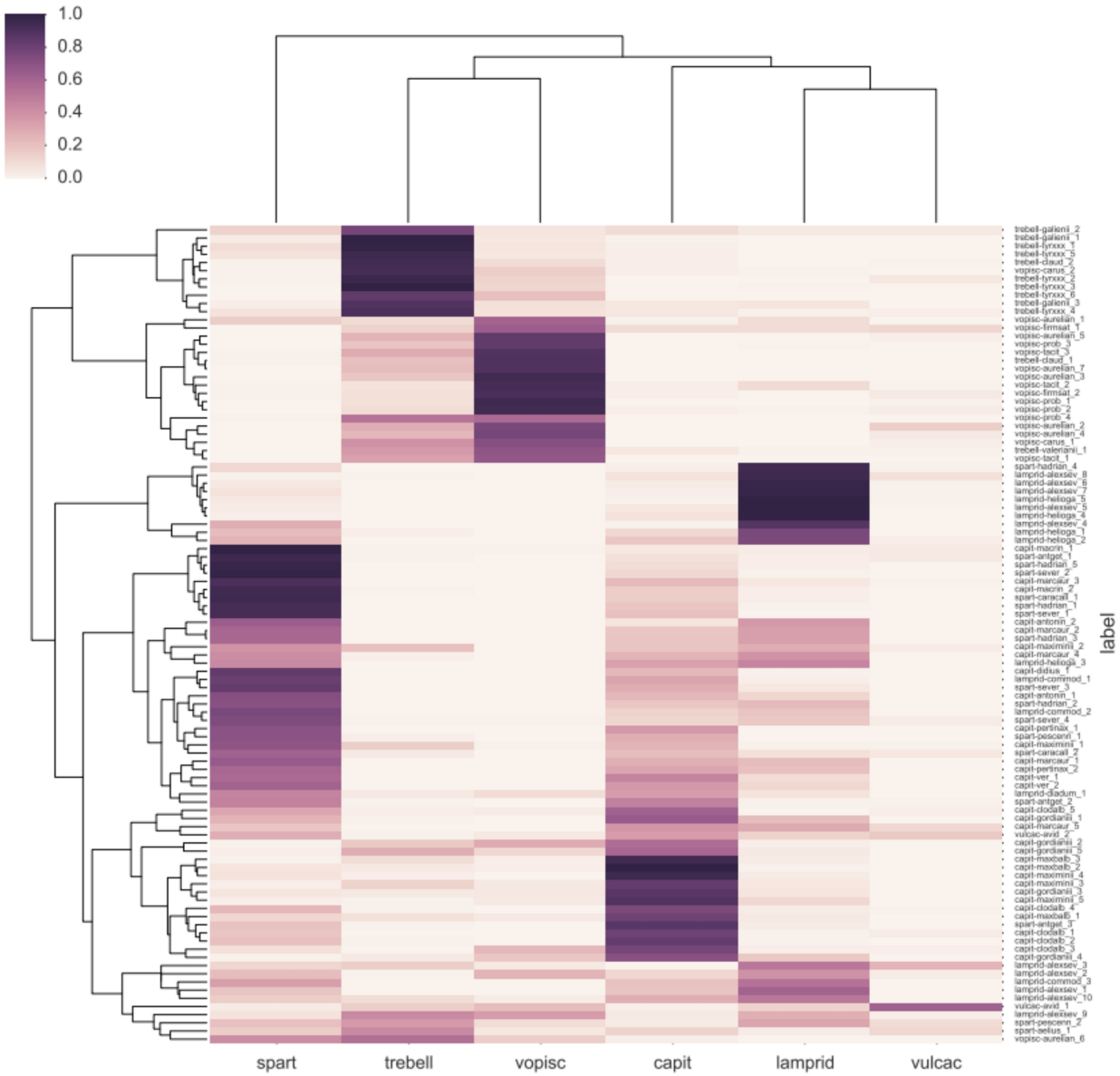

Figure 1: A heatmap visualisation of the result of the GI algorithm for 1,000 word samples from the lives in the *HA* (cf. *rows*). Cell values (darker colors mean higher values) represent the probability of each sample being attributed to one of the alleged *HA* authors (cf. *columns*), rather than an imposter from a random selection of distractors. To the left, a clustering has been added on top of the rows, reflecting which groups of samples behave similarly.

IV INTERPRETING THE RESULTS

The first result that emerges from the GI verification is that the corpus displays two sets of authorial signals. The four authors writing before the lacuna – Capitolinus, Spartianus,

Vulcacius and Lampridius (hereafter *HA*-a) – are indistinguishable from each other; likewise the two later authors – Trebellius and Vopiscus (*HA*-b) – are only marginally distinct from each other. There is, however, a measurable difference in style between these two groups of authors, and the GI attributes texts to one or the other with a high degree of confidence.[30] Nonetheless, some samples from before the lacuna are sometimes attributed to the second set instead of the first, while others are never so attributed. First the latter:

A. <u>Lives with no samples attributed to *HA*-b</u>

Hadrian
Antoninus
Verus
Commodus
Pertinax
Didius

All of these belong to the category of *Hauptviten*; this correlation becomes even stronger when we include lives with one sample only rarely attributed to *HA*-b (see the appendix for the numerical values):

B. <u>Lives with no more than one sample rarely attributed to *HA*-b (<.02)</u>

Hadrian
Antoninus
Marcus Aurelius
Verus
Commodus
Pertinax
Didius
Septimius Severus
Heliogabalus

[30] Den Hengst (2010) had this same impression, 182.

