Computational authorship verification method attributes a new work to a major 2nd century African author

Reference:

Stover Justin Anthony, Winter Yaron, Koppel Moshe, Kestemont Mike.- Computational authorship verification method attributes a new work to a major 2nd century African author
Journal of the Association for Information Science and Technology / Association for Information Science and Technology - ISSN 2330-1635 - 67:1(2016), p. 239-242
Full text (Publishers DOI): http://dx.doi.org/doi:10.1002/asi.23460

Computational authorship verification method attributes new work to major 2nd century African author

Reference:

Stover J., Winter Y., Koppel M., Kestemont Mike.- *Computational authorship verification method attributes new work to major 2nd century African author*

Journal of the Association for Information Science and Technology / Association for Information Science and Technology - ISSN 2330-1643 - (2015)

Handle: http://hdl.handle.net/10067/1220690151162165141

Title:

Computational Authorship Verification Method Attributes New Work to Major 2nd Century African Author

Authors and addresses:

J. A. Stover[a], Y. Winter[b], M. Koppel[b], M. Kestemont[c]

[a]All Souls College, University of Oxford, High Street, Oxford, OX1 4AL, United Kingdom; [b]Department of Computer Science, Bar-Ilan University Ramat-Gan, 52900 Ramat-Gan, Israel; [c]CLiPS Computational Linguistics Group & Institute for the Study of Literature in the Low Countries, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium

Send correspondence to: Mike Kestemont, University of Antwerp, City Campus, Prinsstraat 13, room D.118, 2000 Antwerp, Belgium, Tel: 0032477918668, Email: mike.kestemont@uantwerp.be

Key index words:

Authorship Verification, Computational Linguistics, Classical Antiquity, Apuleius, Stylometry, Forensic Linguistics

Abstract:

We discuss a real-world application of a recently proposed Machine Learning method for authorship verification. Authorship verification is considered an extremely difficult task in computational text classification, because it does not assume that the correct author of an anonymous text is included in the candidate authors available. To determine whether two documents have been written by the same author, the verification method discussed uses repeated feature subsampling and a pool of impostor authors. We use this technique to attribute a newly discovered Latin text from Antiquity (the *Compendiosa expositio*) to Apuleius. This North African writer was one of the most important authors of the Roman Empire in the 2nd century and authored one of the world's first novels. This attribution has profound and wide-reaching cultural value, because it has been over a century since a new text by a major author from Antiquity was discovered. This research therefore illustrates the rapidly growing potential of computational methods for studying the global textual heritage.

Introduction

In computer science, authorship studies are nowadays a prominent field of study, in which algorithms are optimized to automatically identify the authors of texts using linguistic features related to writing style (Juola, 2008; Stamatatos, 2009; Koppel, Schler, & Argamon, 2008; Van Halteren, Baayen, Tweedie, Haverkort, & Neijt, 2005). In the field a distinction is commonly made between the attribution and verification set-up (Koppel, Schler, & Bonchek-Dokow, 2007; Kestemont, Luyckx, Daelemans, & Crombez, 2012). In the attribution scenario, resembling a police line-up, an algorithm selects the author of an anonymous text from a list of potential candidate authors, which is known beforehand to include the target author. The verification setup is more demanding: given any two texts, the task is to determine whether or not they should be attributed to the same author. Authorship verification therefore is a much more complicated task than authorship attribution, because this set-up does not presuppose that the true author of the text is actually included in the set of candidate authors available.

Koppel and Winter (2014) have recently introduced an innovative method for authorship verification which substantially outperformed more naïve approaches on a large bechmark data set. Here, we apply this technique to a newly discovered text from Roman Antiquity and attribute it to the North-African writer Apuleius (2nd century AD). The textual heritage of Greek and Roman antiquity has decisively shaped the development of global, intellectual culture. This heritage has been transmitted to us through copies handwritten during the Middle Ages. From the Renaissance through the nineteenth century, there has been a concerted international effort to recover, identify, and publish the works contained in these manuscripts. For a

long time now that project has been considered complete. It has been over a century since the last discovery of a new text by a major classical Latin author and specialists have long believed that there were no more texts to be uncovered (Dolbeau, 1998-1999; Reeve, 2000).

