Reassessing the Apuleian Corpus : a computational approach to authenticity

Reference:

Stover J.A., Kestemont Mike.- Reassessing the Apuleian Corpus : a computational approach to authenticity
The classical quarterly - ISSN 1471-6844 - (2016), p. 1-44

REASSESSING THE APULEIAN CORPUS: A COMPUTATIONAL APPROACH TO AUTHENTICITY

The renaissance of Apuleian studies of the past few decades shows no signs of abating.[1] The summer of 2014 may well be the highest water mark yet recorded in the tide of interest in Apuleius: June and July alone saw the release of two monographs, one each from Oxford University Press and Cambridge, and one edited conference volume, from Routledge.[2] The clearest sign that the sophist of Madauros has come into his own is his admission into the exclusive club of the *Oxford Classical Texts*: the first volume of his complete works containing the *Metamorphoses* edited by Maaike Zimmerman came out in 2012. One of the

[1] This is not the place to provide a complete bibliography of Apuleius; nonetheless, a few of the more important monographs should be noted. Contemporary Apuleian studies take off from J. Winkler's monograph, *Auctor and actor: A narratological reading of Apuleius's* Golden Ass (Berkeley, 1985). Recent studies of the *Met.* include R. May, *Apuleius and drama: The ass on stage* (Oxford, 2006); L. Graverini, *Le Metamorfosi di Apuleio: Letteratura e identità*. (Pisa, 2007) and S.A. Frangoulidis, *Witches, Isis and narrative: Approaches to magic in Apuleius'* Metamorphoses (Berlin, 2008). On the reception of Apuleius, see R.H.F. Carver, *The Protean Ass: The Metamorphoses of Apuleius from Antiquity to the Renaissance* (Oxford, 2007) and J. H. Gaisser, *The Fortunes of Apuleius and the* Golden Ass: *A Study in Transmission and Reception* (Princeton, 2008).

[2] These are S. Tilg, *Apuleius' Metamorphoses: A Study in Roman Fiction* (Oxford, 2014); R. Fletcher, *Apuleius' Platonism: The Impersonation of Philosophy* (Cambridge, 2014); and B.T. Lee et al. (edd.), *Apuleius and Africa* (New York, 2014). Fletcher's monograph appeared too late for us to use it in this study.

most salutary effects of this renewed interest has been the re-appraisal of the 'whole Apuleius': Apuleius has more to offer than just the *Metamorphoses*, and recent scholarship on the *rhetorica* and the *philosophica* have shown not only how these *opera minora* can help us understand the *opus maius*, but also how they are important and interesting documents in their own right.[3]

Perhaps then it is an auspicious time to revisit some old questions that have bedevilled scholarly treatments of the 'whole Apuleius.' Which texts should be accepted as authentic and which should be rejected? The Apuleian corpus is riddled with vexing problems of authenticity and transmission. To give an incomplete list:

(1) Is the poem ascribed to Apuleius in the *Anthologia Latina* (712 Riese) authentic?[4]

(2) Is the *spurcum additamentum* found in the margin of the most important manuscript of the *Metamorphoses* (10.21) authentic?[5]

---

[3] On the 'whole Apuleius', see B.L. Hijmans, Jr., '*Apuleius Philosophus Platonicus*', *ANRW* 2.36.1 (1987), 395–475; G. Sandy, *The Greek World of Apuleius: Apuleius and the Second Sophistic* (Leiden, 1997); S.J. Harrison, *Apuleius: A Latin Sophist* (Oxford, 2000); and now Fletcher (n. 2). On the opera minora, see C. Marangoni, *Il mosaico della memoria: Studi sui Florida e sulle Metamorfosi di Apuleio* (Padua, 2000) and M. Baltes, et al., *Apuleius: De deo Socratis. Über den Gott des Sokrates* (Darmstadt, 2004).

[4] For a positive view, see S.J. Harrison, 'Apuleius eroticus: *Anth. Lat*. 712 Riese', *Hermes* 120 (1992), 83–9.

[5] On the positive side, see E. Lytle, 2003. 'Apuleius' *Metamorphoses* and the *spurcum additamentum* (10.21)', *CPh* 98 (2003), 349-65; for a response, see V. Hunink, 'The *spurcum additamentum* (Apul. *Met*. 10,21) Once Again', in W.H. Keulen et al. (edd.), *Lectiones*

(3) While it is almost universally believed that the *Herbarius* and the *Physiognomia* are not by Apuleius, how did they acquire their ascription?[6]

(4) How did a high medieval political tract called *De monarchia* come to be ascribed to Apuleius?[7]

(5) What is the status of the rhetorical fragments transmitted at the beginning of the *De deo Socratis*, the so-called 'False Preface'?[8]

(6) Are the *De Platone et eius dogmate* and the *De mundo* (both undoubtedly by the same author) authentic?[9]

---

scrupulosae. Essays On the Text and Interpretation of Apuleius' Metamorphoses In Honour of Maaike Zimmerman (Groningen, 2006), 266-79. Zimmerman, in the introduction to her OCT (Oxford, 2012), provides a full discussion at xxiii-xxv.

[6] On the *Herbarius*, see G. Maggiulli and M.F. Buffa Giolito, *L'altro Apuleio. Problemi aperti per una nuova edizione dell' Herbarius* (Naples, 1996), and V. Hunink, 'Apuleius and the *Asclepius*', *Vigiliae Christianae* 50 (1996), 288-308 at 300-1; for the *Physiognomia*, see Hunink, ibid. 301, considering points raised by F. Opeku, 'Physiognomy in Apuleius', in C. Deroux (ed.), *Studies in Latin Literature and Roman History* I (Brussels, 1979), 467-74.

[7] See B. Kohl and N. Siraisi, 'The *De monarchia* Attributed to Apuleius', *Medievalia* 7 (1981), 1-39; and Gaisser (n. 1), 122-4.

[8] V. Hunink maintains that the 'False Preface', though truncated, is an integral part of the *DdS* ('The prologue of Apuleius' *De deo* Socratis', *Mnemosyne* 48 [1995], 292-312); most other scholars, e. g. Harrison (n. 3), 91-2, have grouped it with the *Florida*.

[9] The bibliography on this question is vast: for orientation, see Harrison (n. 3), 174-180. The most substantial analyses remain those of J. Redfors, *Echtheitskritische Untersuchungen der*

(7) Is the *Asclepius* authentic, and if not, how and when did it become an interloper in the philosophical corpus?[10]

(8) Is the *Peri hermeneias* authentic, and if so, why does it have a separate transmission?[11]

---

apuleischen Scriften De Platone und De mundo (Lund, 1960), who concludes that the problem is insoluble, and A. Marchetta, *L'autenticità apuleiana del De Mundo* (Rome, 1991), who favours authenticity for the *De mundo* and (by extension) for the *De Platone*. Doubts as to the authenticity of these works, while more muted than in decades past, have been raised as recently as 2007 by N. Holmes, 'False Quantities in Vegetius and Others', *CQ* 57 (2007), 668-86, at 684-6.

[10] The question was re-opened after decades of consensus by Hunink (n. 6); his arguments were responded to by M. Horsfall Scotti, 'The *Asclepius*: Thoughts on a Re-Opened Debate', *Vigiliae Christianae* 54 (2000), 396-416.

[11] The case was put forward most vigorously by D. Londey and C. Johanson, *The Logic of Apuleius* (Leiden, 1987), 8-15. B.T. Lee cautiously accepts the authenticity of the text, and provides the relevant bibliography in his commentary on the *Florida* (Berlin, 2005), 10-11; Harrison (n. 3), 11, rejects it.

(9) What of the so-called *Summarium librorum Platonis* transmitted after the *De mundo* in one thirteenth-century manuscript discovered by Raymond Klibansky?[12]

One could go on. There are few ancient Latin authors who give rise to as many problems as Apuleius, and the authenticity of various works in the corpus have given rise to some of the best detailed philological treatments of questions of authorship in Latin literature. In 1960, Josef Redfors composed an exhaustive study on the question of the authenticity of the *De Platone* and *De mundo*; one major component of that analysis was minute lexical study of particles and other Latin function words. Ultimately, he could reach no conclusion: his analysis uncovered too many contradictory indications to point definitively one way or another. Fifteen years later, in a review of Beaujeu's 1973 Budé edition, Michael McGann cautiously proposed a way out of Redfors's impasse, noting that 'there is perhaps room for a statistical approach to the problem'.[13] More than a decade later still, Londey and Johanson stated much the same thing with regard to the *Peri hermeneias*: 'It is possible that more compelling grounds for accepting or rejecting Apuleain authorship of the *Peri Hermeneias* may eventually emerge from stylometric studies.'[14]

---

[12] See the *Proceedings of the British Academy. Annual Report, 1948-1949* (London 1949), 8. The manuscript is Vatican City, Reg. lat. 1572. For a full discussion of this text, an *editio princeps*, and arguments in favour of its authenticity, see J.A. Stover, *A New Work by Apuleius* (Oxford, forthcoming).

[13] M.J. McGann, *CR* 25 (1975), 226-7, at 227.

[14] Londey and Johanson (n. 11), 17.

Taking up these two decades-old challenges, we will examine the last four questions of those posed above, using the methods of computational philology for authorship attribution. Our intention is not to supplant other modes of analysis, but to shed new light on old problems by using new tools based on old methods. In so doing, we will show how current computational methods can be employed for authorship attribution and authentication in ancient Latin texts, an area that has been surprisingly quiet ever since the controversy over the *Historia Augusta* in the 1980s and 1990s.[15]

METHODOLOGY

The present study's methodology is drawn from the field of computational stylistics or 'stylometry', a vibrant multidisciplinary research domain within Digital Humanities (or Humanities Computing). In this field, scholars and scientists study the writing style of (literary) texts through advanced statistical quantification, typically relying on computational means.[16] Interesting applications of stylometry in literary studies include plagiarism

[15] See S. Hockey, 'An Agenda for Electronic Text Technology in the Humanities', *CW* 91 (1998), 521-42, esp. 524-5. The studies on the *Historia Augusta* include I. Marriott, 'The Authorship of the *Historia Augusta*: Two Computer Studies,' *JRS* 69 (1979), 65-77; B. Meissner, 'Computergestützte Untersuchungen zur stilischen Einheitlichkeit der *Historia Augusta*' in G. Bonamente and K. Rosen (edd), *Historiae Augustae colloquium Bonnense* (Bari 1997), 175-215; and E. Tse, F. J. Tweedie, and B. Frischer, 'Unravelling the Purple Thread: Function Word Variability and the *Scriptores Historiae Augustae*,', *Literary and Linguistic Computing* 13 (1998), 141-149.