Only one of these lives, Heliogabalus, is not considered one of the *Hauptviten*; only one *Hauptvita* is not included in this list, Caracalla. Conversely, if we examine in which *HA*-a lives the samples are more strongly attributed to *HA*-b occur, a complementary picture emerges:

C. Lives with at least one sample more strongly attributed to *HA*-b (>.09)

Aelius
Avidius Cassius
Pescennius Niger
Clodius Albinus
Diadumenus
Alexander Severus
Maximini ii
Gordiani iii
Maximus and Balbinus

All of these lives are *Nebenviten* and intermediate lives. The only lives in *HA*-a not accounted for thus far are those with more than one sample weakly attributed to *HA*-b.

D. Lives with more than one sample rarely attributed to *HA*-b

Caracalla
Geta
Macrinus

In its broad outlines, this taxonomy corresponds with the traditional division of the corpus: the lives which our analysis shows contain very few authorial signals from *HA*-b correlate to the lives considered historically reliable, the *Hauptviten*; the lives which show at least occasional strong authorial signals from *HA*-b are the lives considered mostly or partly fictitious, the *Nebenviten* and the intermediate lives. Hence a taxonomy of the corpus that was formulated on the grounds of content can be reconstructed in almost all of its particulars solely from an analysis of style.

If we exclude the *Nebenviten*, and examine the corpus sequentially, we find that the lives from Hadrian to Didius Iulianus display no influence from *HA*-b, the lives from Septimius

Severus to Heliogabalus display some influence from *HA*-b, and the lives from Alexander Severus to Maximus and Balbinus (that is up to the lacuna) show strong influence from *HA*-b. In other words, the character of the style of the lives seem to undergo some change at the point at which the *HA*'s theoretical source text – lives of the twelve emperors from Nerva to Heliogabalus – gives out.

Alan Cameron presciently noted in 2011 that any attempts at authorship analysis of the *HA* ought to take into account the nature of the corpus as discovered by historians and philologists: "Given the undoubted fact that the earlier lives are largely based on Marius Maximus and/or (if he existed) Syme's *Ignotus*, few individual pages of any given life ascribed to Capitolinus, Lampridius, or Spartianus are likely to contain more than a handful of sentences that are entirely the work of the author (better compiler)."[31] Hence, Cameron doubts that computational methods will ever reveal a definitive answer to the question of multiple or single authorship: "In the circumstances, it is unlikely that the quantitative method will ever yield definitive results, and more traditional linguistic approaches may be more revealing."[32] We have shown, however, that there is no conflict between the results obtained from the two approaches. Indeed when we approach authorship from the perspective of the fourfold classification of the texts into primary, secondary, intermediate and later lives - a classification that likewise emerges from the GI verification, as we have shown above - the results become startlingly clear.[33]

V PRINCIPAL COMPONENT ANALYSIS

[31] Cameron, *Last Pagans of the Rome* (Oxford 2011), 744.
[32] *Ibid.*
[33] Tse, Tweedie, and Frischer briefly examined Syme's four categories, but concluded that their variability was too high to consider them relevant for authorship analysis.

Here we use a principal components analysis to obtain a better interpretation of the stylistic grouping which was supported by the GI experiment in the previous section.[34] Principal Components Analysis (PCA) is an established technique for analyzing multivariate data in statistics.[35] Nowadays, it is increasingly common in stylometry, because it has been shown to be a reliable technique to explore the main stylistic variation in corpora, also with respect to Latin data.[36] When we apply PCA to the function word frequencies in texts, the technique has an outspoken tendency to cluster stylistically similar texts. PCA is currently predominantly considered useful for the visualization of datasets of a moderate size, that do not contain too much different categories. Finally, PCA has the interesting characteristic that it is a so-called unsupervised method: it does not require access to any sort of (potentially biased) prior information about the texts to analyze. This is extremely useful in cases like the *HA*, where we do not wish to impose any pre-existing hypotheses on the material.

If we represent our corpus as a frequency table, we work with a matrix in which each row corresponds to a single text, and each column to a specific variable (e.g. a function word). The cells in this matrix are filled with the relative frequency of each variable in each texts. If we work with a matrix which contains just the 50 most frequent function words in the corpus, this matrix is still too large to inspect manually, since we are dealing with 50 axes or 'dimensions'. PCA is a technique for dimension reduction: it aims to replace the original 50 columns by a much set smaller of newly created columns or 'components'. The general idea is that these components should still offer an as accurate as possible approximation of the original frequency

[34] All analyses reported were carried using the following package: M. Eder, J. Rybicki and M. Kestemont, "Stylometry with R: A package for computational text analysis". *R Journal* 16 (2016): *advance access available*. See https://sites.google.com/site/computationalstylistics/.
[35] J.N. Binongo and W. Smith, "The Application of Principal Components Analysis to Stylometry", *LLC* 14 (1999): 445-66; J.N. Binongo, 'Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution', *Chance* 16 (2003): 9-17.
[36] Cf. J.A. Stover and M. Kestemont, "Reassessing the Apuleian Corpus: A Computational Approach", *CQ* [2016]: forthcoming.

table, i.e. when we attempt to reconstruct the original data from the compressed components, the loss should be minimal.

In stylometry, it is common to restrict analysis to the 2 (or sometimes) 3 components, which offer the best summary of the original data. These new dimensions are also called the 'principal' components, because together the offer the best fit of the original data. After performing the PCA, we are therefore left with just two 'dimensions' that describe our data and this makes it much easier to inspect and visualize our texts. The most common visualization of a PCA is a scatterplot as the one below: in this two-dimensional plot, texts are depicted as little dots. The horizontal and vertical axis correspond to the two principal components. From such scatterplots, much stylistic information can typically be gleaned: texts that adopt a similar style (i.e. have a similar frequency profile when it comes to function words) will be plotted in each other's neighborhood, whereas unrelated texts will be plotted in a different region. In general, it is important that PCA, unlike for instance the imposters technique, will not offer a single straightforward score for each text: the PCA scatterplot needs scholarly interpretation.