A New Text from Antiquity

Here, we analyze a newly discovered text, surviving from a single medieval manuscript in the Vatican library in Rome (Reg. lat. 1572). This manuscript has long been ignored by scholars and contains a previously unstudied Latin text discussing Plato's works, entitled the *Compendiosa expositio* ("Brief Presentation"). Preliminary philological analysis yielded solid indications that this *Expositio* in all likelihood is a newly found text from Antiquity, that somehow undetected slipped through four centuries of scholars looking for ancient texts. Traditional stylistic and metrical analyses moreover suggested that this new text might well be a hitherto unknown work by Apuleius of Madaura. Apuleius was one of the most important authors of the Roman Empire (ca. 125-180 AD) and left behind an astonishingly varied literary corpus, including one of the world's first novels (Harrison, 2000). Especially interesting in this respect, is Apuleius's work *De Platone* ("On Plato"). Codicological evidence from other manuscript copies of Apuleius's works indicates that *De Platone* might originally have contained a third book. Our traditional philological research suggests that the newly found *Expositio* could well have been *De Platone*'s thought lost third book.

Below, we report on a computational analysis in which we attempted to verify this attribution. For this study, we have collected a representative development corpus of

22 works by authors such as Pliny the Younger and Suetonius, who have stylistic, chronological, generic, or thematic similarity with Apuleius. Before moving to the verification method, we briefly inspect the *Expositio*'s position in the development corpus. We visualize the texts' lexical term-frequency vectors via a Bootstrap Consensus Tree (Fig. 1), a well-known technique from phylogenetics (Paradis, 2012; Paradis, Claude, & Strimmer, 2004), which has also been ported to computational stylistics (Eder, 2013; Eder, Kestemont, & Rybicki, 2013) using Burrows's Delta as distance metric (Burrows, 2002; Argamon, 2007). (More details on our corpus, methods and preprocessing are given in the online appendix.)

Apuleius's oeuvre appears to be relatively heterogenous with sub-branches for his philosophical, rhetorical and fictional works. This visualization nevertheless confirms the similarities between the newly discovered *Expositio* and Apuleius's works (in red). Fig. 1 is not free of issues however. Cicero's *Timaeus* clusters with the *Expositio*'s clade, probably on the basis of content rather than style: the *Timaeus*, just like the *Expositio*, is an adaptation of Platonic materials. Seneca's *De Constantia* moreover mingles with Cicero's texts. (The same basic pattern emerges across a variety of experimental settings for the BCT.) The authorship verification method discussed below allows us to verify whether Apuleius is indeed a likely authorial candidate for the *Expositio*. We emphasize that using verification instead of attribution is vital with respect to the *Expositio,* since we cannot presuppose that the true author of the text is actually included in the set of candidate authors available.

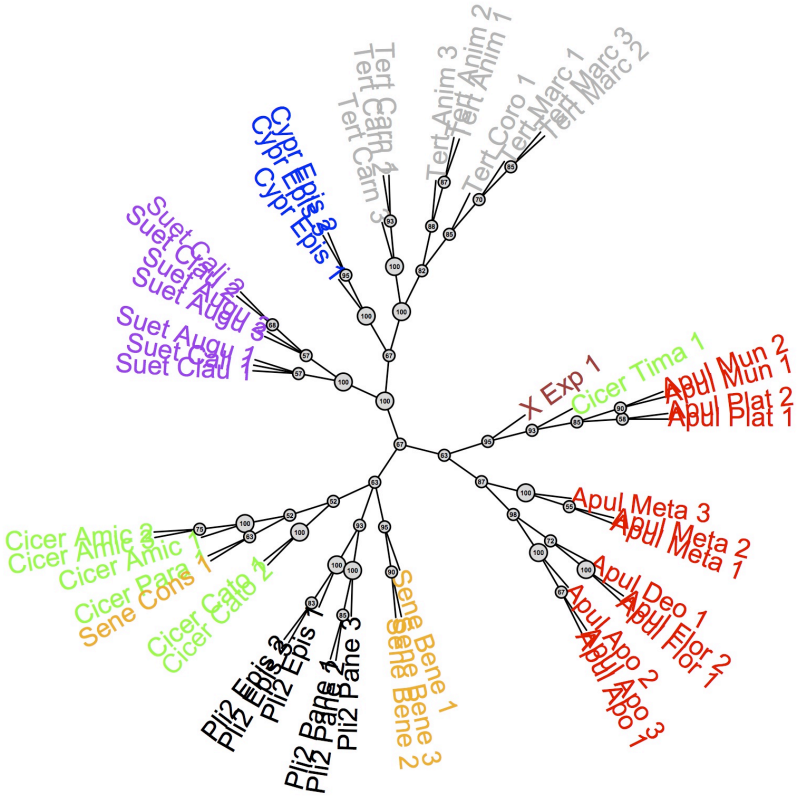

Fig. 1: Visualisation of texts in the development corpus (4,500 word slices). Bootstrap Consensus Tree (majority vote) for 50-word frequency bands up to 3,000 words.

