ANALECTA ROMANA

INSTITUTI DANICI

XLIV

2019
(published 2021)

ROMAE MMXX

ANALECTA ROMANA INSTITUTI DANICI XLIV
© 2020 Accademia di Danimarca
ISSN 2035-2506

Analecta Romana Instituti Danici. — Vol. I (1960) — . Copenhagen: Munksgaard. From 1985: Rome, «L'ERMA» di Bretschneider. From 2007 (online): Accademia di Danimarca.

ANALECTA ROMANA INSTITUTI DANICI encourages scholarly contributions within the Academy's research fields. All contributions will be peer reviewed. Manuscripts to be considered for publication should be sent to: accademia@acdan.it
Authors are requested to consult the journal's guidelines: www.acdan.it

PHILOLOGY

THEN AND NOW

Proceedings of the Conference held at The Danish Academy in Rome, 16 July 2019

Stylometry in a Language without Native Speakers: A Test Case from Early Modern Latin

by Johann Ramminger

Abstract. The paper will use delta measures calculated by stylometric analysis of a corpus of Classical and Early Modern Latin (EML) texts. Within the larger question of whether delta measures can be successfully applied to EML texts, I will address several interconnected questions: (1) Do authors have a consistent style across genres? (2) Do translations from the Greek have a style different from "original" works by the translator? (3) Do translations of the same text by different translators resemble each other (an author versus a translator fingerprint)? And (4) are EML texts close in style to Classical Latin models? It will be shown that stylometry does not necessarily measure indicators of authorship in EML texts, but overall furnishes us with a plethora of stylistic information: Generally works by the same author are grouped together. Translations nearly always stand apart from original works by the same author, translations of the same work by different authors are grouped together. Works that strongly imitate the style of a Classical author (esp. Cicero), may be grouped accordingly. It appears that several humanists in our corpus exhibit such a depth of control over their Latin that they seem to be able to dissociate themselves from their "own" style at will. Thus delta in EML texts is not a reliable indicator of authorship, but rather seems to reflect the stylistic aspirations (and their success) of EML authors.

Introduction

Stylometry,[1] as used in the following, is originally a method of authorship attribution based on the assumption that every author has a characteristic stylistic signature expressed through the words he uses most frequently; it is assumed that the author exerts much less conscious control over that part of his vocabulary than over the lexicon that determines the actual content of a work.[2] The foundational study is Burrows 2002, which tested authorship attribution in Early Modern English poetry and developed the "delta" measure to express the distance between texts as an indication of authorship (now called Burrows' Delta or Classical Delta, since a number of other mathematical models have been proposed which, depending on the language, can produce better results).

Early Modern Latin as L2

Delta measures have been mostly applied to texts written by native speakers (L1) in Indo-European languages, including the Latin of antiquity (in the following called Classical Latin). Stylometric approaches have also been used for L2 texts, i.e. texts written by second-language learners, measuring them against contemporary L1 texts in the same language and in the L1 language of the L2 speakers. It has, however, not been sufficiently appreciated that Early Modern Latin (EML) is the

[1] I would like to thank the anonymous peer reviewer for a thorough reading of the text and stimulating comments.

[2] Bailey 1979 noted that the quantifiable properties of a text "should be salient, structural, frequent and easily quantifiable, and **relatively** immune from conscious control" (quotation from Holmes 1998, 111; my emphasis).

only widely used European language which for several hundred years had no contemporary native (L1) speakers; at the same time, the use of EML (obviously as L2) was governed by a fierce ideology of imitation of L1 texts written more than a millennium earlier. Furthermore, EML authors were extraordinarily self-reflective; even the minutiae of language use were routinely scrutinised within the humanist community.[3] It will be the overarching research question of this study whether and how the complete command of (a specific form of) Latin aspired to by EML authors – even if it surely remained aspirational at times – is visible in the stylometric analysis proposed here.

Test setup

- Corpus

The corpus I have designed consists of a Classical and an EML part. On the Classical side there are Cicero's orations (in three chronologically ordered groups) and letters, generally considered the gold standard of Latin by EML authors.[4] In addition to these, the major part of the data consists of Roman historians: the *Corpus Caesarianum*,[5] Sallust, Livy, Valerius Maximus, Curtius, Tacitus, Ammianus, and some derived texts, i. e. Florus and Eutropius derived from Livy, and Iulius Paris derived from Valerius Maximus (the latter three will provide opposite examples of the stylistic dis/parity between original text and rewritings of the same). I have added Iulius Valerius, although he is hardly a style icon for Quattrocento authors, because this text will allow us to test for the influence of topic similarity with Perotti's translation from Plutarch with similar content. I have bypassed the great historians of the late patristic age (e.g. Cassiodorus, Gregory of Tours, Jordanes), because I wanted to reduce the chronological disparity within the classical part of the corpus and exclude the influence of language change that would have come with their inclusion.

Among EML authors I have for the same reason – chronological coherence – included only authors of the "long" Quattrocento (the youngest text is Bembo's *Rerum Venetarum Historiae,* which finishes in 1513). The distinction between the two parts of our corpus is at the same time a division between L1 (Classical Latin) and L2 speakers (EML). I have selected only EML authors with (some form of) Italian as L1. Selecting authors under the same (or at least a similar) L1 influence is intended to neutralise the influence of L1.[6] To reduce the influence of genre conventions and content, I have selected mostly historical and chorographic authors (for the precise classification of the texts see Appendix 2). In addition there are Biondo's *De verbis romanae elocutionis* and Bruni's *Letters,* which will give us a different perspective on those writers' stylistic flexibility.[7] Since a considerable part of the historiographical writing in the Quattrocento consists of translations from the Greek or Latin rewritings of Greek works, this part of humanist text production is represented here too. An additional criterion of inclusion was the prestige gained among the humanists; this was of course a subjective point of view, but I assume that those authors considered the most prestigious would have the most complete control over their Latin – thus offering us an insight into the limits of conscious control (or its ab-

[3] There is a large number of controversies about Latin style in the Quattrocento alone. A typical and well-documented example is the controversy between Poggio Bracciolini and Lorenzo Valla (1452/1453); see Valla 1978.

[4] While the individual letters are as a rule too short for quantitative analysis, it has been shown that concatenation of short texts is a reliable way to amplify the authorial signal. See Eder 2015, 175.

[5] i.e. a set of six texts, besides *The War in Gaul* and *The Civil War* written by Caesar himself, a supplement to *The War in Gaul* by one of his officers, Hirtius, and accounts of Caesar's campaigns in Africa, Alexandria and Spain written by minor figures connected with Caesar.

[6] For the possibility of an L1 signal in L2 writings see generally Bestgen *et al.* 2012, and Horster 2013, 339–344.

[7] For the genre sensitivity of the Delta procedure see Craig & Burrows 2012, 36.

sence) over Latin. EML authors include Leonardo Bruni, Flavio Biondo, Lorenzo Valla, Niccolò Perotti, and Pietro Bembo.

