Universiteit Leiden
The Netherlands

Computer Science

Finding Seneca in Seneca

using Text Mining techniques
to investigate authorship of
Hercules Oetaeus and *Octavia*.

Luuk Nolden

Supervisors:
Suzan Verberne & Antje Wessels

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

19/07/2019

Abstract

This thesis tries to answer to what extent computational stylistics methods are able to find Seneca minor in the tragedies attributed to him. A variety of methods will be discussed and applied to these ten tragedies. The main focus is to indicate if *Hercules Oetaeus* and *Octavia* are misattributions or not.

Contents

1 Introduction

From the first century CE to this very date a seemingly simple question stands unanswered: *who wrote the only extant Latin tragedies?* The earliest manuscripts containing these plays call them the *Marci Lucii Annei Senecae Tragoediae*,[1] but no writer of that name is known. This leaves modern scholars to believe that the tragedies are the work of Lucius Annaeus Seneca minor, the well-known philosopher, politician and tutor of emperor Nero.[2] Oddly enough this Seneca minor never mentions or cites these plays in his other works, nor do any contemporary or later authors when writing about him.[Kohn, 2003, 271-274] The tragedies do however fit quite well in the corpus of the Stoic philosopher[Levèvre, 1985] and are, also because no other real alternative author exists,[Kohn, 2003, 279] generally attributed to this well-known Seneca the Younger. The oldest known manuscripts of the plays, dating from the late eleventh century, though not associated with Seneca's philosophical works, handed down the following ten tragedies: *Agamemnon, Hercules Furens, Hercules Oetaeus, Medea, Oedipus, Octavia, Phaedra, Phoenissae, Thyestes* and *Troades*. This set of plays in itself poses another problem, as *Hercules Oetaeus* and *Octavia* are generally considered misattributions and therefore not from Seneca's hand.[Kohn, 2003, 274]

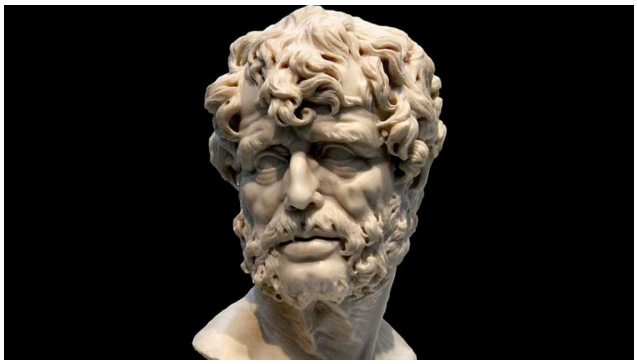

Figure 1: Possible bust of Seneca the Younger.

The scholarly debate almost unanimously agrees that *Hercules Oetaeus* is misattributed and written by another author (although a close imitation).[Boyle, 2013, 290] Problems with the piece are its length, which is more than twice as long as the other plays, and its many small differences in style, which *suggest a fundamentally different approach to playwriting.*[Marshall, 2013, 40] More importantly however, *Hercules Oetaeus* contains a great variety of passages from the other tragedies which have been lifted out of context, reworked and inserted into the text.[Tarrant, 2017, 97] Lastly, the tragedy ends with a rather strong Stoic theme. Although Stoicism is present in the other pieces,[Birt, 1911] such an outspoken passage is not found in the other plays.[Christopher, 2015, 255] It must be noted however that the suspicion exists that the first seven hundred lines of the play might have been written by Seneca,[Edert, 1909, 62-68] whilst the rest of the piece was hastily written and finished later by a rather clumsy imitator.[Walde, 1991, 1][Axelson, 1967, 92]

[1] Regarding the manuscript tradition, take a look at R.J. Tarrant (1976), *Seneca: Agamemnon*, L.D. Reynolds, ed. (1983), *Texts and Transmission* 357-381, R. H. Philp (1968), *Manuscript Tradition of Seneca's Tragedies*, CQ n.s. 18 150-179 and O. Zwierlein (1986), *L. Annaei Senecae Tragoediae* v-xi.