[16] D. Holmes, 'The Evolution of Stylometry in Humanities scholarship', *Literary and Linguistic Computing* 13 (1998), 111-17. For Digital Humanities in general, see, *inter alia*, S.

detection, as well as in 'stylochronometry', or establishing the sequence of an author's works, such as in the well-known studies of the Platonic dialogues.[17] Authorship attribution, where scholars attempt to automatically determine the identity of a (possibly anonymous) text's author by inspecting its stylistic characteristics, nevertheless, remains the most popular branch of stylometry.[18] The fundamental assumption underlying this field is that "by measuring some textual features we can distinguish between texts written by different

---

Schreibman, R. Siemens and J. Unsworth (edd.), *A Companion to Digital Humanities* (Oxford, 2004).

[17] L. Brandwood, *Stylometric Method and the Chronology of Plato's Works* (Cambridge, 1990). For stylochronometry in general, consult the survey in: C. Stamou, 'Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating', *Literary and Linguistic Computing* 23 (2008), 181-99.

[18] Recent surveys of the field include: P. Juola, 'Authorship attribution', *Foundations and Trends in Information Retrieval* 1 (2006), 233-334; M. Koppel, J. Schler and S. Argamon, 'Computational Methods in Authorship Attribution', *Journal of the American Society for Information Science and Technology* 60 (2009), 9-26; E. Stamatatos, 'A Survey of Modern Authorship Attribution Methods', *Journal of the American Society for Information Science and Technology* 60 (2009), 538-556. An inspiring recent contribution is J. Burrows, 'A Second Opinion on 'Shakespeare and Authorship Studies in the Twenty-First Century', *Shakespeare Quarterly* 63 (2012), 355-92.

authors",[19] an idea which has found its most ambitious (and tendentious) formulation as the 'Universal Stylome Hypothesis'.[20]

Empirical research demonstrates that a variety of computational techniques can in many cases discriminate between the writing styles of distinct authors.[21] In experiments where computers are 'trained' on example material from candidate authors, algorithms achieve outstanding performance in attributing previously unseen texts to the correct author, solely based on the target author's stylistic characteristics. Although the prerequisites for such attributions should not be underestimated (one should, for example, have enough example material per author[22]), these techniques can be put to interesting use in philological research. In addition to a series of more technical inquiries focusing on methodological issues,[23] recent literary studies have applied stylometry to diverse domains, including

[19] Stamatatos (n. 18), 538.

[20] H. van Halteren, H. Baayen, F. Tweedie, F., M. Haverkort, M. and A. Neijt, 'New Machine Learning Methods Demonstrate the Existence of a Human Stylome', *Journal of Quantitative Linguistics* 12 (2005), 65-77.

[21] A good methodological survey is offered by Stamatatos (n. 18).

[22] See e.g. K. Luyckx and W. Daelemans, 'The effect of author set size and data size in authorship attribution', *Literary and Linguistic Computing* 26 (2011), 35-55 or M. Eder, 'Does size matter? Authorship attribution, small samples, big problem', *Literary and Linguistic Computing* (forthcoming, advance access via doi:10.1093/llc/fqt066).

[23] M. Eder and J. Rybicki, 'Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?', *Literary and Linguistic Computing* (2011), 315-21.

nineteenth-century German literature, French Enlightenment theatre and twelfth-century Latin literature.[24]

Apart from the advanced degree of computational quantification which is currently common in stylometry, the field shows a number of major differences with respect to earlier practices of authorship attribution.[25] The most important methodological innovation in the field can be traced back to the investigations of the disputed authorship of the pseudonymous *Federalist Papers* by the American statisticians Mosteller and Wallace in the 1960s.[26] In their influential study, Mosteller and Wallace argued that for authorship attribution, scholars should move away from a text's conspicuous characteristics (uncommon nouns, for example, or rare syntactical constructions), which until then had been the customary focus of stylistic inquiry in attribution studies. Instead, they proposed to study a text's most common, yet most inconspicuous, components: its function words.

---

[24] See respectively F. Jannidis and G. Lauer, 'Burrows's Delta and Its Use in German Literary History', in M. Erlin and L. Tatlock (edd.), *Distant Readings. Topologies of German Literature in the Long Nineteenth Century* (Woodbridge, 2014), 29-54; C. Schöch, 'Fine tuning our stylometric tools: Investigating authorship, genre, and form in French classical theatre', in *Digital Humanities 2013: Conference Abstracts* (Lincoln, NE, 2013), 383-6; M. Kestemont, S. Moens and J. Deploige, 'Collaborative Authorship in the Twelfth Century. A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux', *Digital Scholarship in the Humanities* 30 (2015), 199-224.

[25] The broad field of authorship attribution has been surveyed by H. Love, *Attributing Authorship. An Introduction* (Cambridge, 2002).

[26] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist* (Cambridge, MA, 1964).

There is widespread acceptance that a text's function words (for Latin, this applies primarily to particles, prepositions, conjunctions, and some adverbs) offer an exceptionally privileged stylistic category for the stylistic study of authorship.[27] These items are frequent throughout all texts; hence, they yield a statistically reliable base for textual comparison. They are fairly independent from the genre or theme of a text, which is why they are attractive for determining authorship across generic and thematic boundaries. Finally, an often-heard, yet slightly more controversial, claim is that the use of these words, at least when studied on a larger textual scale, are not under an author's conscious control. This is an important idea, because this aspect of function words would make them resistant to imitation (by students or epigones, for example) and outright forgery. The idea underlying the use of function words is that we attempt to reduce texts to a set of features that differ in nothing besides authorship.[28]

In this paper we shall apply a stylometric methodology to a corpus of Latin prose texts from antiquity, in particular texts which stand in some relation to Apuleius. Our corpus includes philosophical texts by Seneca and Cicero; works by authors of the generation preceding Apuleius, including Suetonius and Pliny the Younger; works of Apuleius' contemporaries such as Aulus Gellius and Tertullian, and works by authors in the generation

[27] Accessible surveys of this idea can be found in J. Binongo, 'Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution', *Chance* 16 (2003), 9-17; M. Kestemont, 'Function Words in Authorship Attribution: From Black Magic to Theory?', in A. Feldman, A. Kazantseva and S. Szpakowicz (edd.), *Proceedings of the Third Workshop on Computational Linguistics for Literature Workshop, co-located with the 14th Conference of the European Chapter of the Association for Computational Linguistics* (Gothenburg, 2014), 59-66.

[28] Cf. Juola (n. 18), 264-5.

following Apuleius, such as Cyprian. We have restricted our analyses to the first 9,000 words of each text, in order not to let shorter texts be too heavily outbalanced by some of the longer texts in the corpus, such as the *Met.* Our methodology will solely consider the most common words in the texts analysed.

We have not lemmatized the texts, or applied word stemming to them: our earlier exploratory experiments (which we do not report for the sake of brevity) showed that an approach based on plain, inflected surface tokens, generally yielded much more stable result than lemma-based approaches. More details on the pre-processing are given below. In each of the analyses described below, we have automatically extracted those words which had the highest cumulative frequency in the texts analysed. As will be illustrated below, these lists are typically dominated by (uninflected) function words, such as particles and prepositions. Nevertheless, these lists will occasionally also include high-frequency inflected word forms such as *est*, which typically serve a grammatical function and are thus suited for stylometric analysis. Note that for all the analyses reported below, we have also automatically deleted all frequency information related to personal pronouns in texts, a procedure known as pronoun culling.[29] This removal is meant to minimize the stylistic influence of narrative perspective and also genre to some extent, which are often betrayed by the personal pronouns in a text and thus interfere with analyses which focus on authorship.[30] To normalize any differences in orthography we have also replaced all every *v* with *u* in the corpus (irrespective of their representing vowels or not).

---

[29] Highly relevant in this respect are D. Hoover, 'Frequent Collocations and Authorial Style', *Literary and Linguistic Computing* 18 (2003), 261-86 and D. Hoover, 'Multivariate Analysis and the Study of Style Variation', *Literary and Linguistic Computing* 18 (2003), 341-60.

[30] This aspect as well as other potential shortcomings of function words are discussed by Kestemont (n. 27).

Although this is by no means the first application of stylometry to classical Latin texts, the methodology is still not considered mainstream in contemporary classical scholarship. In the first sections below, we will therefore demonstrate that our method can yield valid results with texts of undisputed authorship, before moving on to the disputed works of Apuleius. We will deliberately adopt a non-technical and introductory language that allows the broader readership of this journal to follow the main argument in our paper.[31]

CLUSTERING

All our analyses below are based on a text representation that is called a 'bag-of-words model' in fields such as information retrieval and computational linguistics.[32] For each text in a corpus, our text representation or model will first lowercase the text, remove all punctuation, and then split the text into individual words along white space. The resulting list

[31] All our experiments reported in this paper can be easily replicated using the 'Stylometry with R' package, a suite of software scripts for the popular statistical *R* program (http://www.r-project.org/). This package is freely available online in the public domain and is presented by the suite's main developers (the Computational Stylistics Group) in M. Eder, M. Kestemont and J. Rybicki, 'Stylometry with R: a suite of tools', in *Digital Humanities 2013: Conference Abstracts* (Lincoln, NE, 2013), 487-9. A manual for the package can be found on the group's website: https://sites.google.com/site/computationalstylistics/. We have shared a version of our corpus in an online repository (https://github.com/mikekestemont/Apuleius), excluding the texts by Tertullian and Cyprian, which are proprietary data owned by Brepols Publishers (*Library of Latin Texts*). We wish to acknowledge Brepols Publishers for the use of this proprietary material.

[32] C. Manning, P. Raghavan and H. Schütze, *An Introduction to Information Retrieval* (Cambridge, 2008).

of words will then be used to create a frequency table, reminiscent of the way tabular information is displayed in standard spreadsheet applications: the frequency table will have a column for every word that has been attested in the corpus and a row for every text in the collection. The cells in the table will be populated by the relative frequency of every word in every text. All subsequent analyses are then applied to this frequency table only. The bag-of-words model has a number of obvious shortcomings from the point of view of stylistic analysis, the most important one being that the original word order in the document is completely lost under this kind of (seemingly superficial) text representation.[33] Nevertheless, numerous empirical tests have shown that this kind of text model is both extremely efficient and extremely effective for the study of authorship.

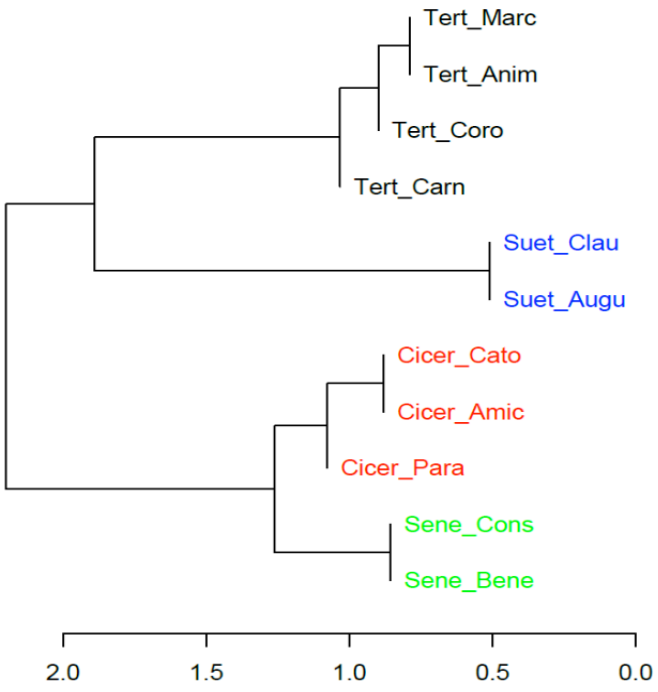


FIGURE 1

[33] See e.g. Stamatatos (n. 18) and Koppel, Schler and Argamon (n. 18), but also W. Daelemans, 'Explanation in Computational Stylometry', in A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing* (Berlin and Heidelberg, 2013), 451-62.