To ensure that the stylometric approach used actually has produced meaningful results, there are certain controls on both parts of the corpus. For the classical part, the coherence of texts by the same author is a reliable indicator of successful authorship attribution. As for the EML authors, I have included Annius da Viterbo, who follows completely different stylistic conventions than his humanist contemporaries even in his chorographic work. Coherence among his works on different topics (Etruscan culture, loans, the end of the world) – which traditionally would have been governed by different stylistic conventions – will be a unique indicator of a common authorial signature (or lack thereof) of works with different content.

Only textual units longer than approximately five thousand words have been included in order to have a dataset that is adequate in view of a list of one thousand style markers used (see below).[8] EML texts have not been orthographically normalised (for the preparation of the texts see Appendix 2). Suffixes in any case show little variation (mostly the ae/-e ending of fem. gen. sg. and nom. pl.). With prefixes, there is some orthographic variety (e.g. with assimilated and non-assimilated compound forms); these have been left "as is" in view of the fact that authorship attribution in Latin is less sensitive to contamination than in other languages (see below).

- Parameters

The delta measure which works most reliably on Latin is the so-called Würzburg or Cosine Delta, i.e. a cosine distance with a z-scored matrix of values (i.e. scaled word frequencies).[9] This is also the delta variant used in the following. The use of delta measures became more easily accessible in 2016 with the development of the "stylo" script by Maciej Eder of the Computational Stylistics Group at the University of Krakow. This has been widely applied by researchers without a mathematical background and has also been used here.[10] Since the texts are of very unequal length, it is important for the quality of the results that the similarity is computed on a scaled (z-scored) matrix of word frequencies.[11] Eder has shown that for (Classical) Latin, text samples give reliable results already from about 2,500 words (based on the 200 most frequent words; see below).[12] Latin (i.e. the Latin of Antiquity, in the following called Classical Latin) behaves somewhat unusually compared to other languages insofar as delta measures work even on highly contaminated texts (with up to 40 per cent noise).[13] This allows us to include in our data also texts with a certain level of orthographic peculiarities or mistakes derived from OCR.

Style markers

The delta measurements can use two types of style markers, most frequent words (mfw, in raw form or lemmatised) or most frequent sequences of elements (so-called *n*grams).[14] Neither of these approaches is without problems.

- Most frequent words (mfw)

The mfw measure developed out of the original reliance of stylometry on function words,

[8] This excluded a large part of the (forged) texts "by other authors" inserted in Annius's *Antiquitates* which he claims to have found or acquired, which are shorter. Remarkably, preliminary tests did not show incorrect authorship attribution for these shorter texts (which were all attributed to Annius, in conformance with modern scholarship), possibly due to the resilience of the *n*gram measure or topic similarity.

[9] Evert *et al.* 2015; Bütttner *et al.* 2017. Applied e. g. in Hasse & Büttner 2018.

[10] See Eder *et al.* 2016.

[11] Description in Eder *et al.* 2019, 21–22.

[12] Eder 2015.

[13] Eder 2013a.

[14] The *n*gram approach has mostly been used with character *n*grams (POS-tag-*n*grams in Cafiero & Camps 2019, rhythmic patterns in Plecháč 2019). Word *n*grams are usually unreliable (see Eder 2011 and Hoover 2018).

which were assumed to be at the same time the most frequent words. Since measurements of style are supposedly distinct from measurements of content (hence the possibility of identical authorship attribution for works of different content), as a general rule content words are less suitable in measurements of stylistic distance. However, the distinction between (most frequent) function words and most frequent words in general has become increasingly fuzzy, and very few studies have rigorously focused on function words. There is no general rule on how many mfw are needed for valid results. Mandravickaité & Krilavičius 2017 used up to 10,000 in a study of Lithuanian parliamentary speeches, Plecháč 2019 in a study of a Shakespeare play only 500. In the latter, also characteristic orthographies are taken into account. In a recent study concerning Molière, Cafiero & Camps 2019 identified only about a hundred function words in French. The precision of authorship attribution in (Classical) Latin prose increases with the number of mfw, and reaches a plateau somewhere between 750 and 1,500 mfw.[15]

For Latin, no comprehensive definition of function words has been proposed so far, and a strict differentiation between function and content words may not be generally possible.[16] Also, content words can be just as frequent as function words; in the published list of the topmost 100 words in the mfw list used by Deneire 2018, there are words such as *vita*, *pater* (life, father). An estimate based on the frequency list of the archive of the *Neulateinische Wortliste* would put the number of function words in Latin – even with a most extensive definition – at considerably less than 500.[17] In addition, many functions in Latin are grammaticalised, which further limits the number of function words used (e.g. repetition can be expressed by *saepe*, *identidem* etc., but also by the imperfect indicative or by stems with suffixes indicating repetition, e.g. *factitare*). One of the most important Latin function words, *-que*, cannot be measured reliably, since as an enclitic it cannot be easily disambiguated from word forms with the same ending. In EML texts, it can also occur non-enclitically and be indistinguishable from the frequent orthographical variant *que* for the relative pronoun *quae* (adding a further uncertainty to our data). All this means that measurements relying on mfw in Latin will, if unfiltered, often implicitly express similarity of content (or genre) instead of style.

- Ngrams

An alternative is to rely on groups of letters as style markers (so-called "[character] *n*-grams"). For (Classical) Latin, 4-grams and 5-grams (i.e. sequences of four or five letters) have been shown to give better results than the mfw-based approach.[18] While character *n*-grams undoubtedly give very good results, it has often been noticed that no explanation of why they work has been put forward.[19] As far as Latin is concerned, several explanations of what is actually measured by *n*-grams can be suggested. *Prefix n*-grams will measure typical compounds (e.g. 3-grams will catch compounds with the most frequent prefixes *per*-, *pro*-, *sub*-; *prae*-either as PRA or as RAE). *Suffix n*-grams would in Latin (in contrast to English, where they have performed poorly) catch ample stylistic information owing to the large amount of grammatical information contained in Latin suffixes. Both *prefix* and *suffix n*-grams may catch stylistic idiosyncrasies in the work of an author who may have favourites among equivalent modes of expression – though I would hesitate to call those features uncontrolled or uncontrollable. *N*-grams which bridge word divisions might to some degree measure typi-

[15] Eder 2011. Rybicki & Eder 2011.

[16] A list of (Medieval) Latin function words tailored to the Avicenna project is published in Hasse & Büttner 2018, 360–361 n.42.

[17] Ramminger 2003-.

[18] Eder 2011.

[19] Stamatatos 2013; Sapkota *et al.* 2015.

cal word bi-grams (though intuitively one would expect this to require *n*-grams of six to eight characters in length). Mid-word *n*-grams could in Latin in some way identify root words of frequent compounds: it has been shown that in English these are best in single-domain settings (i.e. texts from the same domain), because they convey topic-related information, and this may also be true for Latin.[20] On account of these limitations, the latter two types of *n*-grams have not been used here.