[2] He was also simply known as Seneca minor or Seneca the Younger, to differentiate him from this father Seneca maior (or the Elder), a well known rhetorician.

The authenticity of *Octavia* is more difficult to decide, its problem lying in the manuscripts. Two separate branches of manuscript traditions containing all Seneca's tragedies are handed down, the interpolated A-version, which is of lesser quality, and the *codex Etruscus* (E), which, though more complete, does not contain *Octavia*.[Sluiter, 1949, 11-12] Another problem for its authenticity is the date of writing, which is thought to be past Nero's death in 68CE, three years after Seneca's own forced suicide. The tragedy seems to refer to events from the final year of Nero's life, which would be impossible for Seneca to know.[Ferri, 2003, 5-9] These forecasting passages are however rather vague and it would not have been hard for Seneca to predict Nero's imminent demise.[Sluiter, 1949, 13-18] Regarding style and atmosphere, *Octavia* does not raise any questions of authorship, completely supporting the idea of a Senecan tragedy. G. Carlsson even remarks how the piece would be an absolute masterpiece of imitation if it would not have been written by Seneca.[Carlsson, 1926, 51] The text is however an outlier when compared to the other nine tragedies, as it is the only one written in the Roman style (the others are written in the style of Greek tragedies)[Sluiter, 1949, 2-5] and Seneca himself acts as a character in the play.[Sluiter, 1949, 5-11] In short, although advocates for a misattribution outnumber those adverse to this idea, no real conclusion has been reached yet.

This thesis uses computational stylistics, which is a methodology in stylistic studies implementing computer-assisted corpus analysis. It focuses on the question whether a computer can help finding stylistic features in a text and attribute an author to it.[Deneire, 2018, 246] The focus of this research will thus lie in finding out whether this digital form of authorship attribution is able to support the claim that *Hercules Oetaeus* is indeed misattributed and to help the debate in deciding whether to attribute *Octavia* to Seneca or not. The main question of this paper is as follows: *to what extent are computational stylistics methods able to find Seneca in the tragedies attributed to him?*

My method will be as follows: on the (rather bold) premise that the other eight tragedies are indeed from the same author (whether this is Seneca minor or not does not really affect the procedure), *Hercules Oetaeus* and *Octavia* are compared to these eight pieces. This comparison is done by several statistics, which are explained and embedded in literature in Section 2. These methods are then applied to the dataset (the ten tragedies) and produce their own data. This is described in Section 3. The data produced is interpreted via tables, graphs and other figures in Section 4. If the alleged misattributed pieces differ greatly from the others, explanations for these deviations and conclusions in the direction of misattribution or not will be given where possible. At the end of this paper, the robustness of the used methods on this dataset will be discussed in Section 5 and a final conclusion will be drawn in Section 6, along with ideas for further research.

This research is in my opinion useful for Computer Science as well as for Classics. For the former because the dataset is small (65.000 words) and there exists a strong presumption of the outcome: if our methods result in the same preliminary conclusions, it might help root the robustness of this field of study in the Classical world, as it proves the usability of Text Mining and authorship attribution on small sets of data. Furthermore, it might help to show Computer Science the type of research interesting for Classics, which is often very specific and extremely focused (unlike previous research in this area by Bamman (2012) and Mimno (2012)[Bamman & Smith, 2012][Mimno, 2012]). For the latter because many Classical and Neo-Latin texts and fragments still need proper authorship attribution. Moreover, much research would benefit hugely from the possibilities and time efficiency provided by Computer Science and Digital Humanities, which is — in my opinion — to this date, although growing in popularity, underutilised and not part of the standard curriculum. This thesis will hopefully show the possibilities and accessibility of Text Mining and its methods for small

datasets to the Classical World.