First, we examine all the lives from the standpoint of transmitted authorship, using the 200 most frequent words (MFW), excluding pronouns, with 1000 word samples:

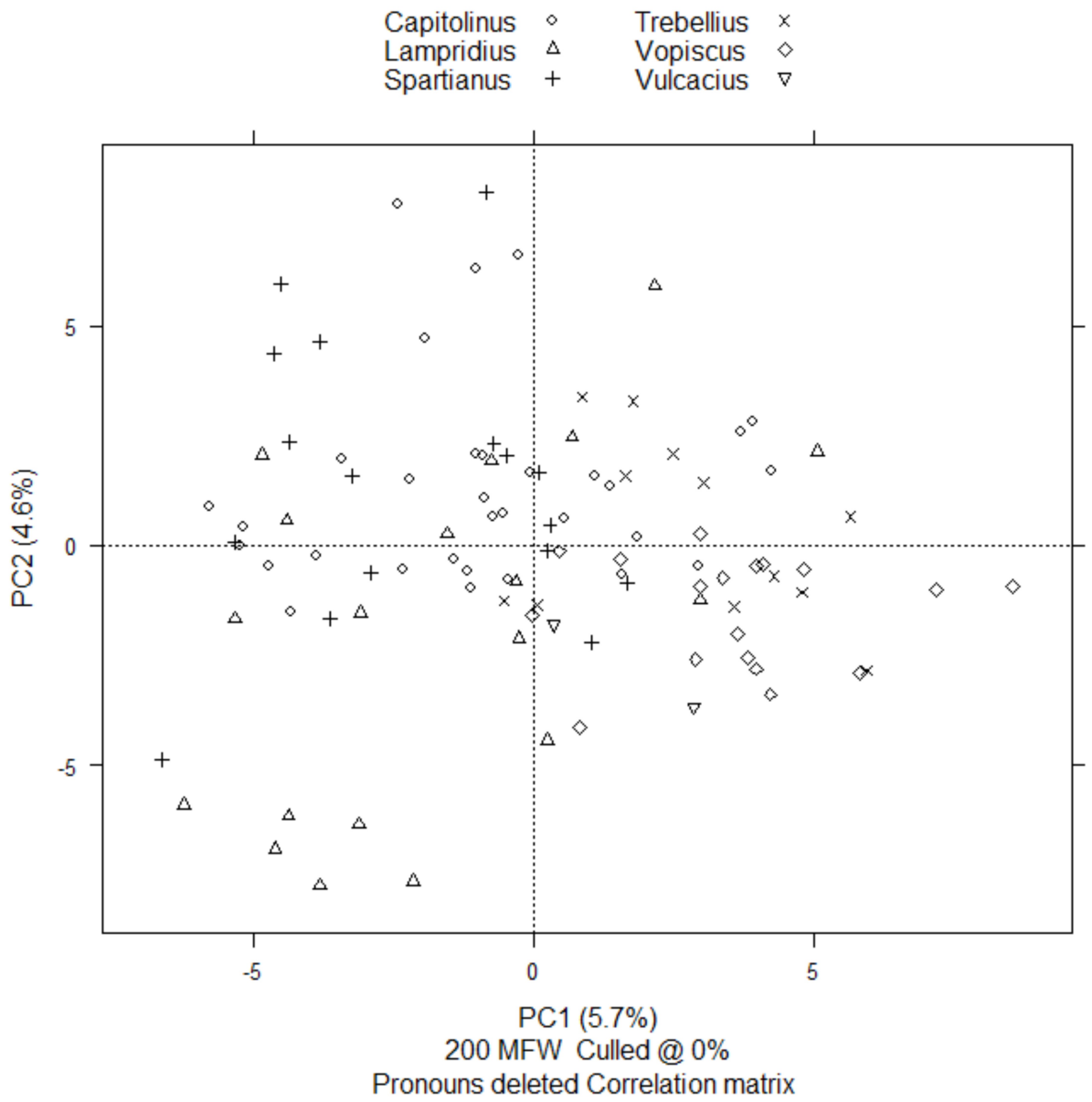

Figure 2

The result is chaos: there is no discernible pattern indicating any firm authorial relationship. The intermingling of samples strongly suggests single authorship. That suspicion grows stronger when two other authors of imperial biography are added for comparison, Aurelius Victor (*De Caesaribus*) and Suetonius (lives of Augustus, Caligula, Claudius and Domitian).