FIGURE 1 shows an example of a dendrogram, the typical output of the first kind of statistical analysis we will discuss here: hierarchical cluster analysis.[34] With this procedure, we pair each text with every other text and calculate the stylistic distance between each pair. These distances are calculated on the basis of the relative frequencies of the most frequent words (MFW) in the entire text collection. Here and throughout, we use a classic distance metric in stylometry: Burrows's Delta, a relatively simple yet demonstrably effective metric for estimating the stylistic distance between texts on the level of authorship.[35]

Burrows's Delta is determined as follows. Suppose that we wish to calculate the deltas between all text pairs in given text collection on the basis of 30 high-frequency words. First, we select the 30 items which have the highest cumulative frequency in the entire text collection and restrict our subsequent distance calculations to them. Next, we compute the relative frequency of each of these items in each text (if a word occurs 5 times in a text that counts 200 words in total, its relative frequency in that text is 0.025). As such, each text will be represented by a list of 30 numerical values. We then calculate the standard deviation of

[34] A good introduction to the advantages and disadvantages of cluster analyses in stylometry can be found in M. Eder, 'Computational Stylistics and Biblical Translation: How Reliable can a Dendrogram be?' in T. Piotrowski and Ł. Grabowski (edd.), *The translator and the computer* (Wrocław, 2013), 155-70.

[35] This metric was introduced in J. Burrows, '"Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship', *Literary and Linguistic Computing* 17 (2002), 267-87. An interesting theoretical discussion is: S. Argamon, 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations', *Literary and Linguistic Computing* 23 (2008), 131-47. Argamon showed in this paper that Burrows's original distance formula can be greatly simplified, both mathematically and conceptually: we have based our discussion of Burrows's Delta in the main text on Argamon's simplified interpretation.

each word's relative frequency in all of the texts analysed, which yields a list of 30 standard deviations, one for each word analysed. (The standard deviation is a statistical measure which captures how strongly the values in a list of values diverge from their mean value.) Finally, we calculate the actual stylistic distance between two texts A and B, by determining the absolute difference in the relative frequency of each word in the two texts, and weigh that absolute difference by dividing it by the standard deviation associated with that word. Burrows's Delta is the sum of the resulting thirty weighted differences. Because of the standard deviation weighting, Burrows's Delta tunes down the contribution of words whose frequencies display significant fluctuation in the text collection. This reduces the effect of content-specific items and focuses the analysis more closely on style-related lexical items.

The resulting distances ('deltas') between all texts are then collected in a large distance table that serves as the input for the actual cluster analysis. The result of a cluster analysis is especially useful to visualize at a glance which texts are more alike than others. One can think of this kind of cluster analysis as a 'bottom-up' procedure, which is data-driven, instead of being guided towards a specific solution by the researcher, what is called an 'unsupervised' technique. First, the cluster analysis will determine which two texts in the corpus are closest to each according to the distance table.[36] It will then merge these two texts into a new, abstract node, representing the 'average' of these two texts in terms of relative word frequencies. In the dendrogram, the analysis will position this new node at a further stage in the tree than the original texts (represented as 'leaves' in the resulting plot). Next, the procedure will work its way up an imaginary tree, iteratively merging these nodes and texts that are most similar to each other, until all texts have been merged at the top tier in the tree.

[36] See e.g. J. Burrows, 'Textual Analysis', in S. Schreibman, R. Siemens and J. Unsworth (edd.), *A Companion to Digital Humanities* (Oxford, 2004), 323-47 at 326.

FIGURE 1 can be thought of as being created right to left: relatively closer text pairs will cluster more to the right. The horizontal difference between two nodes in the graph reflects the distance between nodes: dissimilar texts and nodes will only be joined at a later stage in the procedure (i.e. more to the left) and at a more advanced position in the tree, whereas highly similar groups of texts will tend to form tight clusters from the beginning onwards. The stylistic distance between texts and nodes can therefore be read from the horizontal axis in the cluster plot. FIGURE 1 displays the result of a standard cluster analysis for a fairly random initial selection of texts by four major classical and early Christian authors: Tertullian (*De anima*, *Ad Marcianum*, *De corona militis*, *De carne Christi*), Suetonius (the lives of Claudius and Augustus), Cicero (*Cato maior de senectute*, *Laelius de amicitia*, *Paradoxa stoicorum*) and Seneca (*De constantia*, *De beneficiis*).

As before, the analysis only considers the relative frequencies of the 100 words which are most frequent throughout the texts analysed in this experiment. Moreover, words were only included in this list if they appeared in all of the texts considered here: in technical terms we set the 'culling level' at 100% (i.e. a word's presence in 100% of the texts is required, meaning that it should occur at least once in each text), as an attempt to avoid the interference of content-related artefacts in this analysis. As can be clearly gleaned from the dendrogram, the cluster analysis has no difficulties in grouping the individual texts based on their authorship. This is a remarkable result, because this analysis too is fully unsupervised: it only has access to the 100 MFW frequencies for each text and has no information whatsoever about the provenance of these texts. As one might expect, the oeuvres of Cicero and Seneca seem to have more stylistic affinities and as a group they tend to be relatively different from Tertullian and Suetonius.

An experiment can illustrate how we can use dendrograms for authorship attribution. Say we take Suetonius's *Caligula*, and pretend that it is an anonymous text of which the

authorship is disputed or even unknown. Would a cluster analysis that has no foreknowledge on the text's origin be able to position it under the correct author's branch, even when the analysis only has access to the relative frequencies of the 100 MFW? FIGURE 2 shows that this is indeed the case: the *Caligula* (attributed to the unknown author Q) neatly fits in with the rest the rest of Suetonius's works. This result would give us ample reason to investigate the stylistic similarities between these texts in more depth and indeed – if more corroborating evidence were available – to consider attributing the *Caligula* to Suetonius.

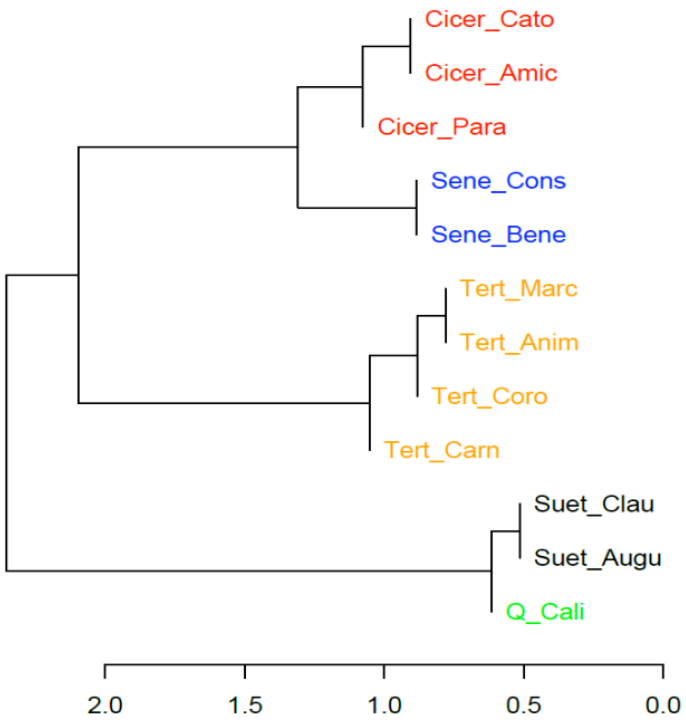


FIGURE 2

Of course, there is no reason whatsoever to doubt the authenticity of the *Caligula* as in the dummy example above. So let us take a slightly more challenging case: the works of Pliny the Younger. His two major works are the *Epistulae* and the *Panegyricus*, but they have entirely separate lines of transmission, and are very different in genre, content and style. Supposing for the moment that the question were still to be settled, what could we learn from a computational analysis?

CONSENSUS TREES

Here, we can introduce a useful extension to traditional cluster analyses in stylometry which takes its inspiration from stylometric studies by Eder.[37] It has been noted that traditional cluster analyses can sometimes be unstable: small changes in the parameters for an experiment (e.g. the exact number of MFW considered) can sometimes yield rather different dendrograms. Therefore it is often helpful to run different cluster analyses for different parameters and combine the results in a single dendrogram. A Bootstrap Consensus Tree (BCT), like the one plotted in FIGURE 3, does exactly this: this (unrooted) dendrogram is based on a series of cluster analysis that are based on different 'frequency bands': a first tree is built for the frequencies of the band of the 1-100 MFW in the texts (band 1), a second tree for the 50-150 MFW band (band 2), etc., all the way up to e.g. the 2950-3000 MFW (the final frequency band). In each of these cluster analyses, we thus perform an experiment on a different slice of high-frequency words, which might introduce subtle discrepancies between individual trees. Note that in each iteration, the same texts are analysed: only the set of words by means of which these texts are represented changes. Next, this series of experiments can be summarized in a BCT like FIGURE 3, which ignores cluster nodes between texts which were not present in at least 50% of the experiments. This type of analysis also considers words with much lower frequencies than the typical function words, which might increase the effect of content-related lexis. Nevertheless, a BCT typically yields very reliable results, because it tests for stylistic similarities across different frequency bands, rendering it relatively insensitive to highly specific, content-related features of texts, such as those in a single specific frequency band.

[37] See Eder (n. 34).