In general, Latin *n*-grams, if unfiltered, share some of the problems of mfw. Shorter pronouns, prepositions and similar will appear intact as *n*-grams, but so will content words (such as *vita,* quoted above). To mitigate this, the following exploration will use a somewhat novel approach using *prefix* and *suffix* 4-grams. Conventionally, stylometrics relies on wordlists generated from the text corpus itself. The procedure to extract the variables from a dataset and then to use them to measure subsets of the same dataset, is favourably biased towards detecting similarities. Whether this is an advantage or disadvantage has not been explored.[21] However, some objections against corpus-internal lists of style markers can be raised. As such a selection cannot take into account those indicators of style that may be rare in the corpus but frequent within a wider range of texts, this approach possibly misses some significant stylistic markers. Equally, a corpus-dependent selection of style markers makes the comparison of different corpora difficult, since the addition of a text will automatically change the pool of significant style markers and thus change the basis for the comparison. Instead, I have based myself on a list of the words in the archive of the *Neulateinische Wortliste* (350 million words) from texts from the fourteenth to sixteenth centuries (excluding those from the present corpus). A part of this corpus has normalised orthography, a part preserves the orthography of the source. I extracted a list of words with six or more letters (224,675 words). These occur in total about five million times in the texts of the NLW (5,539,724 instances). From these, I extracted a list of the thousand most frequent 4-grams occurring either at the beginning (positions 1 to 3) or the end (length minus 6 to 4) of a word.[22] This has the advantage that different or modified datasets can be compared, since the basic *n*-gram list remains unmodified (rather than being recalculated with every change in the data).

Research questions

The corpus designed for this study can cover only a small part of the question of how Early Modern Latin behaves stylistically in comparison to Classical Latin. Bearing the limitations of our dataset in mind, I will in the following address several interconnected questions about style in EML literary texts. (1) Do authors have a consistent style across genres? (2) Do translations from the Greek have a style different from "original" works by the translator?[23] (3) Do translations of the same text by different translators resemble each other (an author versus a translator fingerprint)? And (4) are EML texts close in style to Classical Latin models (especially Cicero as the often-vaunted model of humanist writing)?

Results

- The Classical dataset

In our tests, all L1 authors are clearly recognised (with a secondary genre signature in the

[20] Sapkota *et al.* 2015.

[21] See however Bestgen *et al.* 2012, 132.

[22] E.g. *contrectationis* would supply CONT ONTR NTRE as *prefix n*grams, and TION IONI ONIS as *suffix n*grams, but not TREC RECT ECTA CTAT TATI ATIO. This avoids some content information, but gives information about word composition (ONTR will mostly occur in contra-compounds) and word form (TION will mostly reference a substantive of the third declension, ONIS will often indicate the genitive singular). For the three-letter prefixes *per-*, *pro-*, *sub-* the length of 4-grams will mean that they will be considered as types of compounds including the first letter of the root word, thinning the content information.

[23] See Eder 2013b and Rybicki 2013.

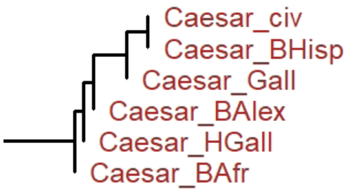


Fig. 1. The *Corpus Caesarianum* (Figs. 1–4 are details of Appendix 1; abbreviations for individual works are explained in Appendix 2)

case of Cicero). As already shown by Deneire, Cicero has a characteristic authorial fingerprint: even the different genres (speeches, philosophical works, letters) are reliably recognised (see Fig. 2).

Notably, the *Corpus Caesarianum* is a separate cohesive unit, largely distant from other texts, even though written by different authors (see Appendix 1). At the moment I have two tentative explanations: (1) that what we are seeing is a strong genre signature (unfortunately we have no other texts to test this hypothesis), and (2) that the vocabulary is homogenous to a degree that overwhelms the "defensive" measures applied in the selection of *n*-grams.

Sallust's *Iugurtha* and *Catiline* and the speeches from the otherwise lost *Historiae* are always together, as is Livy and the texts derived from him (Iulius Paris, Florus, Eutropius). Tacitus is stylistically a fascinating case: the Würzburg Delta produces a deep bifurcation, though still on the same branch of our dendrogram, grouping together the *Annals*, *Histories* and the *Agricola*. The *Germania* is placed at a slightly larger distance, while the *Dialogus*, generally recognised as the most Ciceronian of Tacitus's works, is inserted into the Ciceronian branch of our dendrogram (see Fig. 2).

If this attribution is more than a fluke, it (as well as the sub-groupings of Cicero) would indicate that L1 writers of Latin controlled their language to a degree that overcame the limits of the authorial fingerprint.

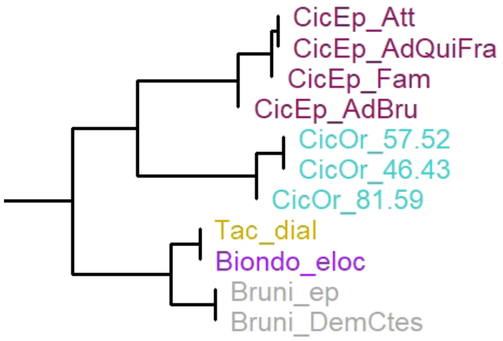


Fig. 2. The *Ciceronian* branch, including Tacitus's *Dialogus* and "Ciceronian" works by Biondo and Bruni

- EML translations

One result of the stylometric analysis is that translations nearly always stand apart from original works by the same author.[24] Valla's *Thucydides* translation is placed far away from his *Gesta Ferdinandi regis*, Perotti's translation of Plutarch's *De Alexandri Magni fortuna aut virtute* is stylistically diverse from his translation of Polybius. This may be evidence for an authorial fingerprint on the part of the Greek author that is strong enough to drown out the stylistic signal of the translator. The only case where the translator is clearly visible is Bruni's translation of Demosthenes's *Pro Ctesiphonte*, which is consistently grouped with his letters and close to Cicero (see Fig. 2).

Confirmation of the existence of an authorial signature remaining in the translation of a text (as opposed to the translator's) may be found in the one case in our corpus where we have two translations of the same text: the translations from *Polybius* by Bruni and Perotti (Perotti's is the later by some thirty years; Fig. 3).