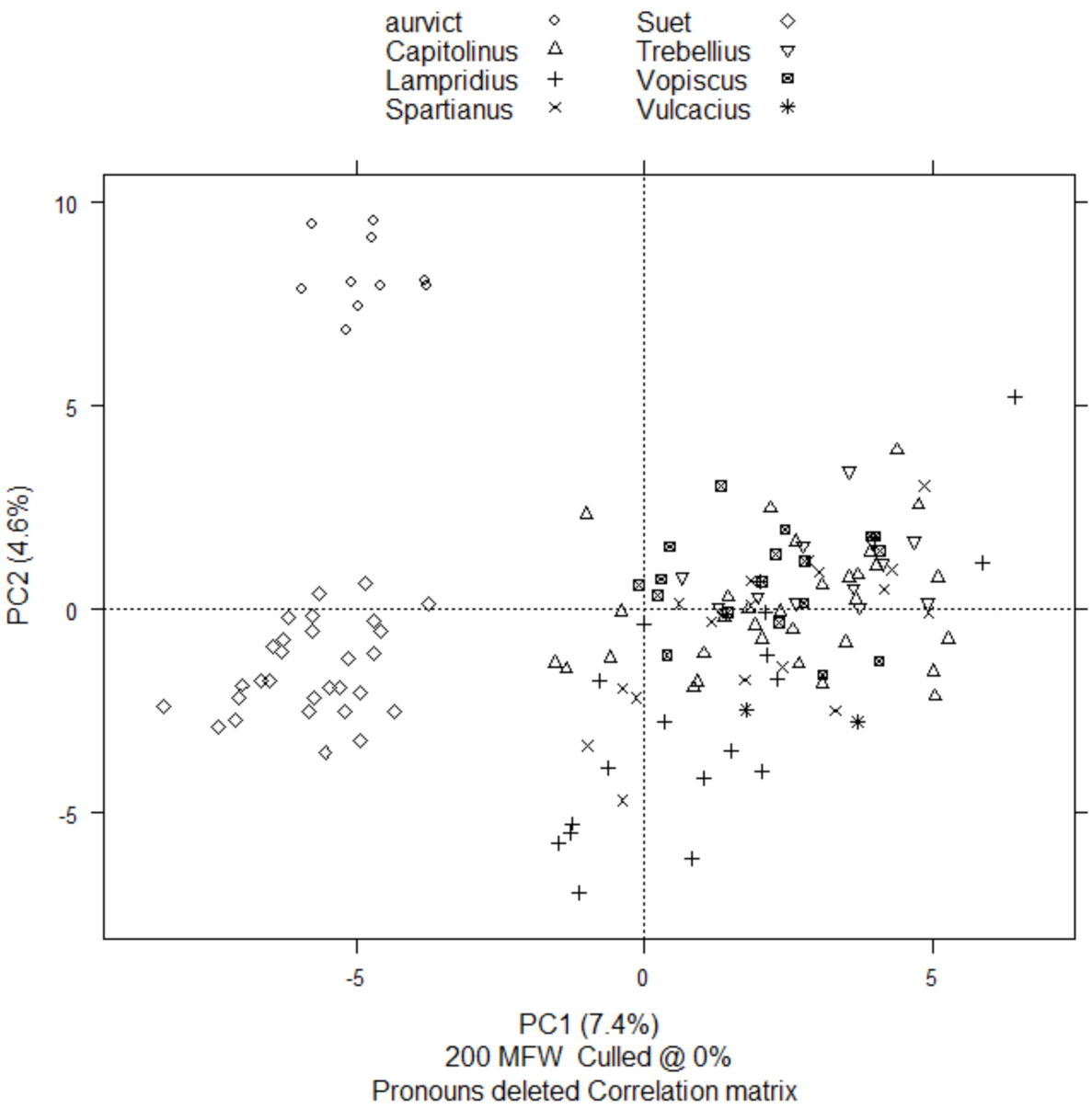

Figure 3

The *HA* corpus behaves here precisely as we would expect a single author text to behave under such experimental conditions: the samples cluster more compactly and form a group obviously distinct from the other two authorial groups.

If, however, we conduct a PCA ignoring the transmitted authorship designations and replace them with our four categories - the authorship tags are purely for visualization and play no role in the analysis of the samples - a much clearer picture emerges:

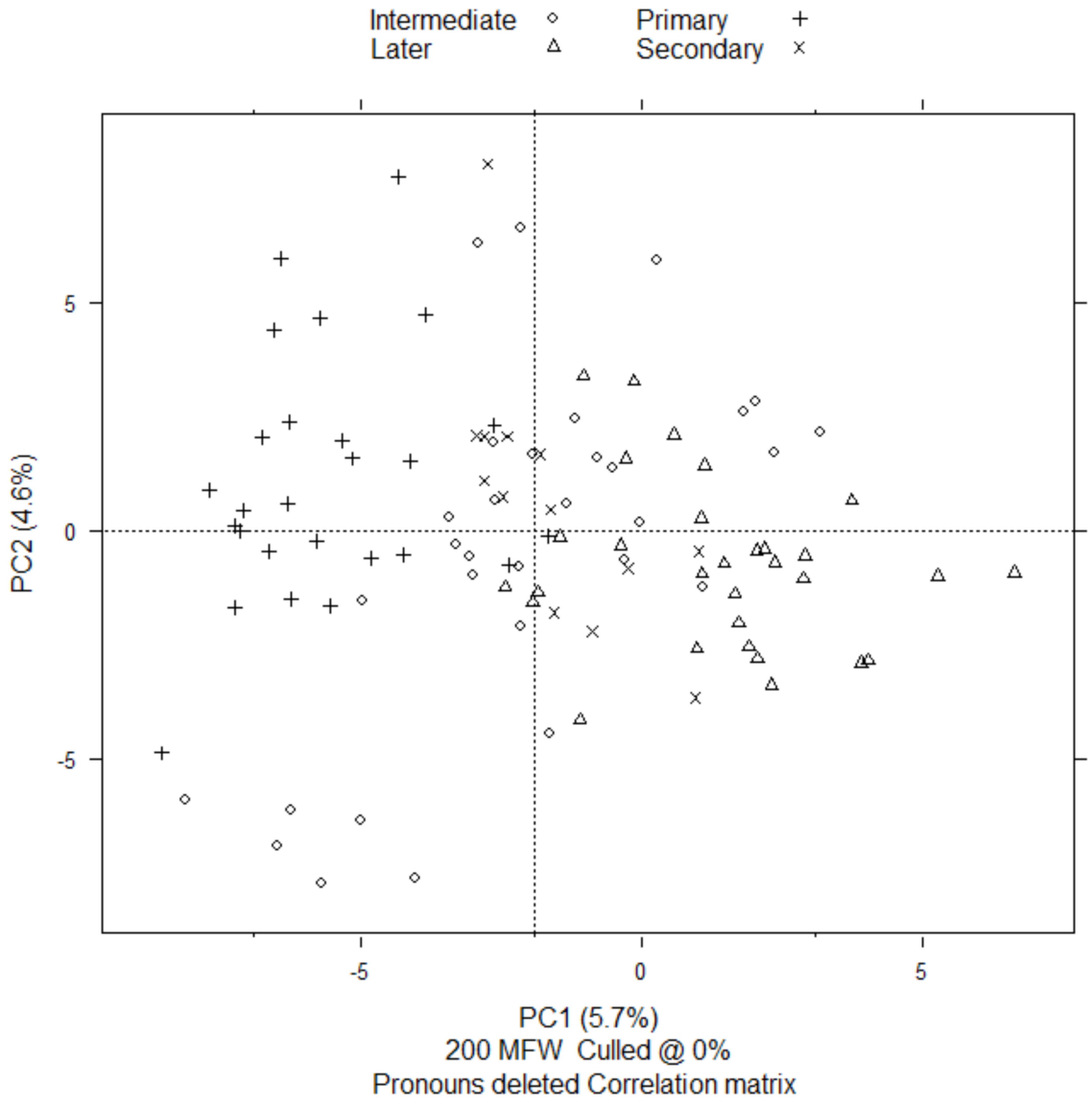

Figure 4

This is precisely the same plot as that presented above (**Figure 2**), with the names of authors - which are not relevant to the computational process - changed to those of the categories. The

result, however, is dramatically clearer: the primary lives all cluster in the top left quadrant, while the the later lives group on the right side of the plot. The secondary and intermediate lives mingle indiscriminately with both groups (the samples at the bottom left are from the lives of Alexander Severus and Heliogabalus, with one sample from Hadrian). Hence, it emerges that the primary and later lives can be considered distinct from one another, while the secondary and intermediate lives cannot be distinguished from each other nor from either ‘author’. An extremely clear picture emerges from a consideration only of the primary lives and the later lives, with Suetonius for comparison:

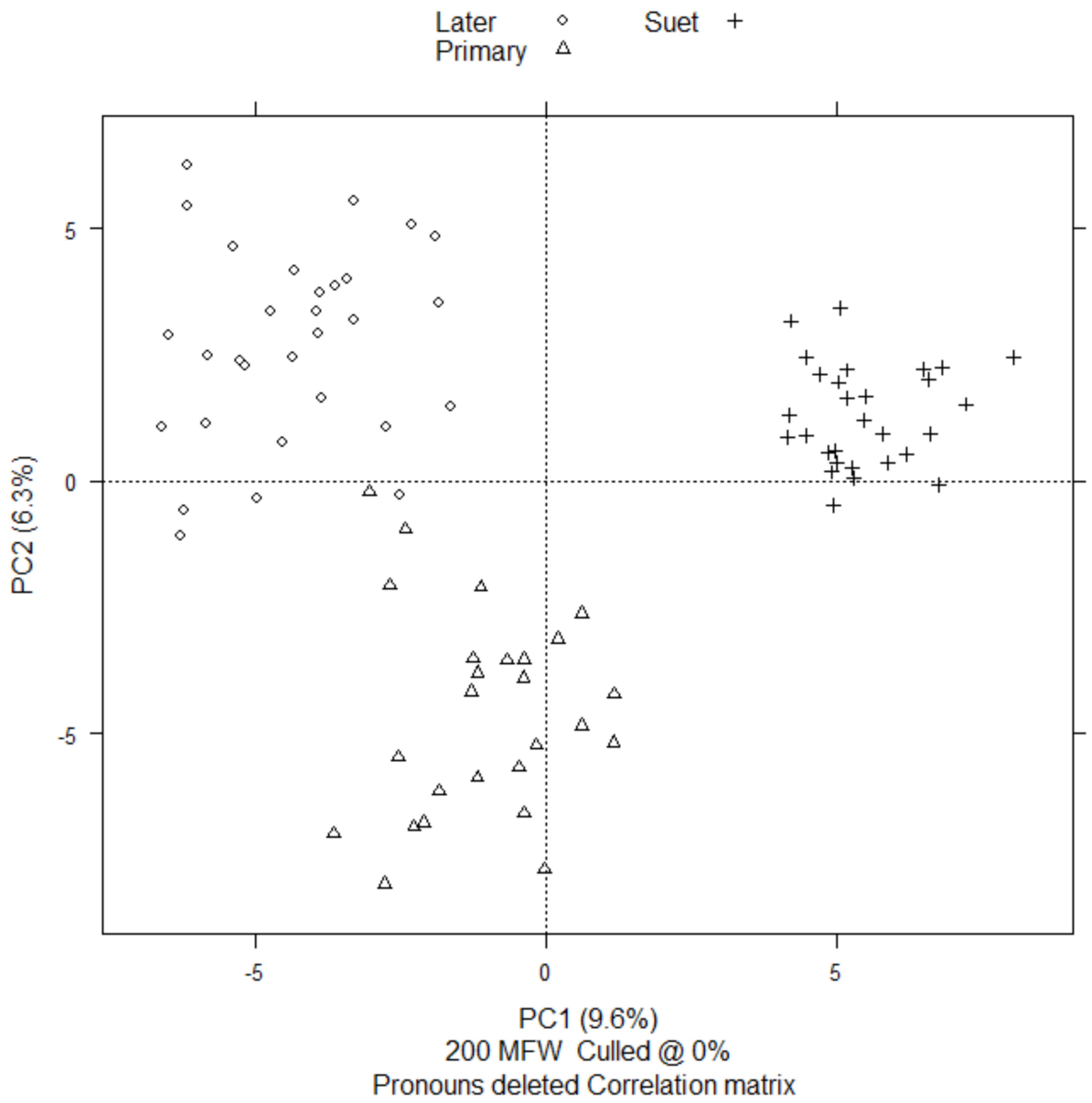

Figure 5

Examining the feature loadings, that is, the actual lexical elements from which the PCA scatterplot is constructed, gives us some indication of the basis for this division.