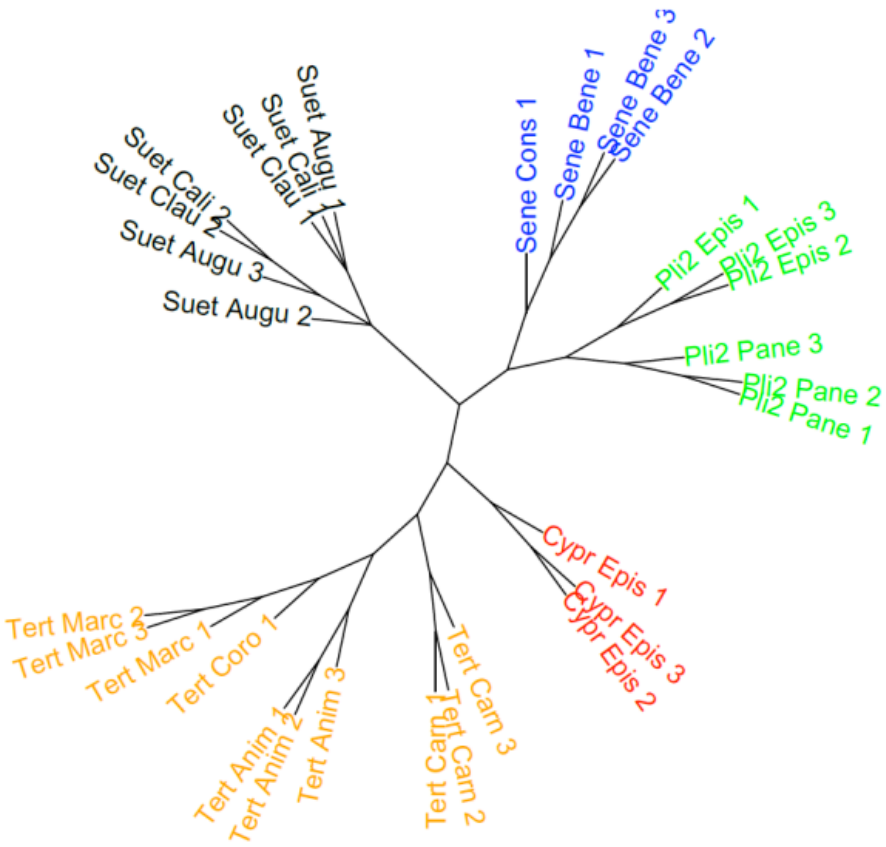


Figure 3

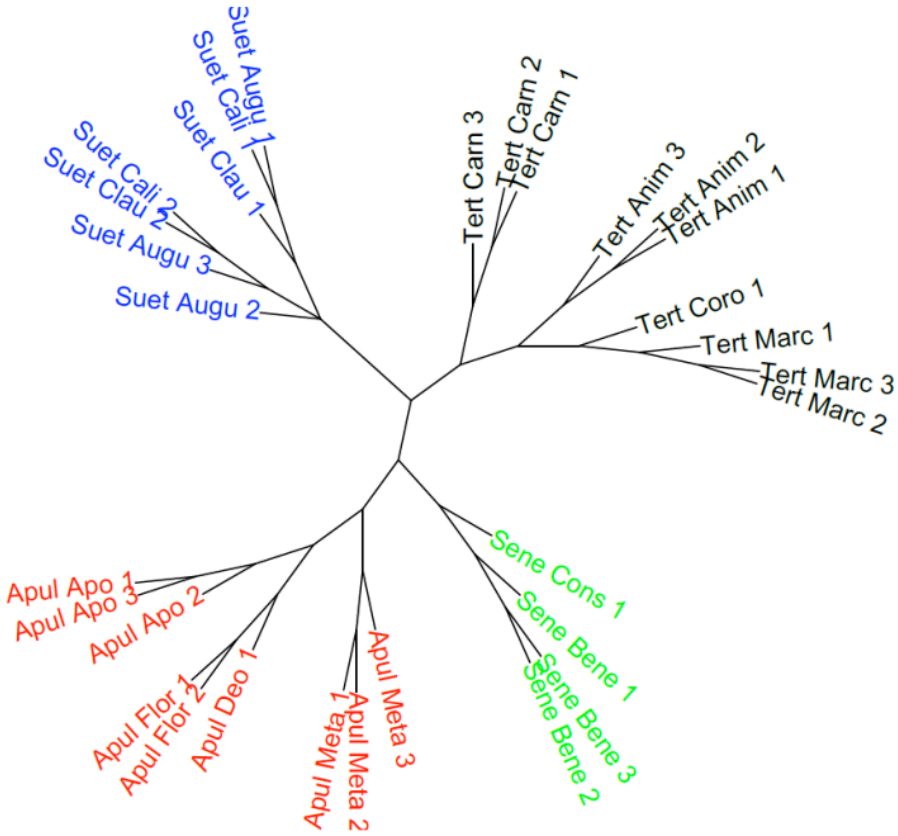
For the experiment in FIGURE 3, we have added both works by Pliny the Younger to the set of works analysed in FIGURE 2. We left out Cicero, however, in order not to overload the analysis and subsequent visualisation with different authors and texts. We added the *Epistulae* by Cyprian to the analysis, however, as a relatively difficult test to assess whether the procedure is able to distinguish between the epistolary production of two authors, which might obviously share many genre-related, rather than author-related, stylistic characteristics. (Note that we did not specify a culling rate now, to be able to obtain a sufficient number of MFW from this varied text collection.) The different frequency bands analysed in FIGURE 3 are 1-100 MFW (band 1), 50-150 MFW (band 2), 100-200, …, 2850-2950, 2900-3000. This analysis includes multiple samples for each text, since it is useful to assess the internal stylistic coherence of texts (here truncated to their 9,000 first words to maximize comparability) into consecutive windows or samples from each text. This will also better

control for the length of the texts analysed. Here, each leaf in the BCT represents a 3,000 word sample from each text (the leaf names are followed by a number indicating the index of the sample).

The branch structure of the dendrogram resulting from a BCT can be read in a manner similar to a standard cluster dendrogram. Although the exact orientation of the circular BCT tree is irrelevant, texts which are stylistically similar will be positioned closer to each other. Additionally, the length of each pair of branches reflects the stylistic distance between the items joined under a particular node. The BCT accurately captures the fact that the samples taken from the *Panegyricus* form a tight cluster and could well be assigned to the same author as the *Letters*, because samples from both texts are, generally speaking, much closer to each other than to any of the other authors studied in this analysis. In other words, the works of Pliny make up a varied but coherent collection of texts.

Similarly, we can use a BCT to show that the works of Apuleius are likewise coherent and therefore suitable for computational analysis. For the analysis in FIGURE 4, we leave out Pliny the Younger, but now include the four works universally accepted as authentically Apuleian: the *Metamorphoses*, the *Apology*, the *De deo Socratis*, and the *Florida*. Here, we see how samples from Apuleius's texts from a tight cluster that is clearly different stylistically than those from the other authors analysed. At this point, we can add *De Platone* and *De mundo*, the two philosophical works whose authenticity is usually but not universally accepted. FIGURE 5 convincingly demonstrates how samples from these two *philosophica* neatly fit in with the other, undisputed texts under Apuleius's clade in the diagram. The attribution of these two works to the Apuleian corpus is extremely stable across a wide variety of algorithmic settings. We have not obtained a single experimental result that would cause one to have any suspicion that these two texts were not written by Apuleius, that is to say the individual responsible for authoring the *Metamorphoses*, the *De deo Socratis*, and the

Florida. Since this analysis rests on the MFW, most of which are inconspicuous function words, we can rule out the possibility of deliberate, skilled imitation. Further, due to the relative length of the two taken together, we have considerable confidence in the robustness



FIGURES 4 and 5

of our results for the *De Platone* and *De Mundo*. This is not always the case for the smaller texts which we will discuss in the next sections, since they offer a less substantial set of features to test.

PRINCIPAL COMPONENTS ANALYSIS

In present-day stylometry, it is common not to limit a study to a single technique, but to compare the output of different methodologies. In the virtual absence of ground truth in historical corpora, especially when it comes to authorship, it is worthwhile to assess the stability of experimental outcomes using different methodologies, which all have their strengths and weaknesses. Thus here we introduce a third technique, called Principal Components Analysis (PCA), which is complementary to the previous clustering approaches. This method is taken from multivariate statistics and has often been successfully applied to authorship attribution.[38] Techniques for the clustering of texts, like the BCT discussed above for instance, are suitable for the stylometric analysis and visualisation of larger corpora, but they have one major drawback: it is difficult for someone using them to find out how specific stylistic characteristics have contributed to the placement of texts in a graph. PCA has the drawback that, as a visualisation technique, it can only be reliably applied to a relatively small set of authors at the same time (typically three or four). It does have the advantage, however, that it tends to give very reliable results for such smaller sets, and it clearly

[38] A seminal application of this technique can be found in J. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels* (Oxford, 1987). Accessible introductions to the application of PCA to authorship attribution include Binongo (n. 27), but also J. Binongo and W. Smith, 'The Application of Principal Components Analysis to Stylometry', *Literary and Linguistic Computing* 14 (1999), 445-66.

visualizes the features from which it is constructed – in the present case, these features are the relative frequencies of the most common words in texts.

Like cluster analysis, PCA operates on the relative frequencies of a small number of MFW in the frequency table of a corpus. PCA will attempt to visualize the main stylistic variation in a text collection by projecting the texts into a two-dimensional scatterplot, in which each text (or sample from it) is plotted as a dot.[39] Generally speaking, the placement of these dots reflects the stylistic similarity or dissimilarity between texts: texts that are close to each other in style will form tight clusters of dots in a plot, whereas distant dots can be said to be more stylistically dissimilar. (Note that the 'dot' is placed at the middle of the text tag.) FIGURE 6 is such a scatterplot resulting from a PCA of 3,000 word samples extracted from the previously analysed texts by Cicero, Seneca and Suetonius. This PCA was applied to the 50 MFW in this corpus. The position of each sample is indicated by a grey label (*Sene_Bene_2*, for example, is the second sample extracted from Seneca's *De beneficiis*). Samples written by the same author form relatively tight clusters, although PCA is also an entirely unsupervised technique. This means that the PCA does not try to place Suetonius's works in a similar region in the plot (it does not even have access to the potentially preconceived meta-information about these texts), but rather lets the frequency information speak for itself.

[39] Here, we will restrict the PCA scatterplots to the first two dimensions (or principal components), which is common in present-day stylometry. Because only so much information can be captured in a two-dimensional analysis, including more than three oeuvres in a PCA should be generally avoided. The underlying theoretical assumption is that, because of this restriction, each dimension has the potential to contrast one author with the other authors included.