These have overlapping content: Bruni's text comprises only part of the Greek text to the middle of the second book of Polybius, whereas Perotti's is a translation of all five books of Polybius's *History* that were known in the Renaissance. It has been shown that Perotti knew Bruni's translation and, in some

[24] See Rybicki 2012. A recent discussion of the translator's "invisibility" is in Konjhodžić 2018.

details, used it.[25] On the other hand, they were written with different stylistic agendas. Perotti's is what has been termed a "domesticating" translation, Bruni's *Polybius* more a paraphrase than a translation proper,[26] so that it is not surprising that it is stylistically close to his "original" works. Since we have no original historical works by Perotti, our comparison remains lopsided. To show a possible authorial signature of Bruni in Perotti's text, I have divided Perotti's *Polybius* into two parts, with the first one corresponding to the part also used by Bruni. However, from a stylometric point of view, Perotti's use of Bruni has no impact on the style: both parts of Perotti's *Polybius* are stylometrically indistinguishable. Thus, the stylometric closeness of Bruni's and Perotti's *Polybius* may be an indication of the common authorial fingerprint of Polybius rather than of text reuse by the later translator.

- Original EML texts

Stylometric authorship attribution seems to work well with original EML texts. Bruni's two original historical works, the *Rerum suo tempore gestarum commentarius* and the *Historiae Florentini populi*, are recognised as cognate texts.

Equally, stylometry offers proof (if any were still needed) that the supposed older texts that Annius "edited" and commented upon in his *Antiquitates* are forgeries written by himself.[27] The only text long enough for analysis is the *Berosus*, proffered as a Chaldean text supposedly translated by Armenian monks; stylometrically this is indistinguishable from the texts (the commentaries) acknowledged by Annius as his own (Fig. 4).

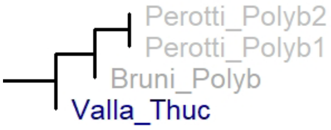


Fig. 3. The *Polybius* translations by Bruni and Perotti (part one of Perotti corresponds to the part of Polybius used by Bruni).

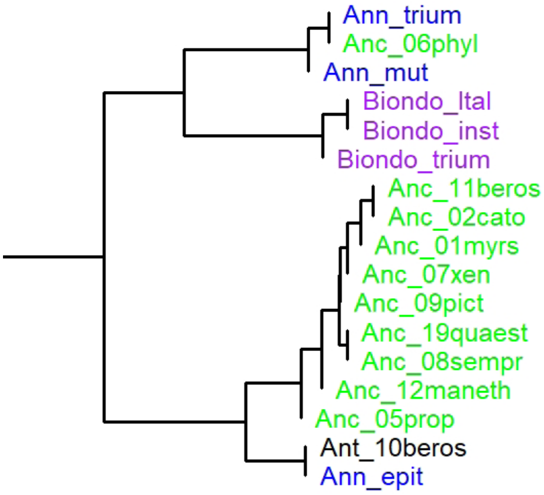


Fig. 4. The *Annius* branch, including works by Biondo (the *Berosus* forgery is in black).

That this (and other shorter texts in the *Antiquitates*) is a forgery has long been undisputed in Annius scholarship; our evidence shows that Annius did not even try (or have the command of Latin necessary) to impart a different stylistic signature on texts supposedly not written by him. As with the *Corpus Caesarianum*, it may be due to an overwhelming similarity in content or a genre signature that the two works of Annius in our corpus that are not historical, *De futuris christianorum triumphis in saracenos* (his earliest known work) and *De mutuo iudaico*, are largely separated from the *Antiquitates*, although still on the same general branch, that is, correctly attributed to the same author.

Flavio Biondo is, just like his fellow humanists, a writer consciously modulating his Latin in relationship to the classical idiom. He, probably to a higher degree than his contemporaries, articulates a pragmatic and thus at times unclassical approach to writing Latin.[28] This may be reflected in the stylometric analysis. Not only are his (chorographic)

[25] See Pade 2008. Charlet 2011, 13 n. 4 lists further relevant literature.
[26] For the stylistic implications see Pade 2018.
[27] On Annius's Etruscan studies see Rowland 2016; a general introduction is Fubini 2012.
[28] Ramminger 2014.

works correctly grouped together, they are also near Annius, who – whether intentionally or not – wrote in a distinctly unclassical idiom (see Fig. 4).

- Imitation of Classical Latin

As concerns the imitation of Classical Latin authors by the humanists, the stylometric evidence is uneven. That Perotti's *De fortuna Alexandri* is close to Iulius Valerius's translation of the *Historia Alexandri Magni* has so far no independent support. Rather, the supposed proximity may result from topic similarity. That Bruni's *Pro Ctesiphonte* is somewhat Ciceronian in style fits well with what we know about other aspects of Bruni's translation activity (see Pade in this volume). The stylometric analysis shows an impressive linguistic achievement on Bruni's part (Fig. 2).

Equally well placed are Valla's *Gesta Ferdinandi,* at a distance from Cicero and in the large group that also comprises the historiographers of the post-Ciceronian period. From a different angle, this accords well with Tunberg's observations on the style of the work.[29]

- Genre signature

While the dataset was explicitly construed so as to minimise a genre signature (by having texts mostly from the same genre), it is nevertheless hard to ignore that one of the top divisions in our dendrogram coincides with a genre division, locating all the historiographical texts on the same major branch (except for Annius and Biondo). We can exclude this being a chronological signature, because Classical and EML texts appear grouped together. Whether this division is (1) a consequence of the EML authors' imitative approach to writing, thus indicating that humanist historiographers (successfully) imitated classical works in the same genre (what we might call a "secondary" genre signature), or (2) a genuine genre signature of historiography despite the huge chronological disparity can only be decided with a larger corpus of text in different genres.

Conclusion and further research

The basic assumption of stylometric analysis – that there are parts of language which are "under the radar" in an author's writing – must remain in doubt as a general rule in humanist Latin literature. Several of the humanists in our corpus exhibit such a depth of control over their Latin that they seem to be able to dissociate themselves from their "own" style at will. Thus delta is in several cases unreliable as a diagnostic tool for authorship attribution. On the other hand, we saw that it was extraordinarily useful in measuring the stylistic aspirations of our authors (and their success). Two of the authors of our corpus produce distance measures that are clearly additional indicators of authorship: Annius, who comes from a scholastic tradition of writing Latin, and Biondo, who has a pragmatic approach to Latin in which imitation is of little importance (although Biondo can at will shift his style into an imitative direction). It needs to be emphasised that these results were obtained from authors of what might be called the "Golden Age" of Latin humanist writing. Later authors, and authors who by choice or capacity exert less control over their writing, might yield different results. Furthermore, it remains to be seen whether authors with other L1 languages will exhibit a discernible L1 signal in their Latin. In addition, the use of a more diverse corpus will bring questions of the genre fingerprint – which have played a marginal role above – to the fore. The behaviour of verse texts is so far unexplored.