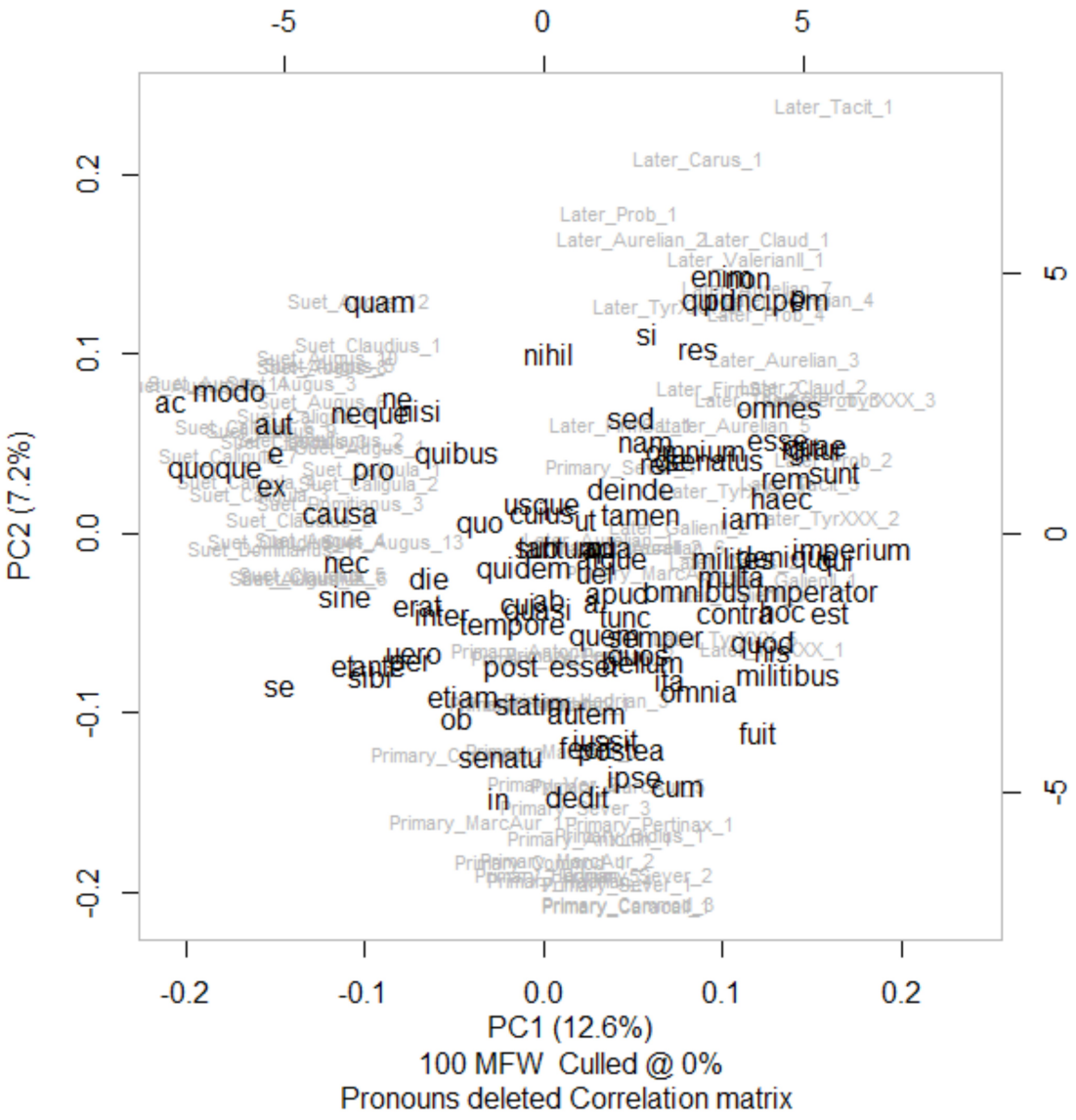

Figure 6: Note only 100 MFW, for the sake of the visualisation of the loadings

Words preferred by the *Hauptviten* include *vero*, *etiam*, *autem*, and *statim*; words disfavoured include most notably *enim*, but also *sed, deinde*, *tamen*, and *iam.* To give two examples: we find *enim* at a frequency of about seven times every ten thousand words (.00073) in the *Hauptviten*, whereas we find it twenty times every ten thousand words in the later lives (.0020). For *autem*, by contrast, it occurs four times every thousand words (.0045) in the *Hauptviten* and just twice every thousand in the later lives (.0021). In other words, the later lives use *enim* three times as

often as the *Hauptviten*, and *autem* half as often. By themselves these two points would be hardly probative, but it is from these kinds of usages analyzed for the two hundred most frequent words that the plots above are constructed.

VI CONCLUSIONS

From this analysis the following conclusions emerge:

(1) *Pace* Tse, Tweedie, and Frischer, the authorial structure presented by the manuscripts - six authors, four working before the lacuna and two after - is not compatible with the features of the text, given the strong results we have achieved from the GI verification.