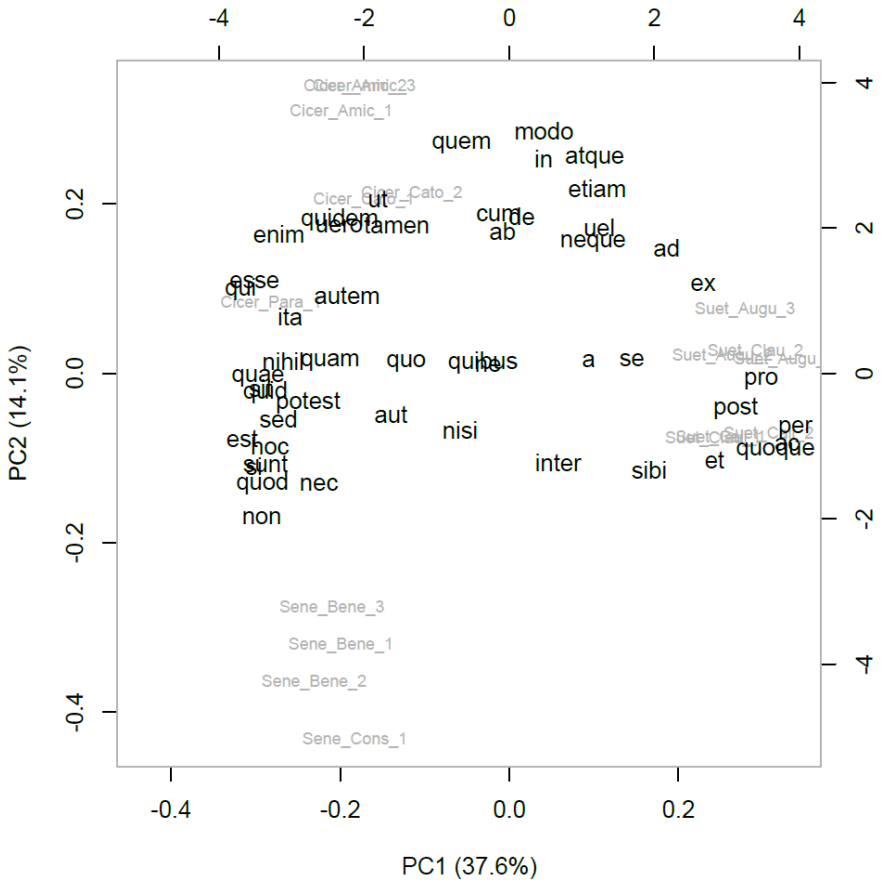


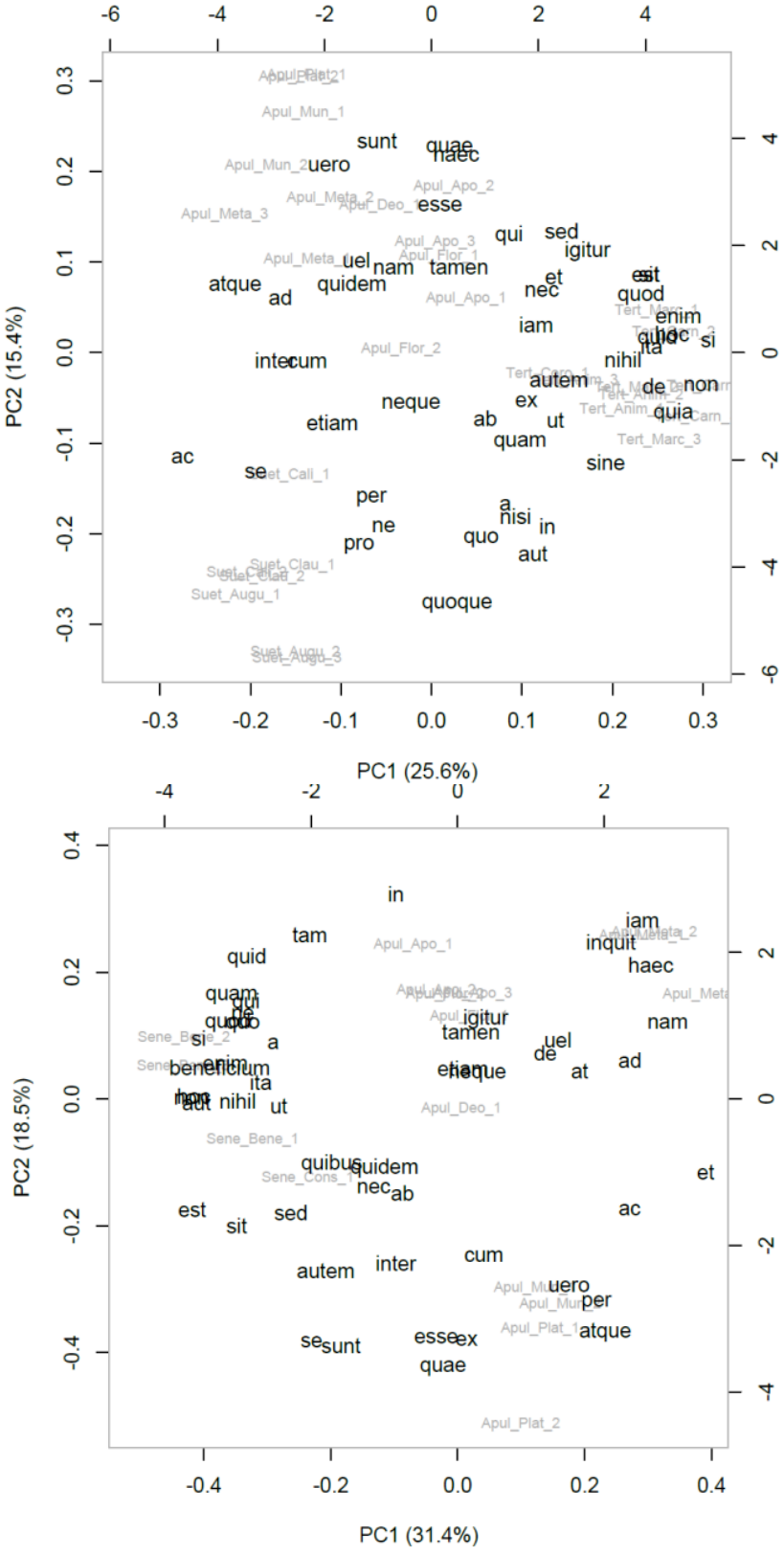
FIGURE 6

In FIGURE 6, note the Latin words scattered across the plot. These so-called 'feature loadings' relate to the original relative word frequency which were fed to the analysis. They offer us an important insight into which word combinations are responsible for the particular placements of the text in the scatterplot. Each feature loading can be thought of as a magnet, so to speak. For example, *et*, is found in the far right half of the plot: therefore, samples which display an elevated frequency of *et* are hence assigned positions in the plot further toward the right edge. Put another way, if the corpus contained only two samples and each sample had an equal frequency of the word *sed*, it would be placed directly between the two samples. If one of these two samples contained *sed* twice as often as the other, *sed* would be placed twice as close to it, and so on. In FIGURE 6, the loadings reveal how Cicero prefers *etiam*, where Suetonius strongly favours *quoque*. This example illustrates how the PCA loadings will often focus on stylistic oppositions that can be found in texts. From the

perspective of studying authorship, the oppositions created between alternatives such as *etiam* and *quoque* are uniquely important, because such subtle oppositions can often be more easily linked to differences in authorial style.

This visualization technique can now be used to inspect which words are typical of the Apuleian corpus, including now the *De Platone* and *De mundo*. Conducting a PCA at 50 MFW of this corpus with, for instance, Suetonius and Tertullian shows visually both its range and coherence (FIGURE 7). The three authors' works cluster relatively tightly together – the Apuleian works are the most widely separated, yet still fairly coherent – and *De Platone* and *De mundo* fall right among the other works by Apuleius, directly between the rhetorical works and the *Metamorphoses*. Visualizing the PCA with the loadings clearly shows some of the characteristic words of Apuleius' lexicon across both the *Met.* and the *rhetorica*. This is extremely clear if we compare the Apuleian corpus with a single control author: in this case, the horizontal opposition in the first component will typically be used will be used to set both oeuvres apart (placing them on the left and the right), whereas the PCA will invariably reserve the vertical spread in the second component to represent the considerable variation we find inside Apuleius's oeuvre. A representative example of this trend can be found in FIGURE 8 where we have compared Apuleius and Seneca (50 MFW): here the threefold, genre-driven stylistic division we typically encounter in Apuleius's texts seems to have been pushed to the extreme by the PCA.

Conducting analyses at various levels of MFW with different parameters and different combinations of control authors always revealed the same effects. Importantly, FIGURES 7



FIGURES 7 and 8

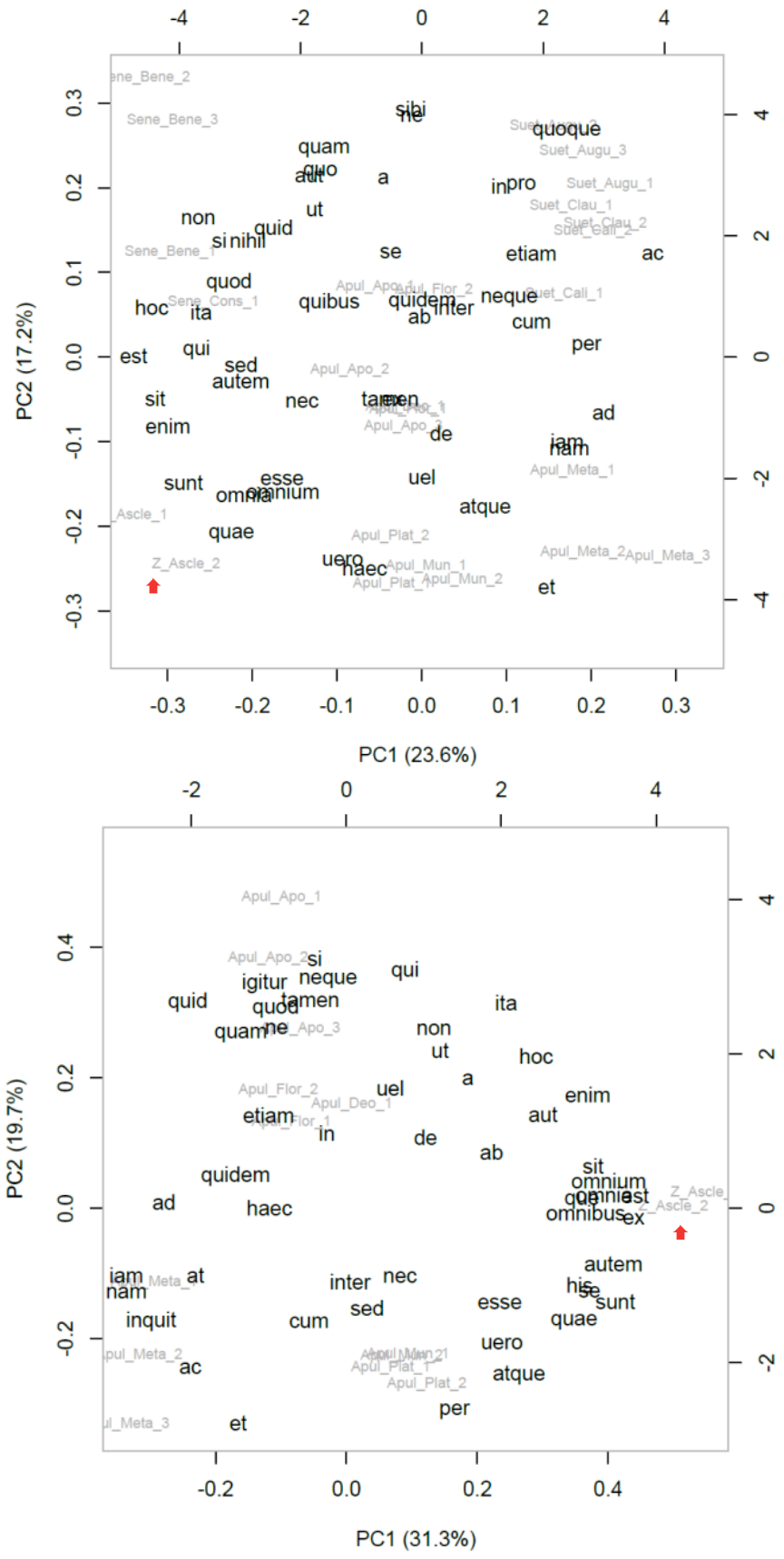
and 8 demonstrate that our methods are not able to completely distinguish the stylistic differences related to authorship and those related to genre. Contemporary stylometric scholarship suggests, nevertheless, that the authorial signal tends to be stronger than the genre-related signal; hence, the genre-induced separation between Apuleius's different sorts of works quickly vanish, if we add other authors.[40] In FIGURE 7, for instance, it is clear that the threefold division of Apuleius's oeuvre tends to collapse, whenever a third author is added to the analysis.