As for the method used, fixed-length *n*-grams have been very successful. Nevertheless the method of selecting most frequent *n*-grams from a larger corpus rather than the

[29] Tunberg 1988.

texts under purview here will need validation in comparison both to other methods of selection and to other research corpora.[30] But further research will have to add other approaches. Variable-length *n*-grams should be explored.[31] To increase reliability, the minimum text length (in relation to the number of style markers and the differences in length between the texts of the corpus) needs to be established more carefully.[32] Most importantly, a list of Latin function words (and not-topic-related open-class words) appropriate to Early Modern Latin will have to be devised so as to gain a different perspective on these distance measurements: one that uses words rather than *n*-grams (this will need texts that are orthographically normalised as far as function words are concerned). A further promising direction is an approach recently successfully applied by Cafiero & Camps 2019, who have measured POS *n*-grams.[33] Part-of-speech tagging of EML texts, however, is not a trivial matter, owing to lexical diversity and the large amount of data needed. Tests will establish how much of a given text has to be successfully POS-tagged for delta measurements to produce results. Finally, questions of lexical richness and intensity of imitation will need to be explored – though these may lie beyond stylometric approaches proper.

Johannes Ramminger
Thesaurus Linguae Latinae
Munich

[30] I would like to thank the anonymous reviewer for reminding me of the necessity of ample validation.

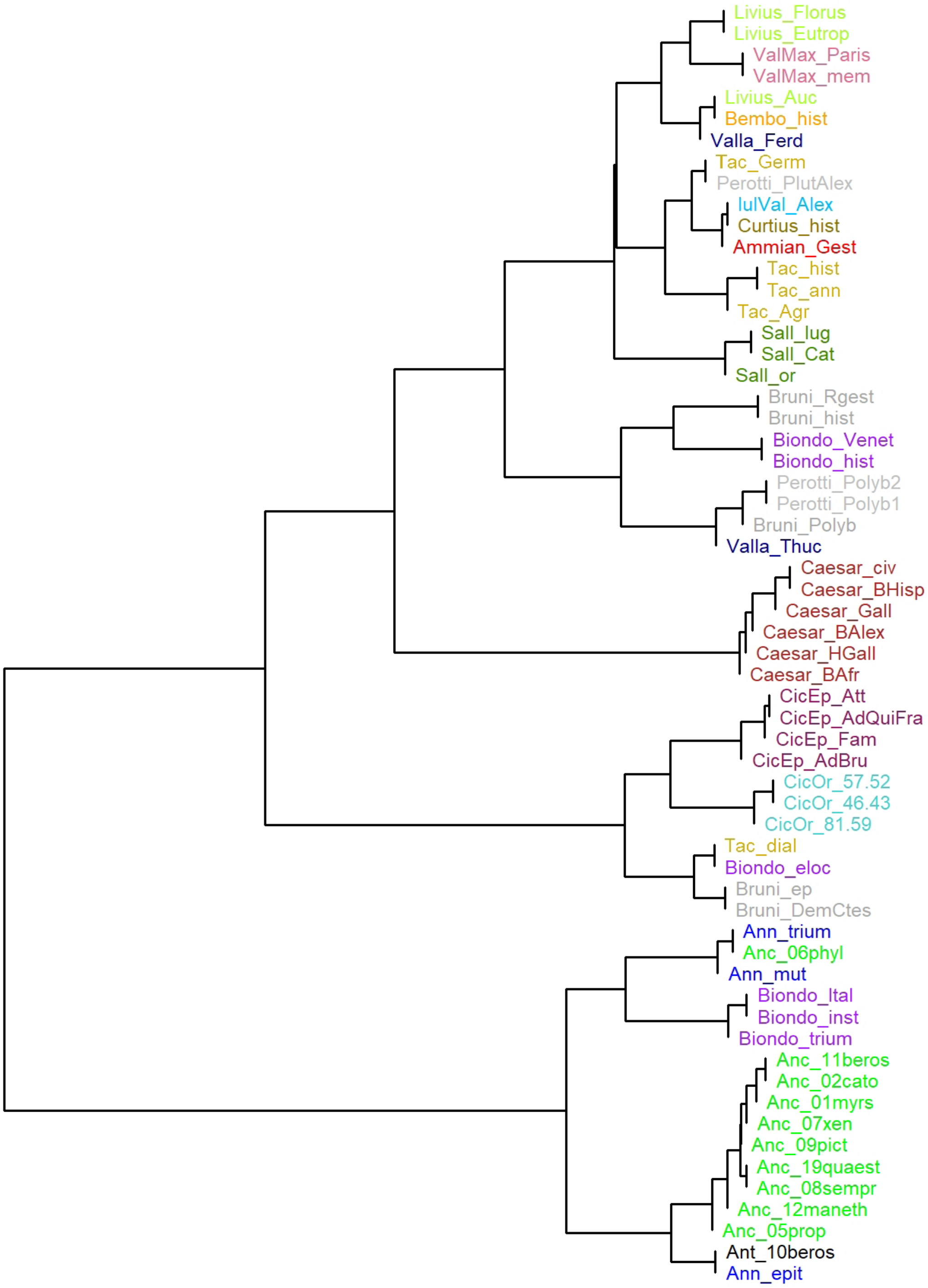
[31] See Kestemont *et al.* 2019.

[32] Scaling the frequencies of style markers in small-size documents has implications for the reliability of the data; see Moisl 2011.

[33] On syntactic patterns as authorial fingerprint see Stamatatos 2009; for POS-tag *n*grams of (Classical) Latin poetry see Eder 2015.

APPENDIX 1

Complete dendogram



Historiographic and chorographic texts of the Quattrocento. A graph produced with Stylo (see Eder *et al.* 2016). Cosine Delta distance (aka Würzburg), 1,000 character *4*grams (abbreviations see Appendix 2).

APPENDIX 2

The text corpus

The study is based on a corpus of 3,658,176 words, divided into Classical Latin, 2,098,494 words (shortest text Sall_or 4,163 words, longest Livius_Auc 561,121), and EML (six authors), 1,559,682 words (shortest Ant_10beros 5,042 words, longest Biondo_hist 368,618).

The Classical texts used are from Perseus (<http://www.perseus.tufts.edu/>). The titles and dates given are those found in the Index Librorum of the *Thesaurus Linguae Latinae* (<http://www.thesaurus.badw.de/tll-digital/index.html>). All dates are AD unless otherwise indicated.

The EML texts are, where not otherwise indicated, my own, and have been produced with OCR4all (<https://github.com/OCR4all>) from the editions named in the following.[34] All abbreviations were resolved, hyphenated words at linebreaks were joined together (including words separated without a hyphen), scanning mistakes were corrected. Running titles, page numbers, marginalia and catchwords were eliminated. Accented vowels were replaced by vowels without accent (i/j and u/v are harmonised, and capital letters are treated as small letters by Stylo during the execution of the script). Otherwise the texts were not normalised orthographically. I am aware of the fact that several of the EML texts used would be available in modern editions; in order to avoid problems of copyright in a subsequent publication of the sources, I have not used these.