To summarize:

- Ten tragedies attributed to Seneca minor are known to exist.
- In general, eight of these are considered to be from the same author, be it Seneca minor or not. These shall be called authentic, genuine or Senecan and are in alphabetical order:
  - Agamemnon (ag)
  - Hercules Furens (herf)
  - Medea (med)
  - Oedipus (oed)
  - Phaedra (phae)
  - Phoenissae (phoe)
  - Thyestes (thy)
  - Troades (tro)
- Two are considered misattributions. These are, as mentioned:
  - Hercules Oetaeus (hero)
  - Octavia (oct)

2 Background

2.1 Related Work

This thesis takes inspiration from Kestemont, Moens, & Deploige (2015) *Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux*, [Kestemont *et al.* , 2014] in which the oeuvre of Hildegard of Bingen (1098-1179), an influential female author of the Middle Ages, was investigated. She dictated her texts to scribes, which were then written in Latin. As she did not master this language completely, these scribes were allowed to correct her where necessary. In the article, the *Visio ad Guibertum missa* and *Visio de Sancto Martino* are examined, which are both attributed to Hildegard. Using a corpus with letters from Hildegard and two of her main scribes, Guibert and Bernard of Clairvaux, and a number of stylometric techniques, it is convincingly shown that the texts attributed to Hildegard are created by extensive collaboration, displaying remarkable differences within her works. It is even shown that these texts attributed to Hildegard must have been reworked in such a way that style-oriented computational procedures attribute the texts to Guibert.

Another important inspiration is the paper from Tom Deneire (2018) *Filelfo, Cicero and epistolary style: a computational study*,[Deneire, 2018] in which he compared the works of the humanist writer Filelfo (1398-1481) to those of Cicero (106-43BCE). The later humanist Erasmus (1466-1536) claimed in his *Ciceronianus* that Filelfo's letters earned praise because they where successful imitations of Cicero's letters and that his orations did not because the opposite was true. Using the same tools used in the aforementioned Kestemont paper, Deneire shows that the difference between the letters and the orations of the two is not big, thus refuting Erasmus' claim.

2.2 Metrics in computational stylistics

All the used methods are described below. Their application and analysis are printed in Section 4.

2.2.1 Average Word-length

In 1851 Augustus de Morgan, well known as the mathematician of the laws of set theory bearing his name, proposed the average word-length of a text as an indicator of authorship. His ideas were tested by T.C. Mendenhall, an American geophysicist at Ohio State University, culminating in the two papers *The Characteristic Curves of Composition* (1887) and *A Mechanical Solution to a Literary Problem* (1901).[Grieve, 2002, 8-9] In these papers Mendenhall concluded that word-length can not be considered to be a good indicator of authorship (a conclusion supported by C.B. Williams (1970) and M.W.A. Smith (1983)), *for it is possible, although not probable, that two (different) writers might show identical characteristic curves.*[Grieve, 2002, 10] However, as an entirely different average word-length might still raise suspicion, this method has been included in this paper.

2.2.2 Syllable Distribution

To improve upon the method of average word-length, Forsyth, Holmes and Tse (1999) suggested to measure the percentage of one-, two-, three-, four-, five-syllable words.[Forsyth *et al.* , 1999] This statistic was used to see if either Cicero or Sigonio had written the sixteenth century edition of the

Consolatio. They concluded that this text resembled Sigonio the most from all the investigated Neo-Latin writers and was, more importantly, very unlike Cicero.