(2) Certain stylistic features of the corpus point toward single authorship, and these indications are complementary to the analysis by J. N. Adams who found an unobtrusive but curious stylistic tic throughout the corpus.[37]

(3) At the same time, two distinct authorial layers seem to be present which correspond, more or less, to the categories of *Hauptviten* and later lives. The *Nebenviten* and the intermediate lives seem to contain a style mixed from that of the two authorial layers. These categories can be reconstructed solely on the basis of the stylometric data.

(4) The lacuna after Maximus and Balbinus corresponds to a discernible stylistic break in the text. While that does prove that the lacuna is deliberate, it does provide further indication that its presence is not a happenstance of transmission.

(5) The major exception to these two authorial layers are samples from the lives of Alexander Severus and Heliogabalus which form a distinct and compact clade in Fig. 1 (a darkly coloured square in Lampridius' column in our visualization) and separate from the

[37] J. N. Adams, "On the Authorship of the *Historia Augusta*," *CQ* 22 (1972) 186-194.

other samples in our PCA plot (Fig. 2 and Fig 4). This could mean that these lives are drawn from a third source whose stylistic features have persisted in the text as we have it.

One way to interpret these conclusions is that a single author incorporated an earlier source for the lives of the emperors, making very few changes in the lives of senior emperors through to Didius Iulianus, while adding the lives of junior emperors with material partially of his own composition. After Didius, his collection shows increasing evidence of his own hand at work, until after the lacuna, where the lives seem to be primarily of his own composition. This theory neatly accords with Syme's conclusions about the collection, and gives strong indication that there is no necessary disjunction between the conclusions arising from computational studies and traditional literary and historical analysis.

Appendix: Full probability table resulting from applying a GI analysis to 1000 token samples from the *HA* (cf. Figure 1)

label	capit	lamprid	spart	trebell	vopisc	vulcac
capit-antonin_1	0.24	0.12	0.72	0	0	0
capit-antonin_2	0.06	0.37	0.64	0	0	0
capit-clodalb_1	0.79	0.01	0.19	0	0	0.03
capit-clodalb_2	0.83	0.01	0.21	0.01	0	0
capit-clodalb_3	0.77	0.02	0.08	0	0.24	0.02
capit-clodalb_4	0.75	0.03	0.24	0.01	0	0.01
capit-clodalb_5	0.62	0	0.31	0.02	0.03	0.02
capit-didius_1	0.24	0.01	0.84	0	0	0
capit-gordianiii_1	0.66	0.22	0.25	0.01	0	0
capit-gordianiii_2	0.56	0.02	0.01	0.17	0.29	0
capit-gordianiii_3	0.87	0.06	0.06	0.04	0.05	0
capit-gordianiii_4	0.71	0.17	0.01	0.04	0.18	0
capit-gordianiii_5	0.59	0.04	0.02	0.28	0.11	0
capit-macrin_1	0.05	0.03	1	0.01	0.01	0.04
capit-macrin_2	0.16	0.02	0.95	0.01	0	0
capit-marcaur_1	0.27	0.22	0.66	0	0	0
capit-marcaur_2	0.18	0.34	0.59	0	0	0
capit-marcaur_3	0.23	0.05	0.91	0	0	0.01
capit-marcaur_4	0.26	0.39	0.42	0.02	0	0
capit-marcaur_5	0.38	0.3	0.18	0	0	0.11
capit-maxbalb_1	0.8	0.04	0.11	0.08	0.04	0
capit-maxbalb_2	1	0.01	0.06	0.04	0.02	0
capit-maxbalb_3	0.95	0.03	0.01	0.09	0.02	0
capit-maximinii_1	0.25	0.01	0.69	0.15	0.01	0.01
capit-maximinii_2	0.19	0.28	0.39	0.21	0	0.03
capit-maximinii_3	0.83	0.05	0.02	0.13	0.04	0
capit-maximinii_4	0.96	0.03	0.07	0.02	0	0
capit-maximinii_5	0.89	0.11	0.02	0.01	0.04	0
capit-pertinax_1	0.37	0	0.7	0	0	0
capit-pertinax_2	0.31	0.21	0.59	0	0	0
capit-ver_1	0.46	0.09	0.58	0	0	0
capit-ver_2	0.37	0.11	0.61	0	0	0
lamprid-alexsev_1	0.19	0.61	0.19	0	0	0
lamprid-alexsev_10	0.28	0.51	0.16	0.09	0.03	0
lamprid-alexsev_2	0.11	0.4	0.22	0.01	0.26	0.03
lamprid-alexsev_3	0	0.51	0.12	0.15	0.04	0.26
lamprid-alexsev_4	0.01	0.88	0.28	0	0	0
lamprid-alexsev_5	0.05	1	0.03	0	0	0
lamprid-alexsev_6	0.03	0.98	0.04	0	0	0.01
lamprid-alexsev_7	0.02	0.98	0.07	0	0	0.01
lamprid-alexsev_8	0.07	0.96	0.01	0.01	0	0.08
lamprid-alexsev_9	0.04	0.27	0.07	0.43	0.35	0.02
lamprid-commod_1	0.32	0.02	0.81	0	0	0
lamprid-commod_2	0.13	0.19	0.76	0	0	0

