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

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

REFERENCES

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 [17]. 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 i^{th} 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 [26–27]. We now use Ward's minimum variance method to cluster the documents

[28–29]. 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. 2, 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* in Fig. 2a 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 i^{th} 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:

$$\text{minmax}(\vec{X},\vec{Y}) = \sum_{i=1}^{n} \min(x_i, y_i) / \sum_{i=1}^{n} \max(x_i, y_i) .$$

Software

The BCT clustering experiments (Fig. 2) 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 Figures)

- 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.