2.2.3 Average Sentence-length

In response to Mendenhall's previously mentioned publication, H.T. Eddy (1887)[Eddy, 1887] offered alternative textual measurements for attributing authorship, average sentence-length being the most prominent one.[Grieve, 2002, 12] His methods were improved by W.B. Smith in 1888[Mascol, 1888] and L. Sherman in 1893.[Sherman, 1893] The first real result of this idea was published in G.U. Yule's 1939 paper *On Sentence-length as a Statistical Characteristic of Style in Prose.*[Yule, 1939] In this paper, Yule showed that the sentence-length distribution of *De Imitatione Christi*, an anonymous religious treatise from 1418, resembled the text characteristics of Thomas à Kempis more than it did those of Jean Charlier de Gerson.[3] This conclusion was and is in accordance with the scholarly debate.[Grieve, 2002, 14] Gerson's texts have a mean 23 words/sentence and a median of 19 words/sentence, whereas Kempis' corpus has a mean of 18 words/sentence and a median of 15 words/sentence. *De Imitatione Christi* has a mean of 16 words/sentence and a median of 14 words/sentence.[Grieve, 2002, 14] As Kempis' results are more similar to the text in question, it might be a good indication that he is more likely to be the author than Gerson.

W.C. Wake (1957)[Wake, 1957] showed a remarkable similarity in sentence-length for a selection of works by Plato, Aristotle and Xenophon, all of which were fairly stable.[Grieve, 2002, 15] From this he concluded that Plato's disputed *Seventh Letter shows an agreement [in sentence-length] so good that statistical proofs of the insignificance of the differences are superfluous.*[Wake, 1957, 243] The disputed *Ethics* ascribed to Aristotle however did not show the same agreement, from which Wake concluded that the text was written by multiple authors. Despite these apparent successes, damning critique for this method was provided by Herdan's *Type Token Mathematics*,[Herdan, 1960] in which he argued that the within-author sentence-length variation was far too big to be used as a general indicator of authorship. A similar length does not say as much either, as F. Mosteller showed by comparing Alexander Hamilton's and James Madison's writings, which, although being written by two different authors, are extremely similar (a mean of 34.55 and 34.59 respectively).[Grieve, 2002, 17] Lastly, in the case of Classical texts, punctuation marks are placed by modern editors and are therefore not guaranteed to correspond with the author's original suggestion.[Wake, 1957, 334]

2.2.4 Hapax Legomena

In 1986 A.Q. Morton published a paper in which he debated the use of once-occurring words, the so-called hapax legomena, as discriminatory features.[Morton, 1986] He argued that the position of these words in a sentence would enable to distinguish one writer from another. As the hapax legomena occurred remarkably infrequent at the beginning of a sentence, but at an enhanced rate at its ending, Morton thought that a difference in this variation should be characteristic of an author.[Holmes, 1985, 111] M.W.A. Smith replied in 1987,[Smith, 1987] stating that Morton's work was *not statistically sound*, with no convincing evidence provided that the position of these once-occurring words could assist in attributing authorship.

However, disregarding the position of such a word in a sentence, it might still be useful for authorship attribution. My idea is as follows: as a hapax legomenon is a word only occurring

[3] These two writers were the main candidate authors.

once in a corpus of texts, such a word seems not very characteristic of an author. A bad imitator would therefore not use a word like this, as it might not be regarded typical for the author he is imitating. A low percentage of hapax legomena could therefore be an indication of a rather clumsy imitator. Unlike the statistical and quantitative approach of the other methods, this would be a more qualitative approach to the texts, as it focuses more on the active and concious decisions of an author.

2.2.5 Type/Token Ratio

In 1985 D. Holmes wrote that *the basic assumption is that the writer has available a certain stock of words, some of which he/she may favour more than others. If we sample a text produced by that writer, we might expect the extent of his/her vocabulary to be reflected in the sample frequency profile.*[Holmes, 1998, 334] This sample frequency, denoting the vocabulary richness of a text, might then be used for comparative purposes between texts. One might immediately notice the flaw with this method, as the words used in a text depend far more on the subject in question than on its author.[Grieve, 2002, 21] Although every word used should be in the vocabulary of an author, different subjects require different sections of one's vocabulary (which might in turn differ in richness). However, since the texts compared in this paper are of the same genre from presumably the same author, this method should not suffer from this inherent flaw.