lamprid-commod_3	**0.2**	**0.54**	**0.34**	**0**	**0**	**0**
lamprid-diadum_1	**0.35**	**0.07**	**0.46**	**0.04**	**0.09**	**0**
lamprid-helioga_1	**0.11**	**0.76**	**0.22**	**0.02**	**0**	**0**
lamprid-helioga_2	**0.18**	**0.75**	**0.25**	**0**	**0**	**0.02**
lamprid-helioga_3	**0.33**	**0.45**	**0.43**	**0**	**0**	**0**
lamprid-helioga_4	**0.07**	**1**	**0.02**	**0**	**0**	**0**
lamprid-helioga_5	**0.01**	**1**	**0.04**	**0**	**0**	**0.01**
spart-aelius_1	**0.14**	**0.03**	**0.16**	**0.43**	**0.05**	**0.11**
spart-antget_1	**0.08**	**0.02**	**0.96**	**0.01**	**0**	**0.04**
spart-antget_2	**0.48**	**0.01**	**0.47**	**0.01**	**0**	**0.01**
spart-antget_3	**0.87**	**0.03**	**0.2**	**0.01**	**0**	**0**
spart-caracall_1	**0.15**	**0.02**	**0.96**	**0**	**0**	**0**
spart-caracall_2	**0.22**	**0.08**	**0.62**	**0.04**	**0**	**0.05**
spart-hadrian_1	**0.18**	**0.01**	**0.93**	**0**	**0**	**0**
spart-hadrian_2	**0.18**	**0.23**	**0.71**	**0**	**0**	**0**
spart-hadrian_3	**0.19**	**0.34**	**0.59**	**0**	**0**	**0**
spart-hadrian_4	**0.02**	**0.95**	**0.11**	**0**	**0**	**0.01**
spart-hadrian_5	**0.09**	**0.02**	**0.98**	**0**	**0**	**0**
spart-pescenn_1	**0.28**	**0**	**0.7**	**0.02**	**0**	**0**
spart-pescenn_2	**0.03**	**0.32**	**0.21**	**0.36**	**0.09**	**0.1**
spart-sever_1	**0.21**	**0**	**0.93**	**0**	**0**	**0**
spart-sever_2	**0.11**	**0.01**	**0.99**	**0**	**0**	**0**
spart-sever_3	**0.28**	**0.05**	**0.83**	**0**	**0**	**0**
spart-sever_4	**0.12**	**0.17**	**0.74**	**0.01**	**0.01**	**0.03**
trebell-claud_1	**0**	**0**	**0.01**	**0.21**	**0.9**	**0**
trebell-claud_2	**0.02**	**0**	**0.01**	**0.94**	**0.1**	**0.01**
trebell-galienii_1	**0.01**	**0**	**0.03**	**1**	**0.05**	**0**
trebell-galienii_2	**0.1**	**0.04**	**0.13**	**0.77**	**0.06**	**0.04**
trebell-galienii_3	**0.05**	**0.06**	**0.02**	**0.89**	**0.08**	**0**
trebell-tyrxxx_1	**0.02**	**0**	**0.09**	**0.99**	**0.06**	**0**
trebell-tyrxxx_2	**0**	**0**	**0.01**	**0.97**	**0.14**	**0.05**
trebell-tyrxxx_3	**0**	**0.01**	**0**	**1**	**0.12**	**0**
trebell-tyrxxx_4	**0.05**	**0.01**	**0.07**	**0.91**	**0.07**	**0.02**
trebell-tyrxxx_5	**0.02**	**0**	**0.07**	**0.96**	**0.05**	**0**
trebell-tyrxxx_6	**0**	**0**	**0**	**0.84**	**0.21**	**0**
trebell-valerianii_1	**0.05**	**0.01**	**0**	**0.36**	**0.69**	**0.01**
vopisc-aurelian_1	**0.01**	**0.09**	**0.15**	**0.1**	**0.61**	**0**
vopisc-aurelian_2	**0**	**0**	**0**	**0.28**	**0.76**	**0.15**
vopisc-aurelian_3	**0**	**0**	**0**	**0.18**	**0.95**	**0**
vopisc-aurelian_4	**0**	**0**	**0**	**0.24**	**0.77**	**0.04**
vopisc-aurelian_5	**0**	**0**	**0**	**0.25**	**0.84**	**0.02**
vopisc-aurelian_6	**0.02**	**0**	**0.42**	**0.55**	**0.15**	**0**
vopisc-aurelian_7	**0**	**0**	**0**	**0.22**	**0.9**	**0**
vopisc-carus_1	**0**	**0**	**0**	**0.4**	**0.72**	**0.02**
vopisc-carus_2	**0.02**	**0.01**	**0.01**	**0.93**	**0.16**	**0**
vopisc-firmsat_1	**0.1**	**0.09**	**0**	**0.13**	**0.64**	**0.12**
vopisc-firmsat_2	**0.02**	**0**	**0.01**	**0.08**	**0.92**	**0.04**
vopisc-prob_1	**0**	**0**	**0**	**0.08**	**0.96**	**0.01**
vopisc-prob_2	**0.02**	**0.01**	**0**	**0.09**	**0.95**	**0.03**
vopisc-prob_3	**0**	**0.01**	**0**	**0.19**	**0.84**	**0**
vopisc-prob_4	**0**	**0**	**0.01**	**0.53**	**0.58**	**0**
vopisc-tacit_1	**0**	**0**	**0**	**0.33**	**0.67**	**0**

vopisc-tacit_2	**0.02**	**0.1**	**0.01**	**0.07**	**0.93**	**0.01**
vopisc-tacit_3	**0**	**0**	**0.01**	**0.29**	**0.89**	**0**
vulcac-avid_1	**0.02**	**0.14**	**0.03**	**0.15**	**0.22**	**0.62**
vulcac-avid_2	**0.35**	**0.14**	**0.28**	**0.01**	**0.04**	**0.17**