Verifying the Apuleian Authorship of the *Expositio*

We limit ourselves to a concise discussion of the verification method and we refer to the original paper for more details on the method's performance on bench mark data sets (Koppel & Winter, 2014). For the verification method we proceed as follows. Given two texts, *X* and *Y*, as well as a pool of similar background texts, we use different trials to assess whether *X* and *Y* are written by the same author. In each trial,

we select as a feature set at random a subset of 125,000 word unigrams and bigrams from among a total pool of approximately 250,000 such features.

By this repeated subsampling, we can sample from both higher-frequency and lower-frequency bands, while keeping the method less sensitive to content-specific features, because each time different feature combinations are selected. We can now determine whether *X* is more similar to *Y* than it is to any of the background texts, according to the min-max distance measure. We run multiple trials, each time using a different randomly generated set of features. If the proportion of trials in which *X* is more similar to *Y* than to any of the background texts is greater than some threshold σ^*, we conclude that the author of *Y* is identical to the author of *X*.

We apply the verification method to each pair of texts in the development corpus. For any given pair of texts, *X* and *Y*, we use 50 texts (not by the same author(s) as *X* or *Y*) randomly chosen from among a pool of 180 background texts by 36 authors writing in similar genres and/or periods as the texts in our corpus. For each pair, we run 100 trials. The pool of background texts thus serves as a set of impostors in each trial, which is a well-established practice in e.g. speaker identification and which has recently been introduced in authorship studies (Koppel & Winter, 2014).

We assign a text pair to the class *same-author* if its score (i.e., the proportion of trials for which the pair are more similar to each other than to any text in the background set) exceeds a threshold σ^*. We vary the σ^* threshold to obtain a recall-precision curve for the class *same-author* as shown in Fig. 2. Of 199 *different-author* pairs, only three obtain a score above .20 and only one above .50: the pair of Cyprian's *Epistles*

and Tertullian's *De corona*, which obtains a score of .71. In contrast, of 32 *same-author* pairs, 15 obtain a score above .50. These results proved stable across a large variety of other parameter settings (e.g. number of impostors) and feature types (e.g. character *n*-grams) with which we experimented on the development corpus.

With respect to Apuleius's works, each pair among the *De deo Socratis*, *Florida* and *Apologia* obtains a score of .85 and above, as do the pair *De mundo* and *De Platone*. The *Metamorphoses* obtains a score above .50 with the *Florida*, but with no other text by Apuleius. (Note that the verification method does not guarantee transitivity.) No pair involving a text by Apuleius and a text by another author obtains a score above .35. Even different-author pairs such as Cicero's *Timaeus* and Apuleius's *De Platone*, which are very similar in genre and subject matter, never surpass the .35 threshold. The verification method is therefore characterized by a high precision and a relatively lower recall for same-author pairs. This shows that the method is a "non-greedy" attributor, which is typically a desirable quality in real-world forensic application (e.g. to avoid false accusations in a legal context). If a new text is paired with a work by Apuleius and obtains a σ^* score above .20, it would be extremely likely that Apuleius authored the new text.