ASCLEPIUS, PERI HERMENEIAS, EXPOSITIO

In this section, we report experiments with two texts which are attributed to Apuleius by the manuscripts, but whose ascriptions have been debated, the *Asclepius* and the *Peri hermeneias*, and also examine the authorship of another text transmitted with Apuleius. The length of these texts (and particularly of the second) naturally warrants caution from a statistical point of view: for shorter texts, quantitative techniques are necessarily less reliable than for longer texts – 'there's no data like more data', as the maxim goes. Nevertheless, we hope to demonstrate that a number of stylometric experiments provide powerful indications as to the authenticity of these materials.

First, the *Asclepius*. We have assigned the text to unknown author Z, and conducted a PCA with the accepted corpus, with Seneca and Suetonius for initial comparison of 50 MFW (FIGURE 9). Even this preliminary analysis is highly suggestive: the *Asclepius*' samples cluster together at the margins, relatively far from the other samples. So next we exclude the

[40] Kestemont (n. 27) provides relevant references to studies of function words in relation to genre in the field of computational linguistics. A short, yet highly relevant contribution on this topic is: C. Schöch, 'Validating and interpreting Principal Component Analysis: A Case-Study from the Analysis of French Enlightenment Plays', in *Digital Humanities 2014: Conference Abstracts* (Lausanne, 2014), 136-7.



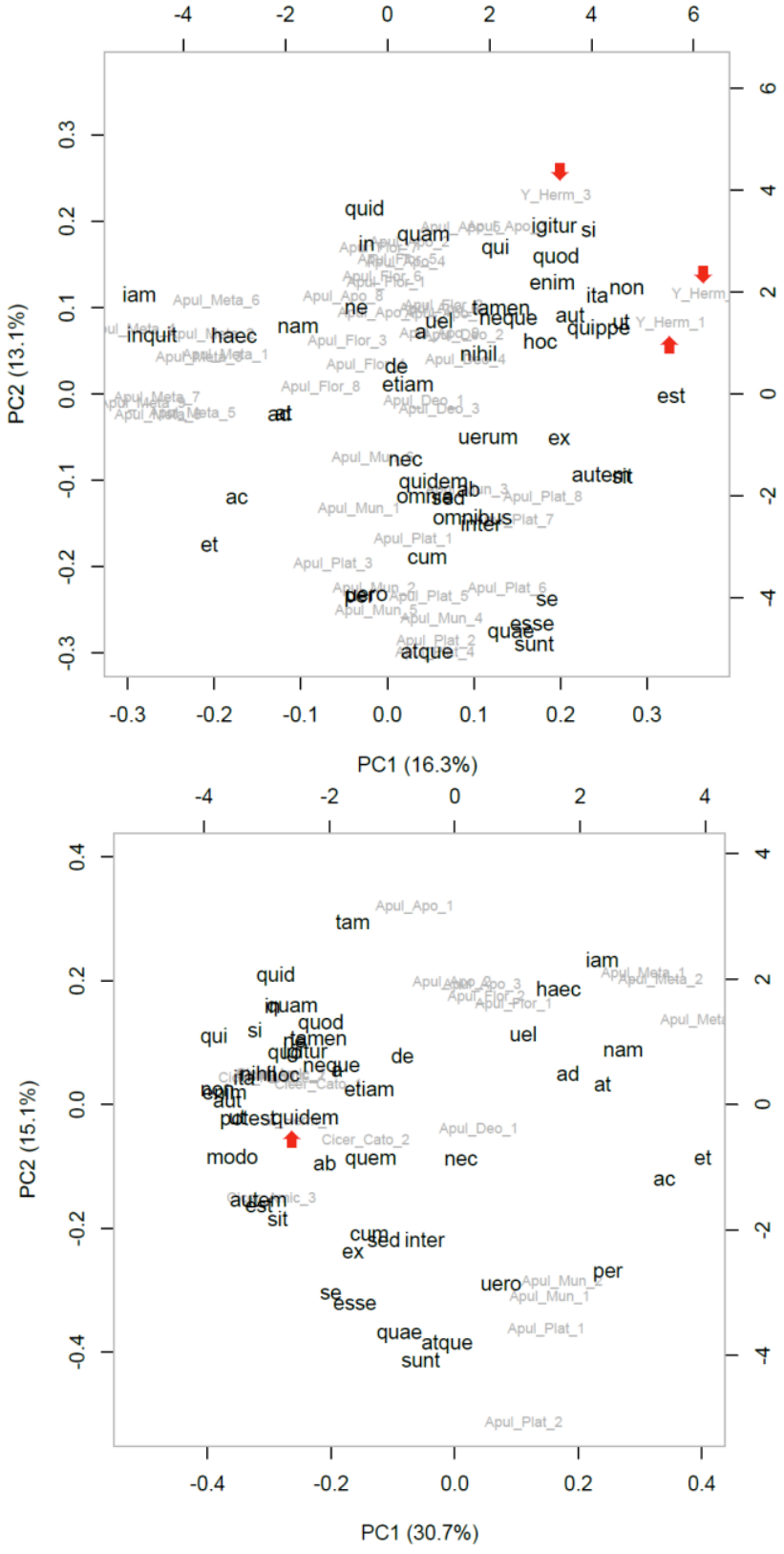
FIGURES 9 and 10

comparison authors, and just examine the *Asclepius* with the accepted corpus, once again at 50 MFW (FIGURE 10). Here we find considerable diversity across the corpus, but general consistency. The *Asclepius* samples again cluster at the extreme right margin. The horizontal dimension and left-right oppositions in PCA scatterplots are relatively more important than the vertical dimension. (This can be gauged from the percentages listed near the axis labels, which indicate how much of the original variation is captured by a particular dimension.) Therefore, it is important that the *Asclepius* is generally separated from the accepted corpus on the horizontal axis, as is for instance the case in FIGURE 10. To ensure that these results are not only describing differences in content, we have run the similar experiments with different settings. The results were unchanged and the *Asclepius* never mingled with the accepted works. The trends in our experiments corroborate the results obtained in traditional scholarship which argue against the authenticity of the *Asclepius*.[41]

Testing the *Peri hermeneias* gives us similar results. Comparing it to the Apuleian corpus (adopting a small sample size of 1,000 words), we find that it is less internally coherent (the samples tend to be widely separated) but still gravitates toward the margins of the PCA plot (FIGURE 11). When we increase the sample size to 3,000 words again and add a third author for comparison, here Cicero, the result is striking (FIGURE 12). Unlike the *Asclepius* which tended to cause the samples of Apuleius and the background authors to cluster together more closely, the *Peri hermeneias* is directly intermingled with the works of Cicero, with a fairly clear vertical zone separating it from Apuleius' works.

Notwithstanding its brevity, the relative ease with which the *Peri hermeneias* can be lured away from Apuleius' texts suggests that it is not by Apuleius, but perhaps also that it is stylistically closer to Cicero and Apuleius (that is to say, to classical prose) than the *Asclepius*

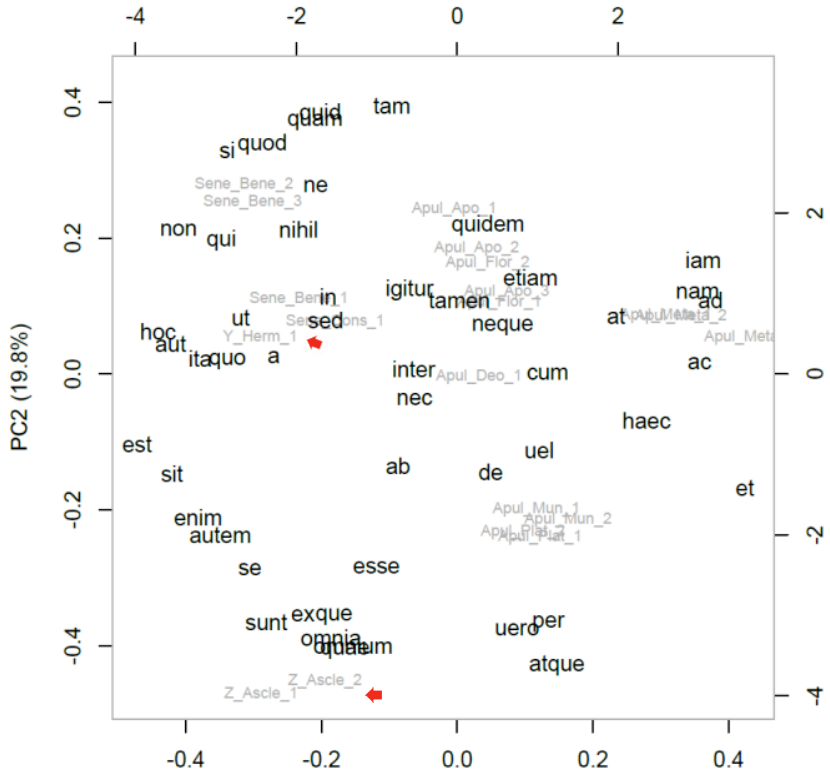
[41] See Horsfall Scotti (n. 10).