This statistic, called the Type/Token Ratio (TTR), is retrieved by dividing the number of word-types V (different words) by the number of word-tokens N (their occurrence).[Grieve, 2002, 22] A score of zero means that all the words in a text are the same, a score of one means that all the words are unique. As the value of V depends on the number of word-types in a text, it is greatly dependent on the length of that text: caution should therefore be taken when comparing texts of different lengths.

2.2.6 Zipmethod

A more abstract approach to authorship attribution is the Zipmethod. As seen in the previous paragraphs, the problem of this attribution is *a typical classification problem which depends on discriminant features to represent the style of an author.*[Oliveira *et al.* , 2013, 100] The focus often lies with features such as vocabulary richness and lexical repetition, but these are, as observed by Madigan *et al.*, strongly dependent on the length of the text. It is therefore difficult to apply these features reliably on texts with deviating lengths.[Madigan *et al.* , 2005] To circumvent this, compression models have been suggested for authorship attribution.[Kukushkina *et al.* , 2003][Malyutov, 2005]

The idea behind this is that compression algorithms (as used for example by WinRAR and xarchiver) build a dictionary of the files they process.[Oliveira *et al.* , 2013, 100] An archive (a compressed set of files) with such a dictionary will be created of our authentic tragedies, whereafter for each individual tragedy (including the tragedies in question) a new archive will be created consisting of the original Senecan archive and the unknown document (the testing sample). Roughly speaking, *two documents/archives are deemed close if we can significantly compress one given the information in the other, the idea being that if two pieces are more similar, we can more succinctly describe one given the other.*[Cilibrasi & Vitanyi, 2005, 1524] In short, there is a candidate author Seneca (with his eight tragedies) and a text sample of unknown authorship *(Octavia)*: the lower the difference in size of the combined archive of candidate author and text sample, the more similar

the two are.

The main advantages of this method are its ease of use, its independence of parameters and its ability to judge a text in its entirety in an abstract way (treating the text as a black box).[Oliveira *et al.* , 2013, 101] The compression method used in this paper is bzip2, as this method was found to be the most reliable and yielding the best results by Oliveira (2013).[Oliveira *et al.* , 2013, 103] The used method is called the Normalised Compression Distance (NCD), which gives a score greater than zero to denote how similar a text is to the original archive.[Cilibrasi & Vitanyi, 2005] The score is normalised to cope with texts of different lengths.

2.2.7 Word Frequency

The aforementioned W.B. Smith used the frequency of function words[4] in his 1888 stylometry study of the *Pauline Epistles*.[Mascol, 1888] Comparing their frequencies he produced his so-called *curve of style*, which he compared to the corresponding curves for *Ephesians, Philippians* and *Colossians*, after which he concluded that these epistles had not been written by Saint Paul.[Grieve, 2002, 32] For this study, Smith used function words exclusively, since they *constitute the fibrous tissue of speech, spreading through and permeating the entire organism of discourse (...) they are the current co-ordinates in the equation of style, determining by their mutual relations the law of form for the writers thoughts.*[Mascol, 1888, 454] His idea was proven to work well by the statisticians Mosteller and Walles (1963, 1964, 1984) when they successfully used function words to investigate the origins of *The Federalist Papers*, which where known to be written by either James Madison, Alexander Hamilton or John Jay.[Grieve, 2002, 34][5] Their conclusions are however adverse to the research of A. Ellegård, who in his 1962 books *A Statistical Method for Determining Authorship* and *Who was Junius* argued that *some words, phrases, and turns of expression are felt as vaguely "typical" of a particular author, while other words and turns of expression are such as he "would never have used".*[Ellegård, 1962a, 12] As Yule did before him,[Yule, 1944] Ellegård limited his research to content words[6] instead of function words, justifying his choice by stating that *the words most frequently used in the language—articles, prepositions, conjunctions, and pronouns, as well as the commonest verbs, nouns, adjectives and adverbs—are necessarily about equally frequent in texts, whoever the author.*[Ellegård, 1962a, 15-16] As their frequency is the same, one should exclude words as prepositions and pronouns, as they have little significance when regarding style.[Yule, 1944, 21]