Abbreviations

EDIT16: EDIT16. Censimento nazionale delle edizioni italiane del XVI secolo (<http://edit16.iccu.sbn.it>)

ISTC: Incunable Short Title Catalogue (<https://data.cerl.org/istc/>)

VD16: Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts (<https://www.bsb-muenchen.de/sammlungen/historische-drucke/recherche/vd-16/>)

VD17: Das Verzeichnis der im deutsche Sprachraum erschienenen Drucke des 17. Jahrhunderts (<http://www.vd17.de>).

Sigla

Classical Latin

Ammian_Gest: Ammianus Marcellinus (c. 390), Rerum gestarum libri [History]. 125,282 words.

Caesar_BAfr: Unknown author (contemporary of Cesar), De bello Africo [The African War]. 13,766 words.

Caesar_BAlex: Unknown author (contemporary of Cesar, Hirtius?), De bello Alexandrino [The War in Alexandria]. 11,094 words.

Caesar_BHisp: Unknown author (contemporary of Cesar), De bello Hispaniensi [The Spanish War]. 112,062 words.

Caesar_civ: C. Iulius Caesar (100-44 BC), Commentarii belli civilis [An Outline of the Civil War] (ca. 45 BC). 33,893 words.

Caesar_Gall: Commentarii belli Gallici [An Outline of the War in Gaul] (52/51 BC). 33,432 words.

Caesar_HGall: A. Hirtius, Liber 8 Caesaris commentariorum belli Gallici libris septem additus [The Eight' Book of Cesar's War in Gaul, a Supplement] (after Cesar's death). 6,907 words.

[34] I would like to thank C. Reul, U. of Würzburg, the lead developer of OCR4all, for his patient advice throughout the digitization of the EML texts in this corpus.

CicEp_AdBru: M. Tullius Cicero (106–43 BC), Epistulae ad M. Iunium Brutum [Letters to Brutus] (43 BC). 6,741 words.

CicEp_AdQuiFra: Epistulae ad Quintum fratrem [Letters to the brother Quintus] (60/59-54 BC). 20,279 words.

CicEp_Att: Epistulae ad Atticum [Letters to Atticus] (68-44 BC). 138,326 words.

CicEp_Fam: Epistulae ad familiares [Letters to Acquaintances] (62-43 BC). 101,280 words.

CicOr_46-43: Orationes [Speeches] (46-43 BC). 67,635 words.

CicOr_57-52: Orationes [Speeches] (57-52 BC). 117,458 words.

CicOr_81-59: Orationes [Speeches] (81-59 BC). 257,950 words.

Curtius_hist: Q. Curtius Rufus (1st cent.), Historiarum Alexandri Magni libri [History of Alexander the Great]. 75,854 words.

IulVal_Alex: Iulius Valerius (4th cent.), Historia Alexandri Magni [History of Alexander the Great] (from a Greek original). 33,966 words.

Livius_Auc: T. Livius (59 BC–17), Ab urbe condita [History from the Foundation of the City]. 561,121 words.

Livius_Eutrop: Eutropius, Breviarium ab urbe condita [Digest of Livy's *Ab urbe condita*] (c. 369). 19,268 words.

Livius_Florus: L. Annaeus Florus (2nd cent.), Epitoma de Tito Livio [Digest of Livy's *Ab urbe condita*]. 26,875 words.

Sall_Cat: C. Sallustius Crispus (86–34 BC), De coniuratione Catilinae [The Plot of Catiline] (c. 42/41). 11,549 words.

Sall_Iug: De bello Iugurthino [The War against Jugurtha] (c. 40 BC). 22,743 words.

Sall_or: [Orations and letters from the *Historiae*] (39–34 BC). 4,163 words.

Tac_Agr: (P.) Cornelius Tacitus, De vita Iulii Agricolae [The Life of Julius Agricola] (98). 7,908 words.

Tac_ann: Annalium (ab excessu divi Augusti) quae exstant [Annals from the Death of the Emperor Augustus] (beginning 2nd cent.). 104,773 words.

Tac_dial: Dialogus de oratoribus [Dialogue about Orators] (beginning 2nd cent.). 10,684 words.

Tac_Germ: De origine et situ Germanorum [The Origin and Condition of the Germans] (beginning 2nd cent.). 6,527 words.

Tac_hist: Historiae [History] (beginning 2nd cent.). 60,379 words.

ValMax_mem: Valerius Maximus, Facta et dicta memorabilia [Memorable deeds and sayings] (shortly after 31). 82,311 words.

ValMax_Paris: Iulius Paris (4th cent.), Epitome Valerii Maximi [A Digest of Valerius Maximus]. 24,268 words.

Early Modern Latin

Annius da Viterbo (1432–1502)

Antiquitates [Antiquities] (1498). Ed. used: Auctores vetustissimi. Romae: Eucharius Silber, 1498 (ISTC ia00748000).

Anc_01myrs: Commentaria super Myrsilum. 5,438 words.

Anc_02cato: Commentaria super fragmenta Catonis. 15,246 words.

Anc_05prop: Commentaria super Vertunnianam Propertii. 6,045 words.

Anc_06phyl: Commentaria super Phylonis Breuiarium de temporibus. 11,023 words.

Anc_07xen: Commentaria super Xenophontem de equiuocis. 7,446 words.

Anc_08sempr: Commentaria super Sempronium De diuisione & chorographia Italiae. 6,415 words.

Anc_09pict: Commentaria super Fabium Pictorem De aureo seculo. 10,448 words.

Anc_11beros: Commentaria super Berosum. 42,925 words.

Anc_12maneth: Commentaria super supplementa Manethonis. 7,071 words.

Anc_19quaest: Anniae Quaestiones. 22,081 words.

Ant_10beros: Berosus Babylonicus, De temporibus. 5,042 words.

Ann_epit: Viterbiae historiae epithoma [Digest of the history of Viterbo] (1491/92). Text used: I have collated the text given by Baffioni (Annius of Viterbo 1981) with the only ms., BAV lat. 6263, and inserted some corrections. 11,065 words.

Ann_mut: Questiones due disputate super mutuo iudaico et ciuili et diuino [Two questions about loans from Jews, from citizens, and from God] (1492). Ed. used: Pro monte pietatis consilia [Venice : Johannes Tacuinus, de Tridino, 1494/1498] (ISTC im00810300). 9,520 words.

Ann_trium: De futuris christianorum triumphis [The future triumphs of the Christians] (1480). Ed. used: Genuae: Baptista Cavalus, 1480 (ISTC ia00750000). 26,571 words.

Pietro Bembo (1470–1547)

Bembo_hist: Pietro Bembo (1470–1547), Rerum Venetarum Historiae libri [History of Venice] (1487–1513). Text used: Petri Bembi Cardinalis Historiae Venetae libri XII, Venetiis 1551 (EDIT16: CNCE 5037). 129,317 words.