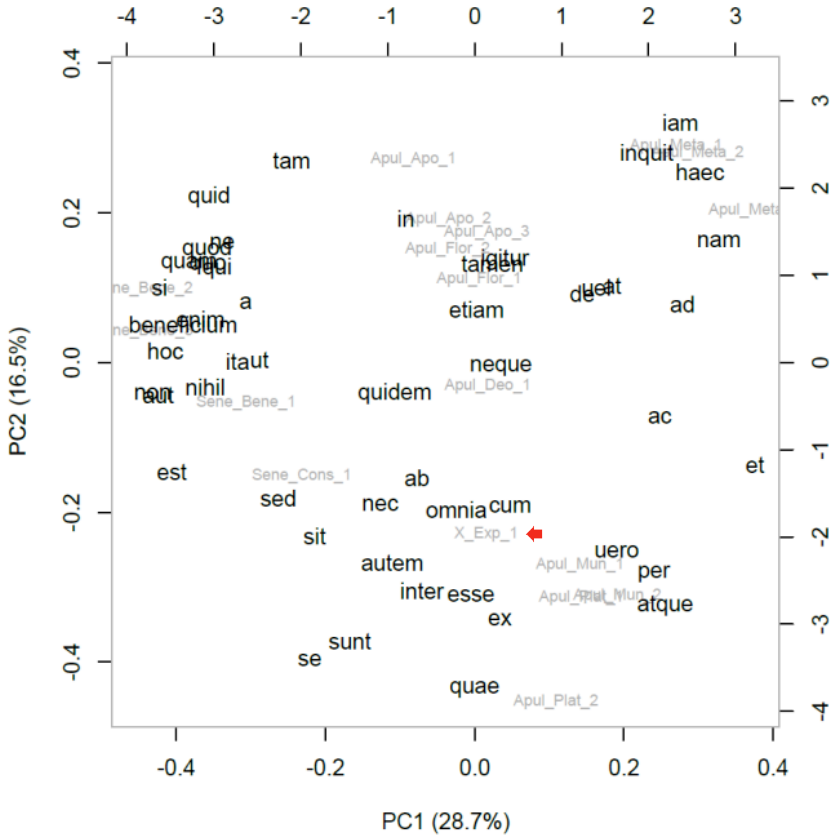


FIGURES 11 and 12 (in this figure, the one sample from the *Peri herm.* is hiding behind *quidem*)

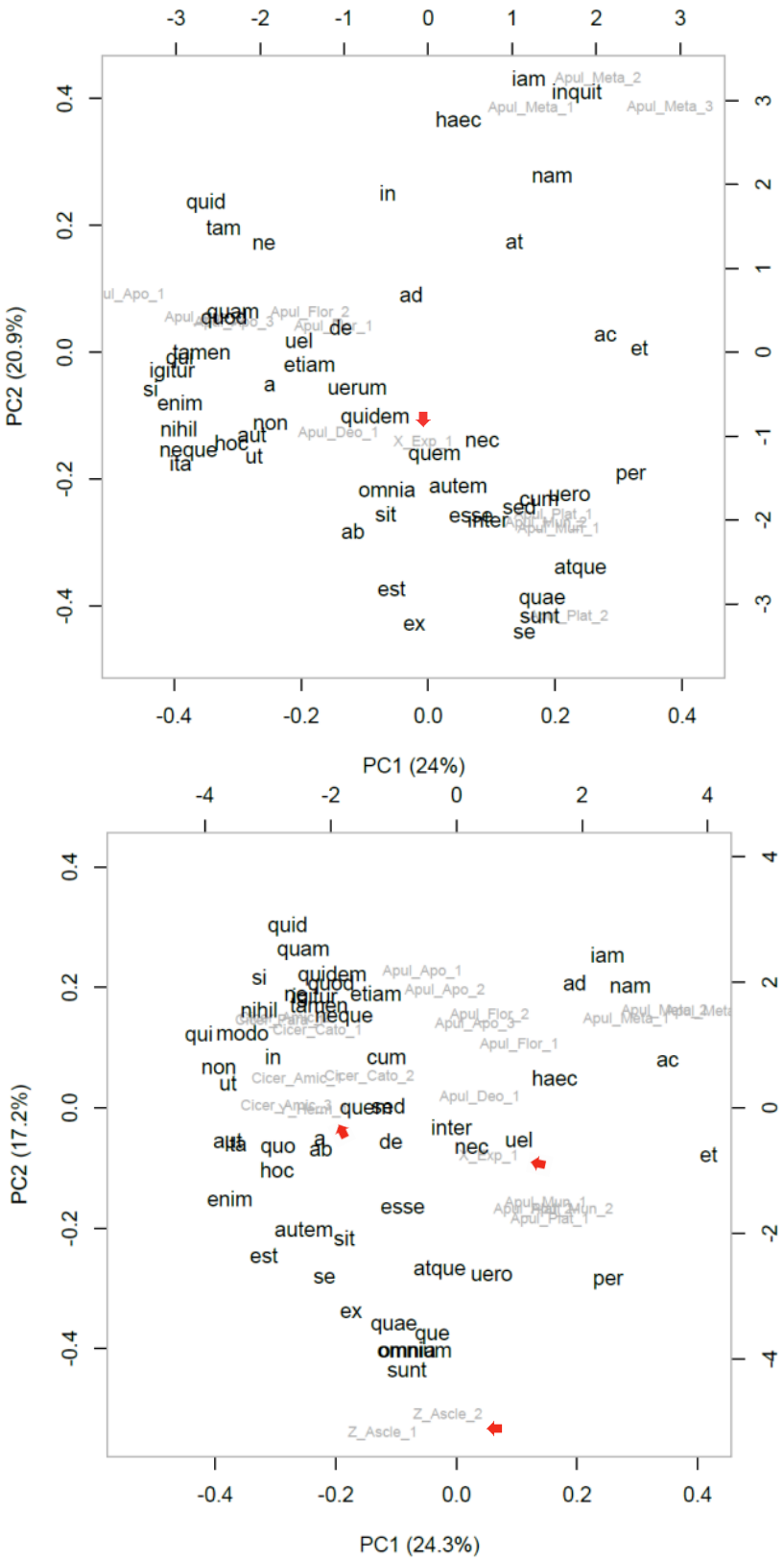
is. For these two works, our cautious conclusion must therefore be that they represent the works of two different authors, neither of whom is Apuleius. Conducting a PC analysis of the two texts with Apuleius and Seneca (FIGURE 13) reinforces this conclusion. This plot also shows once again the *Peri hermeneias* is closer to our classical, or pre-second-century, comparands than the *Asclepius* is.

The *Summarium librorum Platonis*, or to use its more correct title, the *Expositio compendiosa de Platonis pluribus libris* is a summary of fourteen Platonic dialogues found in a single manuscript of the *philosophica* after the end of the *De mundo*. Despite the fact that it has been known for sixty years, it has not yet been published. We have used the text from the forthcoming *editio princeps*. There are many compelling reasons to posit that this text was in fact composed by Apuleius and originally was an integral part of the *De Platone*; they are laid out in detail in the introduction to the forthcoming edition. Here we are only examining the evidence that can be garnered from computational analysis. First we conducted a PC analysis at 50 MFW with Seneca (FIGURE 14): the result is that the *Expositio* closely clusters with the *De Platone* and *De mundo* and there is a somewhat clear vertical axis that separates Apuleius and the *Expositio* from Seneca. Removing Seneca (FIGURE 15), we find once that the works of Apuleius tend to form three distinct clusters – the *philosophica*, the *rhetorica*, and the *Met.* – with the *De deo Socratis* appropriately located between the first two. The *Expositio*, far from gravitating out of the main group, as the *Asclepius* and the *Peri hermeneias* did, remains closely linked with the *De Platone* and *De mundo*. How much closer the *Expositio* is stylistically to the authentic corpus than the two disputed works are can be seen by comparing a PCA plot of all three together with the accepted works and Cicero (FIGURE 16). The *Asclepius* gravitates toward the lower gutter, far from all the other works; the *Peri hermeneias* nestles with Cicero. There is a porous but still relatively clear vertical axis that separates Cicero and the two disputed works from Apuleius; the *Expositio* remains





FIGURES 13 and 14



FIGURES 15 and 16

on the right side, still anchored to the *De Platone* and *De mundo*. Based on this consistent collocation of the *Expositio* with Apuleius in general and the *philosophica* in particular, it should come as no surprise that a BCT with a wide range of imposters still shows the close relationship it shares with the Apuleian corpus (FIGURE 17).[42]

[42] Here we used the same sampling settings (3,000 word samples) as in FIGURE 3. In a separate, much more technical study we have worked together with Yaron Winter and Moshe Koppel (Bar-Ilan University, Ramat-Gan) on the specific topic of the authorship of the *Expositio*: see J. Stover, Y. Winter, M. Koppel, and M. Kestemont, 'Computational Authorship Verification Method Attributes a New Work to a Major 2nd Century African Author, *Journal of the American Society for Information Science and Technology* (forthcoming), DOI: 10.1002/asi.23460. The method described in this paper applies an iterative procedure to verify the authorship of texts: in each iteration, a random set of features is selected to make the algorithm less sensitive to topic-related vocabulary. Texts are only attributed to the same author, if they prove more similar to each other than to a set of similar texts by impostor authors. The results of this approach too demonstrated that the *Expositio* was in all likelihood written by Apuleius.

FIGURE 17

WORD FREQUENCIES

Digging into the data provided by the PCA charts, and particularly their loadings, shows this method's affinities with traditional philological analysis. One can do this two ways: by looking at the words which stand either with or in opposition to the text samples of a given author, and then analysing the absolute frequency of those words in a given text or author; or by examining oppositional pairs, that is words with roughly the same meaning that stand at opposite sides of the chart, such as *nam* and *enim*.

One effective way of examining the evidence from individual words is the use of Craig's Zeta, a comparison method which constructs a double list of words preferred and

avoided when two sets of texts are compared.[43] Instead of looking at relative frequencies, this measure will extract equal-sized slices of words from two corpora, and compare in how many of these slices a particular word is present or not. Using the *Peri hermeneias* as our primary set and the authentic works as our secondary set, the significant words not attested in the former include the following (with the number of times they are found in the authentic works, excluding obviously the *Expositio*):

denique	141
alioquin	36
quin	71
inde	43
prorsus	98
tunc	158
profecto	38
rursus	30
longe	67
usque	34

Anyone who still wants to defend the authenticity of the *Peri hermeneias* needs to consider and explain how these ten words which occur together over seven hundred times in the authentic works, or about twice in every three hundred words, do not occur at all in a work of almost four thousand in length, where we would expect around twenty-seven occurrences. By contrast, there are thirty in the slightly longer *De deo Socratis*, and thirty-four in the first book of the *Metamorphoses*.

The list for the *Asclepius* is equally illuminating:

alioquin	36
quin	71
inde	43
prorsus	98
profecto	38
rursus	30
longe	67
igitur	150

[43] This measure has been proposed in H. Craig and A. Kinney, *Shakespeare, Computers, and the Mystery of Authorship* (Cambridge, 2009).

These nine words occur more than five hundred times together, about once every two hundred words in the corpus as a whole; and yet there is not a single instance in the almost nine thousand words of the *Asclepius*, where we would expect an average of about forty five in an authentic work of this length. (Compare 20 instances in the two books of the *De Platone*, of almost exactly the same length; 50 in the first two books of the *Metamorphoses*; and 54 instances in the shorter *Florida*). Of the items in the lists above, only *inde* and *prorsus* are not in the *Expositio*. To them one could add *ergo*, which is found seventy two times in the authentic works. The other words not found in *Expositio* are too laden with generic and topical significance to be relevant to this inquiry.