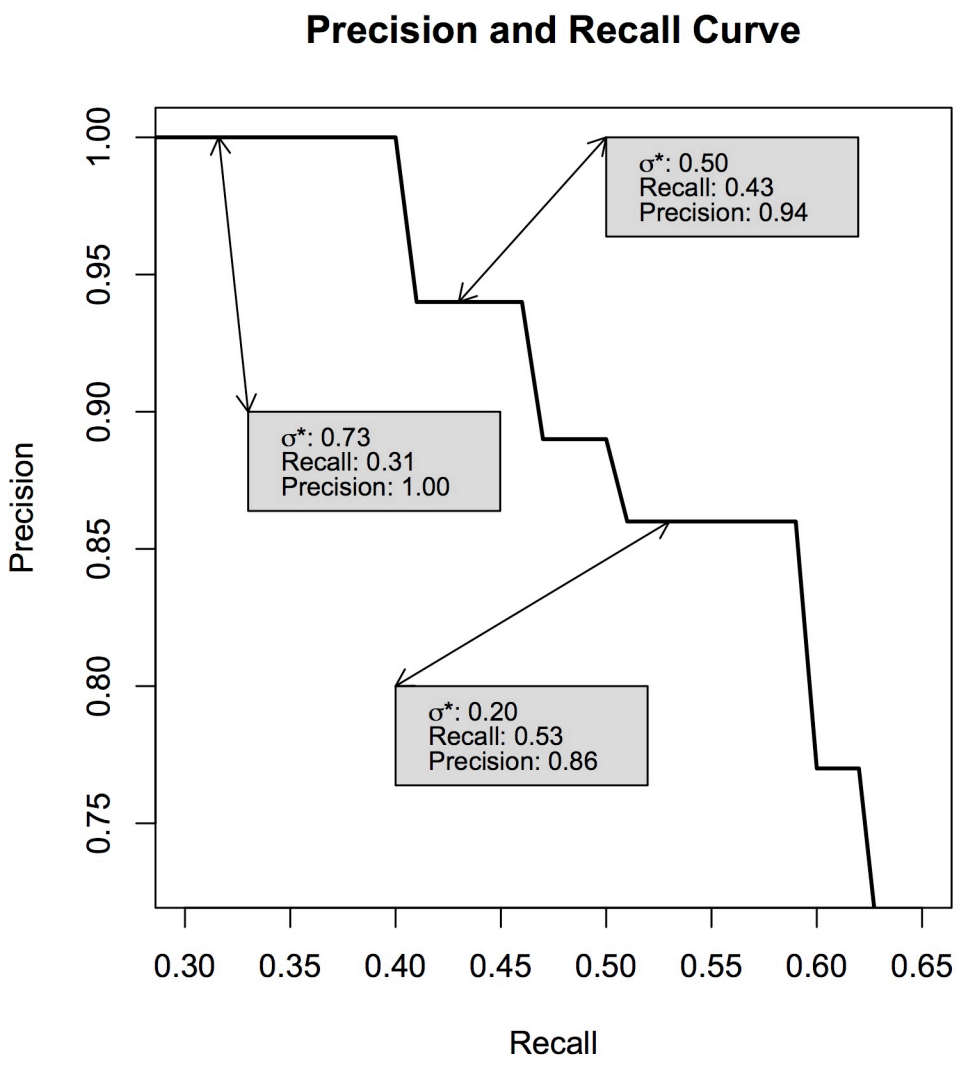

Fig. 2: Precision-Recall curve. The effect of various thresholds σ* for the verification score in terms of precision and recall for the same-author category in the development corpus. For a threshold of σ*=.50, our method reaches a high precision (.94), meaning that only a single text pair is incorrectly classified as same-author.

We now check the newly discovered *Expositio* against the 22 texts in our development corpus. We find that no such pair yields a score above .04 with the single exception of the pair *Expositio* and *De Platone*, which obtains an exceptionally high score of .73. Of the 16 pairs in our development set that obtained a score of .50

or above, 15 (93.8%) are same-author pairs. No different-author pair obtains a score of .73 or more. (Again, the same pattern emerges with different experimental settings and feature types.) These results lend particularly strong support to the hypothesis that the *Expositio* has been written by the same author as *De Platone*, that is Apuleius. This solid result is especially valuable, because of our initial hypothesis that the *Expositio* might be the forgotten third book of *De Platone*.

It deserves emphasis, however, that the *Expositio* does not yield high scores when paired with other texts by Apuleius than *De Platone*, most of which are in different genres than the *Expositio* (Platonic philosophy). This is entirely in line with previous studies which have demonstrated how difficult it is to apply authorship attribution across different genres (e.g. Kestemont, Luyckx, Daelemans, & Crombez, 2012). Strictly speaking, one could therefore only reconstruct Apuleius's oeuvre indirectly on the basis of our results (the *Metamorphoses* is attributed to the *Florida*, the *Florida* in turn to *Apologia*, the *Apologia* to *De deo Socratis* etc.). Nevertheless, the identification of the authors of *De Platone* and the *Expositio* is a solid result, especially because the *Expositio* is never attributed to other authors writing in the same genre (e.g. Cicero's *Timaeus*). In combination with the results from the traditional philological analysis, our quantitative results thus lend quantitative support to the thesis that the *Expositio* is a newly discovered text from Antiquity, which is very likely to have been authored by Apuleius of Madaura.