Biondo Flavio (1388–1463)

Biondo_eloc: De verbis romanae elocutionis [The lexicon of speech in Rome] (1435). Text used: ALIM (<http://www.alim.dfll.univr.it/>). 5,339 words.

Biondo_hist: Historia ab inclinatione Romanorum imperii [History from the Decline of the Roman Empire]. Text used: Basileae 1531 (VD16 B 5541). 368,618 words.

Biondo_inst.txt: Roma instaurata [Rome restored]. Text used: Basileae 1531 (VD16 B 5541). 34,081 words.

Biondo_Ital.txt: Italia illustrata [Description of Italy]. Text used: Basileae 1531 (VD16 B 5541). 84,663 words.

Biondo_trium.txt: Roma triumphans [The Triumph of Rome]. Text used: Basileae 1531 (VD16 B 5541). 144,879 words.

Biondo_Venet: De origine et gestis Venetorum [Origin and History of Venice]. Text used: Basileae 1531 (VD16 B 5541). 12,358 words.

Leonardo Bruni (1370/74–1444)

Bruni_DemCtes: Translation of Demosthenes, De corona vel pro Ctesiphonte [About the Crown or For Ctesiphon] (1407). Ed.: Cicero, De oratore et al., Venetiis 1485 (ISTC ic00662000). 18,740 words.

Bruni_ep: Epistolae [Letters]. Ed.: Leonardi Bruni Arretini Epistolarum libri VIII, recensente Laurentio Mehus, Florentiae 1741. 80,071 words.

Bruni_hist: Historiae Florentini Populi [History of the Florentines]. Ed.: Leonardi Aretini Historiarum Florentinarum libri XII : quibus accesserunt quorundam suo tempore in Italia gestorum & de rebus Græcis commentarii. Argentorati 1610 (VD17 23:231905N). 154,328 words.

Bruni_Polyb: Polybius De primo bello Punico [Polybius on the First Punic War] (1419). An adaptation of the first and part of the second book. Text used: Brixiae: Iacobus Britannicus 1498 (ISTC ib01254000). 31,110 words.

Bruni_Rgest: Rerum suo tempore gestarum commentarius (after 1440). Ed.: Rerum Italicarum Scriptores 19, ed. L. A. Muratori, Mediolani 1731. 11,735 words.

Niccolò Perotti (1430–1480)

Perotti_PlutAlex.txt: Niccolò Perotti, Translation of Plutarch, De Alexandri Magni Fortuna aut Virtute (1459/50). Ed.: Barberini Latin Manuscripts 47-56 and Niccolò Perotti's Latin Version of the De Alexandri Magni fortuna aut virtute of Plutarch, by Bernard J. Cassidy Diss. Fordham University. New York 1967.

<https://fordham.bepress.com/dissertations/AAI6803683/> (License: Open access). 9,556 words.

Perotti_Polyb: Polybii historiarum libri [History, by Polybius] (1454). Text used: Romae, Sweynheym & Pannartz, 1472, copy of the Biblioteca Apostolica Vatican (ISTC ip00907000). Split into two parts with 29,600 (coextensive with Bruni_Polyb) and 72,679 words.

Lorenzo Valla (1407–1457)

Valla_Ferd: Lorenzo Valla, Historiae Ferdinandi Regi Aragonum (1445/46). Text used: Parisiis 1521. 42,497 words.

Valla_Thuc: Thucydidis Historiarum Pelopennensium libri [The War on the Peloponnes by Thucydides] (1452). Text used: [Treviso : Johannes Rubeus Vercellensis, 1483?] (ISTC: it00359000). 142,187 words.

Classification of texts

Historical or chorographic texts:

Classical Latin:

Ammian_Gest.txt Caesar_BAfr.txt Caesar_BAlex.txt Caesar_BHisp.txt Caesar_civ.txt Caesar_Gall.txt Caesar_HGall.txt Curtius_hist.txt IulVal_Alex.txt Livius_Auc.txt Sall_Cat.txt Sall_Iug.txt Sall_or.txt Tac_Agr.txt Tac_ann.txt Tac_Germ.txt Tac_hist.txt Valla_Ferd.txt Valla_Thuc.txt ValMax_mem.txt ValMax_Paris.txt

Early Modern Latin:

Anc_01myrs.txt Anc_02cato.txt Anc_05prop.txt Anc_06phyl.txt Anc_07xen.txt Anc_08sempr.txt Anc_09pict.txt Anc_11beros.txt Anc_12maneth.txt Anc_19quaest.txt Ann_epit.txt Ant_10beros.txt Bembo_hist.txt Biondo_hist.txt Biondo_inst.txt Biondo_Ital.txt Biondo_trium.txt Biondo_Venet.txt Bruni_hist.txt Bruni_Polyb.txt Bruni_Rgest.txt Perotti_PlutAlex.txt Perotti_Polyb1.txt Perotti_Polyb2.txt Picc_co3.txt

Others:

Classical Latin:

CicEp_AdBru.txt CicEp_AdQuiFra.txt CicEp_Att.txt CicEp_Fam.txt CicOr_46-43.txt CicOr_57-52.txt CicOr_81-59.txt Tac_dial.txt

Early Modern Latin:

Ann_mut.txt Ann_trium.txt Biondo_eloc.txt Bruni_DemCtes.txt Bruni_ep.txt

BIBLIOGRAPHY

Annius of Viterbo
1981 *Viterbiae historiae epitoma, opera inedita di Giovanni Nanni da Viterbo*. [Ed. crit. G. Baffioni. *Annio da Viterbo*], *Documenti e ricerche* I, Roma.

Bailey, R. W.
1979 "Authorship Attribution in a Forensic Setting." In: Ager, D. E. & Knowles, F. E. & Smith, J. (eds.), *Advances in Computer-Aided Literary and Linguistic Research*, Birmingham, 1-20.

Bestgen, Y. *et al.*
2012 "Error Patterns and Automatic L1 Identification." In: Scott Jarvis, S. & Crossley, S. A. (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-Based Approach*, Bristol *et al.*, 127-153.

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

Büttner, A. *et al.*
2017 "*Delta* in der stilometrischen Autorschaftsattribution," *Zeitschrift für digitale Geisteswissenschaften* 2017, 2. DOI: 10.17175/2017_006.

Cafiero, F. & Camps, J.-B.
2019 "Why Molière most likely did write his plays," *Science Advances* 5, 11 <advances.sciencemag.org/content/5/11/eaax5489/tab-pdf>.

Charlet, J.-L.
2011 "Nicolas V, Niccolò Perotti et la traduction latine de Polybe: le mécène, l'humaniste et son public." In: Secchi Tarugi, L. (ed.), *Mecenati, artisti e pubblico nel Rinascimento: atti del XXI Convegno internazionale, Pienza-Chianciano Terme 20-23 luglio 2009*, Firenze, 13-24.