Overall, this individual word approach is fairly robust and insensitive to genre and content. Nonetheless, it is still open to the charge of selective bias: it is possible (if not likely) that Apuleius simply never felt it was quite the right place to use *alioquin*, *inde*, *prorsus*, *profecto*, or *longe* in the *Peri hermeneias* or *Asclepius*. The other approach, examining oppositional pairs, complements the potential weaknesses in the individual word approach. Given two words which are generally interchangeable, how often does a given author prefer one over the other? These pairs are in opposition when they occur on opposite sides of the PCA plots above.

A good example here is *nam* and *enim*, roughly equivalent in meaning, and consistently in opposition on the charts above.[44] Here, instead of dealing in absolute

[44] Studies of Latin particles have developed significantly in recent years, especially since C. Kroon's *Discourse particles in Latin: a study of nam, enim, autem, vero, and at* (Amsterdam, 1995); see also her updated discussion in 'Latin Particles and the Grammar of Discourse,' in J. Clackson (ed.), *A Companion to the Latin Language* (Malden, MA, 2011), 176-96. With considerable detail, Kroon lays out how *nam* and *enim* differ, concluding in the latter study

frequency, we can look at the relative frequency of one to the other. In the accepted corpus, of roughly a hundred thousand words, the frequency of *nam* to *enim* is 9:10 (0.90), although in the philosophical corpus this ratio drops to 3:4 (0.76). By contrast, the ratio in Pliny the Elder is much lower at 2:5 (0.42), in Cicero is even lower at almost 1:3 (0.37), and in Tertullian lower still at 3:10 (0.30); in Tacitus, by contrast, the ratio is considerably higher at 8:5 (1.68), and astonishingly higher in Sallust at almost 12:1 (11.7). Now in the *Expositio* and the *Peri hermeneias*, the ratio is not significantly different than that in the *philosophica* at roughly 6:10 (0.58 and 0.63, respectively). In the *Asclepius*, however, the ratio is a miniscule 6:100 (0.06), lower, in fact, than any other author we tested. In absolute terms, *nam* occurs only six times, while *enim* has 103 attestations; in the *De Platone*, by contrast, a work of roughly the same length, there are 19 instances of *nam* and 25 of *enim*. Hence, we can conclude two things: that the author of the *Asclepius* uses *nam/enim* far more frequently than Apuleius, he also has a massive preference for *enim*. The *Expositio* has seven instances of

---

that: '*enim* is not, or not primarily, a connective-particle that is more or less synonymous with *nam*, but a rather a conversation-management particle which seeks to establish a bond between speaker and hearer' (192). Without denying the validity of Kroon's arguments, which are many and persuasive, there is still a sense in which the interchangeability of *nam* and *enim* can be maintained. Take two roughly contemporaneous historians, Livy and Velleius Paterculus; the former uses *nam* to *enim* at a rate of about 7:10 (0.695), the latter at 14:10 (1.375), and if we push back to the previous generation, we find Sallust at a rate of 12:1 (11.7). Whatever one might say about individual cases, it cannot simply be true that semantics demanded Velleius use *nam* twice as often as Livy, or Sallust seventeen times more often. Rather, it is the unique emotional and rhetorical tenor of each word – the features which Kroon has identified – that makes an author favour one over another.

nam and 12 of *enim*; from this we can conclude that both the absolute frequency of the pair and their relative frequency to one another is roughly consistent with Apuleian usage.

Another pair worth examining is *igitur/ergo*. In the accepted corpus, we find a frequency of roughly 2:1 (2.08), and looking at just the *De Platone*, we find 8 instances of *igitur* and none of *ergo*. We can posit that Apuleius had no strong preference for using this pair at all, but when he did he more often preferred *igitur*. Just for the sake of comparison, we find ratios of 3:1 in Cicero (3.35) and 4:1 in Gellius (4.26), while, on the other side, we find a frequency of 1:5 (0.19) in Pliny the Younger, and 1:50 (0.02) in Seneca. For the disputed works, we find nothing inconsistent in the *Expositio*, which has one attestation of *igitur* and none of *ergo*. But in the *Peri hermeneias* there is a very high absolute frequency of the pair (unsurprising, given the genre) and a large preference for *igitur* over *ergo*, to the tune of 10:1 (9.75). In the *Asclepius*, *igitur* is not attested at all, while *ergo* is found 55 times. Anyone who wants to defend the authenticity of the *Asclepius* would have to account for this glaring discrepancy with Apuleius' usual style.

None of these little case studies is meant to offer definitive proof one way or another. Rather they illustrate the kind of data stylometric techniques analyse when they produce the results they do. And while we have only looked at the data here for a small number of cases, the program analysed fifty or more of the MFW. Hence, the examination of the Apuleian corpus we have conducted is similar in method to that conducted by Redfors in 1960, which examined many of the same terms, such as *nam*, *enim*, *igitur*, and *ergo*.[45] But extending his manual calculations with computational methods vastly increases the quantity of data available for analysis. Thereby we can go beyond his judgement that the question of the *De*

[45]Redfors (n. 9), 39-46.

Platone and *De mundo* represents an *unlösbares Echtheitsproblem.*[46] Computational methods allow us to make a definitive claim in favour of their authenticity; indeed, they allow us to recharacterize his question as a *lösbares* (and even *gelöstes*) *Echtheitsproblem*.

CONCLUSIONS AND PERSPECTIVES

The conclusions we have reached indicate first that the authorship of the *De Platone* and the *De mundo* should no longer be questioned, as indeed a majority of scholars have for some time held. The only remaining objections to the attribution of these handbooks to Apuleius are based particularly on rhythm, both on the presence of *cursus mixtus*, or a combination of metrical and rhythmic period terminations, and on the frequency of false quantities. But all the scholars who have studied these questions – Redfors, Oberhelman and Hall, and Holmes – have cautioned that we simply have too little surviving literature from the end of the second century to have suitable comparanda. Since the works are in fact authentic, their attribution to Apuleius should stand as the starting point for examination into the history of later Latin prose rhythm. It is undoubtedly interesting that we can fix the beginnings of *cursus mixtus* in North Africa in the second half of the second century, and it is even more interesting that at its beginning it was used in alternation with the more formal system of *clausulae*. Hence, the development of *cursus* should not be thought of only as an unconscious evolution, but at its beginning a deliberate stylistic choice, dependent on register, genre and audience. Further, perhaps 'false quantities' are not simply 'false', but are rather deployed or not deployed according once again to register, genre and audience.

The second conclusion we can draw is that the *Asclepius* is probably not by Apuleius – despite certain superficial similarities with Apuleius' style, in a comparison with other texts

[46] Redfors (n. 9), 115-17.

it leaps out too often to warrant an attribution. In other words, in word choice, and particularly in the use of function words, Apuleius, Cicero, Seneca, Tertullian, and a number of other authors, are closer to each other than any of them are to the *Asclepius*. This fact corroborates the hypothesis of a fourth-century date for the translation.

The third conclusion we can provisionally draw is that there is no good reason to accept the attribution of the *Peri hermeneias* to Apuleius, despite the fact that it generally seems closer in style to Apuleius than the *Asclepius*. One must emphasize that its small size makes definitive judgement on statistical grounds alone impossible. Nonetheless, we have identified a number of potential problems with the attribution which admit no easy riposte. Given the serious objections that have been raised to Apuleian authorship, the results we have reached, though not conclusive, are still damning. Computational methods have given us no particular reason to support the attribution, and any number of reasons to doubt it, a fact which makes the burden of proof on those who would accept the attribution even heavier. More positively, however, the *Peri hermeneias* consistently remained more closely collocated with the classical texts under analysis than the *Asclepius*; more research is obviously needed, but as a preliminary hypothesis, our findings suggest that dating it to a generation after Apuleius would not be inappropriate.

The fourth reasonably solid conclusion we can draw is that there is no reason to dismiss the authenticity of the *Expositio*, that it is consistently more closely linked to Apuleius in general and the *philosophica* in particular than either the *Asclepius* and the *Peri hermeneias* or any of our background texts, and that there is a high probability that it is by the same author on statistical grounds. This is not to say that there are not stylistic differences – they are many and obvious – but on the aggregate these differences are not sufficient to disprove Apuleian authorship nor to overwhelm the stylistic similarities. We must stress that computational methods alone will not give us any definitive answers; rather, they work to

corroborate or undermine other types of evidence. Elsewhere the whole case for Apuleian authorship has been laid out: let it suffice here to mention that there are undeniable intertextual links between it and the *De Platone*, that it is certainly a product of the second century, and that it contains some of the conspicuous and obvious features of Apuleius' style. Here we add that stylometry offers no case for rejecting the attribution, and several good reasons for accepting it.

In addition to these specific conclusions regarding the authenticity of individual works, our data also suggest a model for classifying Apuleius's different works. In the plots above, the corpus is usually split into three relatively distinct groups: the *Metamorphoses*, the *rhetorica*, and the *philosophica*. The *De deo Socratis*, being a rhetorical performance piece on a philosophical topic, tends to fall between the latter two. This dovetails with the traditional classification of Apuleius's works, and once again demonstrates the sensitivity of our computational methods. Examining just Apuleius's works (with the *Expositio* included), as presented in FIGURE 15 above, we can get some sense in technical terms of how this division comes about: *nam* is strongly preferred in the *Met.* and *enim* avoided, while the *rhetorica* in particular, but also the *philosophica*, use both, without so strong a preference for *nam*. In adversative particles, a threefold division can be seen: the *Met.* prefers the strong (and sometimes narratological) *at*,[47] while the *rhetorica* goes for the strong *tamen*, while the *philosophica* tend toward *sed*, *autem*, *verum*, and *vero*.

We have only scratched the surface of the possible applications of modern stylometry to classical texts. We have cleared up some of the problems plaguing the works attributed to Apuleius. We have also demonstrated how stylometric experiments for authorship attribution in Classical Latin texts can be designed and executed. Finally, we have shown some of the additional benefits that can be reaped from the application of these kinds of analyses for the

[47] On *at*, see Kroon (2011, n. 44).

study of style and genre. Computational experiments designed to test generic questions may well in the future be able to give even more precise results as to the lexical choices that are one constituent of genre. Methods similar to the ones we use here may also be effective in sorting through the difficulties of other vexed corpora, such as the *corpus Caesarianum* or the *Major Declamations* of ps-Quintilian. At the least, we hope that the use of computational methods in conjunction with traditional philology in exploring classical literature will continue to expand and shed light on problems and questions both new and old in the field of classics more broadly.

All Souls College, Oxford
JUSTIN STOVER
justin.stover@classics.ox.ac.uk

University of Antwerp / Research Foundation of Flanders
MIKE KESTEMONT
mike.kestemont@uantwerp.be