Conclusion

Classical philologists have long used lexical analyses (e.g. of stop words) for the attribution and dating of Latin texts, but usually only selectively and by hand.

Unfortunately, such analyses have previously been unable to conclusively resolve questions of authenticity in the Apuleian corpus (Harrison, 2000; Redfors, 1960). The ongoing mass digitisation of classical works, for instance in the open source Perseus project or in the Brepols Library of Latin Texts, has enabled and stimulated the large-scale computational analysis of ancient documents (Crane, 2006; Kestemont, Moens, & Deploige, 2013). Our research illustrates the rapidly growing potential of computational methods for humanities disciplines in studying our global textual heritage.

The results of the verification method corroborate our traditional philological analyses, which reveal a probable dating of the new text to the second century AD, well within the lifetime of Apuleius, and which show a close intertextual relationship between the *Expositio* and Apuleius's works. Our results are also in line with codicological evidence which suggests that the *Expositio* could have been originally the thought lost third book of Apuleius's *De Platone*. The quantitative results allow us to support the thesis that the *Expositio* is in fact by Apuleius, and probably was an integral part of the *De Platone*. This discovery has profound and wide-reaching cultural value, because of the scarcity of Latin texts surviving from the late second century and the new light the discovery sheds on the output of an author of major importance. Especially promising is the application of our methods to other unresolved cases of disputed authorship in classical literature.

References

Argamon, S. (2007). Interpreting Burrows's Delta: Geometric and probabilistic foundations. Literary and Linguistic Computing, 23(2), 131– 147.

Burrows, J. (2002). "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing, 17(3), 267-287.

Crane, G. R. (2006). What do you do with a million books? D-Lib Magazine, 12.

Dolbeau, F. (1998-1999). Découvertes récentes d'oeuvres latines inconnues (fin IIIe-début VIIIe s.). Sacris erudiri, 38, 101-142.

Eder, M. (2013). Computational stylistics and Biblical translation: how reliable can a dendrogram be? In Piotrowski, T. & Grabowski, Ł. (Eds.), The translator and the computer (pp. 155-170). Wrocław: WSF Press.

Eder, M., Kestemont, M. & Rybicki, J. (2013). Stylometry with R: a suite of tools. In Digital Humanities 2013. Conference Abstracts (pp. 487-489). (Lincoln, Nebraska: ADHO.

Harrison, S. J. (2000). Apuleius: A Latin Sophist. Oxford: Oxford University Press.

Juola, P. (2008). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3), 233–334.

Kestemont, M., Luyckx, K., Daelemans, W., & Crombez, T. (2012). Cross-Genre Authorship Verification Using Unmasking. English Studies, 93(3), 348-364.

Kestemont, M., Moens, S. & Deploige, J. (2013). Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. Literary and Linguistic Computing, doi:10.1093/llc/fqt063.

Koppel, M. & Winter, Y. (2014). Determining if two documents are written by the same author. Journal of the American Society for Information Science and Technology, 65(1), 178-187.

Koppel, M., Schler, J., & Argamon, S. (2008). Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60(1), 9–26.

Koppel, M., Schler, J. & Bonchek-Dokow, E. (2007). Measuring differentiability: unmasking pseudonymous authors. Journal of Machine Learning, 8, 1261-1276.

Paradis, E. (2012). Analysis of Phylogenetics and Evolution with R. New York: Springer Verlag.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 20(2), 289-290.

Redfors, J. (1960). Echtheitskritische Untersuchung der apuleischen Schriften De Platone und De Mundo, Lund: C.W.K. Gleerup.

Reeve, M. (2000). Cuius in usum? Recent and future editing. Journal of Roman Studies, 90, 196-206.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

Van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M. & Neijt, A. (2005). New Machine Learning Methods Demonstrate the Existence of a Human Stylome, Journal of Quantitative Linguistics, 12(1), 65-77.