Craig, H. & Burrows, J.
2012 "A collaboration about a collaboration: the authorship of King Henry VI, part three." In: Deegan, M. & McCarty, W. (eds), *Collaborative Research in the Digital Humanities,* Farnham, 27-65.

Deneire, T.
2018 "Filelfo, Cicero and Epistolary Style: a Computational Study." In: De Keyser, J. (ed.), *Francesco Filelfo, Man of Letters*, (*Brill's Studies in Intellectual History* 289), Leiden *et al.*, 239–270.

Eder, M.
2011 "Style-Markers in Authorship Attribution. A Cross-Language Study of the Authorial Fingerprint," *Studies in Polish Linguistics* 6, 101-116.

Eder, M.
2013a "Mind your corpus: systematic errors in authorship attribution," *Literary and Linguistic Computing* 28, Issue 4, 603–614. <doi.org/10.1093/llc/fqt039>.

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

Eder, M.
2015 "Does size matter? Authorship attribution, small samples, big problem." *Digital Scholarship in the Humanities* 30/2, 167-182.

Eder, M. *et al.*
2016 "Stylometry with R: A Package for Computational Text Analysis," *The R Journal* 8/1, 107-121.

Eder, M. *et al.*
2019 *Package 'stylo.' Stylometric Multivariate Analyses. Version 0.6.9. Manual.* <github.com/computationalstylistics/stylo>.

Evert, S. *et al.*
2015 "Towards a better understanding of Burrows's Delta in literary authorship attribution." In: *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, Denver, 79-88.

Fubini, R.
2012 "Nanni, Giovanni (Annio da Viterbo)." In: *Dizionario Biografico degli Italiani* 77. <www.treccani.it/enciclopedia/giovanni-nanni_%28Dizionario-Biografico%29/> (accessed December 24, 2019).

Hasse, D. N. & Büttner, A.
2018 "Notes on Anonymous Twelfth-Century Translations of Philosophical Texts from Arabic into Latin on the Iberian Penisula." In: Hasse, D. N. & Bertolacci, A. (eds.), *The Arabic, Hebrew and Latin Reception of Avicenna's Physics and Cosmology*, Boston & Berlin, 313-370.

Holmes, D. I.
1998 “The evolution of stylometry in humanities scholarship,” *Literary and Linguistic Computing* 13, 111-117.

Holmes, D. I.
2012 “Stylometry.” In: Balakrishnan, N. (ed.), *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*, Hoboken, N.J., 310-317.

Hoover, D. L.
2018 “Authorship Attribution Variables and Victorian Drama: Words, Word-Ngrams, and Character-Ngrams.” Abstract for DH2018. <dh2018.adho.org/en/authorship-attribution-variables-and-victorian-drama-words-word-ngrams-and-character-ngrams/>.

Horster, C.
2013 *The Grammar of Imitation. A Corpus Linguistic Investigation of Morpho-Syntactic Phenomena in Fifteenth-Century Italian Neo-Latin*. Unpublished Diss. Aarhus University. <soeg.kb.dk/permalink/45KBDK_KGL/fbp0ps/alma99123153780105763>.

Kestemont, M. *et al.*
2019 “Overview of the Cross-Domain Authorship Attribution Task at PAN 2019.” In: *Acta of CLEF 2019, 9-12 September 2019*, Lugano. <ceur-ws.org/Vol-2380/paper_264.pdf>.

Konjhodžić, E.
2018 “Stylometric Study of Ivo Andrić’s Short Stories.” In: Hadžikadić, M. & Avdaković S. (eds.), *Advanced Technologies, Systems, and Applications II: Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT)*, Cham, 265-270.

Mandravickaité, J. & Krilavičius, T.
2017 “Stylometric Analysis of Parliamentary Speeches: Gender Dimension.” In: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Valencia, 102-107.

Moisl, H.
2011 “Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora,” *Journal of Quantitative Linguistics*, 18/1, 23-52.

Pade, M.
2008 “Niccolò Perotti and the *ars traduce*ndi.” In: Ebbersmeyer, S. *et al.* (eds.), “*Sol et homo.” Mensch und Natur in der Renaissance Festschrift zum 70. Geburtstag für Eckhardt Keßler*, München, 79-100.

Pade, M.
2018 “Greek into Humanist Latin: Foreignizing vs. domesticating translation in the Italian Quattrocento.” In: den Haan, A. *et al.* (eds.), *Issues in Translation Then and Now: Renaissance theories and translation studies today* (*Renæssanceforum* 14), 1-23.

Plecháč, P.
2019 “Relative contributions of Shakespeare and Fletcher in Henry VIII: An Analysis Based on Most Frequent Words and Most Frequent Rhythmic Patterns,” *arXiv*:1911.05652v1 [cs.CL], <arxiv.org/abs/1911.05652>.

Ramminger, J.
2003— *Neulateinische Wortliste. Ein Wörterbuch des Lateinischen von Petrarca bis 1700*. <www.neulatein.de>.

Ramminger, J.
2014 “Neo-Latin: Character and Development.” In: Ford, Ph. *et al.* (eds.), *Brill’s Encyclopaedia of the Neo-Latin World* (*Renaissance Society of America Texts & Studies Series* 3), Leiden, 21-36.

Rowland, I.
2016 “Annius of Viterbo and the Beginning of Etruscan Studies.” In: Carpino, A. & Bell, S. (eds.), *A Companion to the Etruscans*, Chichester, 433-445.

Rybicki, J.
2012 “The great mystery of the (almost) invisible translator: stylometry in translation.” In: Oakes, M. et Ji, M. (eds.), *Quantitative Methods in Corpus-Based Translation Studies*, Amsterdam.

Rybicki, J.
2013 “Stylometric translator attribution: Do translators leave lexical traces.” In: Piotrowski, T. & Grabowski, Ł. (eds.), *The translator and the computer*, Wrocław, 193-204.

Rybicki, J. & Eder, M.
2011 “Deeper Delta across genres and languages: do we really need the most frequent words?,” *Literary and Linguistic Computing* 26/3, 315-321.

Sapkota, U. *et al.*
2015 “Not All Character N-grams Are Created Equal: A Study in Authorship Attribution.” In: *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, s. l., 93-102.

Stamatatos, E.
2009 “A survey of modern authorship attribution methods,” *Journal of the American Society for Information Science and Technology* 60/3, 538-556.

Stamatatos, E.
2013 "On the Robustness of Authorship Attribution Based on Character N-gram Features," *Journal of Law and Policy* 21/2, 421-439.

Tunberg, T. O.
1988 "The Latinity of Lorenzo Valla's Gesta Ferdinandi regis Aragonum," *Humanistica Lovaniensia* 37, 30-78.

Valla, L.
1978 *Antidotum Primum. La prima apologia contro Poggio Bracciolini*, [Edizione critica con introduzione e note a c. di Ari Wesseling], Assen Amsterdam.