Acknowledgements We wish to acknowledge Brepols Publishers for the use of proprietary material. The authors wish to thank James Hankins, Frank Willaert, Folgert Karsdorp, Piet Maas and Walter Daelemans for critical discussions and

reading of the manuscript. Mi.K. was funded for this research as a postdoctoral researcher for the Research Foundation of Flanders.

Notes Supplementary materials of this study are available from https://github.com/mikekestemont/Apuleius. This repository additionally contains more details on our methods, code, data and the exact composition of our corpus.

Supplementary Materials: "Computational Authorship Verification Method Attributes New Work to Major 2nd Century African Author"

J. A. Stover, Y. Winter, M. Koppel, M. Kestemont

Methods

Function words consensus trees

Throughout the experiments, we have adopted a pragmatic definition of words as space-free lowercase character strings, after the removal of non-alphabetic characters and the orthographic normalization of all u's to v's. In the BCT procedure, we proceed as follows. We first truncate each text to the 9,000 first words to maximise comparability and split these in two equal-sized samples of 4,500 words, which is the approximate size of the shortest text in the corpus. Texts shorter than 9,000 words are truncated to a single 4,500 word sample. We select the 3,000 words which are most frequent in the samples and compute their relative frequency in the samples. We now automatically cull all personal pronouns from this feature space in order to avoid interference from factors such as narrative perspective and genre that are often reflected in personal pronouns. For a given feature set, we represent a document as a numerical vector $<x_1, x_2, \ldots, x_n>$, where x_i is the relative frequency in X of the ith feature (for some fixed ordering of the features in our feature set). We define the distance between two documents as $\Delta(\vec{X},\vec{Y}) = \sum_{i=1}^{n} \frac{1}{\sigma_i} |x_i - y_i|$. This is equivalent to Burrows' Delta measure, which has been found to be useful for authorship analysis. We now use Ward's minimum variance method to cluster the documents. This clustering method is used in multiple iterations, each of which uses a different feature

set. The first iteration uses the frequency band of the 1-50 most frequent words (MFW), the second the 50-100 MFW, and so on, until all frequency bands have been analyzed (up to rank 3,000). Finally, the different resulting cluster trees are aggregated into a single consensus tree, by collapsing nodes that are not observed in at least 50% of the analyses (i.e. majority vote). In Fig. 1, the bootstrap values have been added as node labels (readers can zoom in digitally for more detail). Note how the node over-arching Apuleius's texts and the *Expositio* only reaches an 80% consensus level, whereas those over-arching other authors reach 97-100%.

Authorship verification via feature subsampling

For the verification experiment, we proceed as follows. We represent a document X as a vector of values $<x_1, x_2, \ldots, x_n>$ where the value x_i is the relative frequency of the ith feature in our feature set (for some fixed ordering of the word unigrams and bigrams), multiplied by its inverse document frequency. Because the method is based on text pairs, we have to limit each text to a single sample of 4,500 words. We only sample from features that appear at least three times in the corpus. We measure the similarity of two documents using the min-max measure of vector similarity: $\mathrm{minmax}(\vec{X},\vec{Y}) = \sum_{i=1}^{n} \min(x_i, y_i) / \sum_{i=1}^{n} \max(x_i, y_i)$.

Software

The BCT clustering experiments (Fig. 1) were carried out using the R code in the *Stylometry for R* package, which heavily depends on the *APE* package for this functionality. These packages are available from CRAN (http://cran.r-project.org/). The code necessary to replicate the verification experiments is available in the public

domain from GitHub: https://github.com/mikekestemont/Apuleius. This repository also holds the text materials for this study.

Corpus data

All text materials for this study are included in the GitHub repository for this paper (GitHub: https://github.com/mikekestemont/Apuleius), with the exception of three texts by Tertullian and Cyprian, which are proprietary data owned by Brepols Publishers (*Library of Latin Texts*). These texts are non-essential to the replication of our main findings. The other main sources from which we collected these texts are *The Latin Library* (http://www.thelatinlibrary.com/) and the *Patrologia Latina,* ed. J. P. Migne, 217 vol. (Paris 1841-65), which can be consulted online (e.g. www.mlat.uzh.ch). Other sources are indicated below. Please note that the background corpus was used in the experiments to construct impostors for testing pairs of texts from the development corpus; as a result, it was not essential to use the best editions (or even texts with known print provenance).

Development Corpus (author names followed by abbreviations in Fig. 1)

- Apuleius (Apul), *Apologia* (ed. Helm 1912); *Expositio compendiosa* (ed. Stover); *Florida* (ed. Hunink 2001); *Metamorphoses* (ed. Helm 1907); *De mundo* (ed. Beaujeu 1973); *De Platone* (ed. Beaujeu 1973)
- Cicero (Cicer), *De amicitia* (ed. Mueller 1890); *De senectute* (ed. Shuckburgh 1920).
- Cyprian (Cypr), *Epistulae* (ed. Diercks 1994)
- Pliny the Younger (Pli2), *Epistulae* (ed. Mynors 1963); *Panegyricus* (ed. Mynors 1964)
- Seneca the Younger (Sene), *De beneficiis*; *De constantia* (ed. Basore 1928).

- Suetonius (Suet), *Vitae* (ed. Ihm 1907)

Background Corpus

- Ambrose, *Epistolae*. Latin Library.
- Ambrose, *De mysteriis*. Latin Library.
- Anonymous, *Rhetorica ad Herennium*. Scrineum. scrineum.unipv.it.
- Arnobius, *Adversus nationes*. Latin Library.
- Asconius, *Orationum Ciceronis quinque enarratio*. Latin Library.
- Augustine, *Contra Academicos*. *De agone Christiano*. *De libero arbitrio*. *De civitate dei*. *Confessiones*. *De magistro*. *Retractiones*. *De trinitate. De vera religione*. Patrologia Latina 32-42.
- Aulus Gellius, *Noctes Atticae*. Lacus Curtius. http://penelope.uchicago.edu/Thayer/E/Roman/home.html.
- Boethius, *Consolatio philosophiae* (prose only). James O'Donnell (faculty.georgetown.edu/jod/boethius/jkok/list_t.htm.). *De divisione. De diffinitione*. *De persona et duabus naturis. De topicis differentiis. Quomodo trinitas unus deus*. Patrologia Latina 64.
- Calpurnius Flaccus, *Declamationes*. forumromanum.org.
- Cicero (M. T), *Academica*. *Pro Archia*. *Brutus. Pro Caecina*. *Pro Caelio*. *De divinatione*. *De fato*. *De finibus*. *Pro Milone*. *De natura deorum*. *De officiis*. *De optimo genere oratoris*. *Orator*. *In Pisonem*. *De republica*. *Topica*. *Disputationes Tusculanae*. Latin Library.
- Cicero (Q. T.), *Commentariolum petitionis*. Latin Library.
- Hilary of Poitiers. *De trinitate. Tractatus super psalmos.* Patrologia Latina 9-10.

- *Historia Apollonii regis Tyri*. Latin Library.
- Hyginus, *Fabulae*. Latin Library.
- Jerome, *Epistolae*. *Contra Joannem*. *Vita Malchi*. *Vita Pauli*. Latin Library.
- Lactantius, *Institutiones divinae*. *De mortibus persecutorum*. Latin Library.
- Macrobius, *In Somnium Scipionis*. Wikisource. la.wikisource.org.
- Claudius Mamertus, *De Anima*. *Epistolae*. Patrologia Latina 53.
- Marius Victorinus, *Adversus Arium. In epistola ad Ephesios. In epistola ad Galatos. De generatione divini verbi. In epistola ad Philippenses*. Patrologia Latina 8.
- Minucius Felix, *Octavius*. Latin Library.
- Nazarius, *Panegyricus*. Patrologia Latina 3.
- Novatian, *De trinitate*. Latin Library.
- Pomponius Mela. *De chorographia*. Latin Library.
- Quintilian. *Institutiones.* Latin Library.
- (ps-) Quintilian. *Declamationes maiores*. Latin Library.
- Seneca the Elder, *Controversiae*. Latin Library.
- Seneca the Younger, *Epistulae morales ad Lucilium*. *Ad Galliam. De ira. Naturales quaestiones. De otio. De providentia*. *De tranquillitate*. *De brevitate vitae*. Latin Library.
- Tacitus, *Agricola*. *Annales*. *Germania*. *Historiae*. *Dialogus de oratoribus*. Latin Library.
- Victorinus of Poetovio. *In Apocalypsin*. Patrologia Latina 5.