It must be noted however that despite the popularity of word frequency as an indicator of authorship, the measurement as a whole has been subject to considerable criticism.[Grieve, 2002, 41] The discussion mostly boils down to the question if we can assume that a used word has a universal stable rate in an author's work. In the case of content words it clearly has not (as these depend on the subject of the text), but the case of function words is still up for debate.[7]

For this thesis the one hundred most frequent words are used. These are mostly function words (for these are the most frequent words in Latin), which work best for authorship attribution, as they are, in my opinion, used in a mostly unconscious and intuitive way. And since the

[4]Function words are words with little lexical meaning, signalling the structural relationships that words have to one another. They include, amongst others, prepositions, pronouns and conjunctions.

[5]Interestingly they concluded that average sentence-length was not a good measurement for stylometry.

[6]Content words are words denoting objects, consisting of nouns, lexical verbs and adjectives.

[7]See F.J. Damereau (1975) for criticisms to this technique and Holmes (1985) for a defence.

mind of every author is unique, the use of these function words should differ from writer to writer.[van Dalen-Oskam & van Zundert, 2005, 212] However, as the tragedies are of the same genre and presumably the same author, content words are not actively removed from this list.

The results are interpreted by two statistics. The first statistic is the Kendall rank correlation coefficient, better known as Kendall's Tau. It is used to measure the ordinal association between two measured quantities. This is done by using rank correlation, which denotes the similarity of the orderings of the data.

The second statistic to interpret these results is the Wilcoxon signed rank test. This statistic takes the scores for each pair of words (e.g. the word *quis*) of an individual tragedy and the genuine tragedies and computes the (absolute) difference. The differences for all these pairs are added to a total sum. The higher this score, the more an individual tragedy deviates from the genuine ones.

2.2.8 Principal Components Analysis

The last statistic used in this paper also uses word frequency and greatly relies on the work done by M. Eder and his colleagues J. Rybicki and M. Kestemont. Their scripts for stylometry for statistical computing have been written in the *R* software environment and have been made freely available on their website.[8] In contrast to the previous chapter in which the output was numerical and statistics were used to interpret the results, the word frequency clustering as used by Eder generates visualisations using Principal Components Analysis (PCA), which are to be interpreted by the end-user.

According to Grieve, PCA *is a statistical procedure for transforming the values of a set of measurements into a smaller set of new uncorrelated measurements or principal components (PC), which are then ordered so that the first two or three PCs account for most of the variation in the dataset.*[Grieve, 2002, 36] In other words, a PCA allows patterns in the dataset (in our case frequent words) to be more easily observed by reducing the number of dimensions that describe the data.[Grieve, 2002, 36] Clustering can hereafter be used to group texts that show similarities. This method has been successfully used by J.F. Burrows (1988)[Burrows & Hassall, 1988] in his paper *Anna Boleyn and the Authenticity of Fielding's Feminine Narrative*, as well as in M.W.A. Smith's 1991, 1992 and 1993 papers for, amongst others, the attribution of authorship of the anonymous 1607 play *The Revenger's Tragedy*.[9] As mentioned in Section 2.1, the word frequency clustering metrics created by Eder work extremely well on Latin datasets, as shown by recent research done by T. Deneire in his paper *Filelfo, Cicero and Epistolary Style: a Computational Study*.

The PCA graphs in this paper will be generated with the help of Eder's tools for word frequency clustering. According to his manual, the idea is as follows: first, a list of most frequent words (MFW) of the entire corpus is created, which includes content and function words.[10] Next, the frequencies of the MFW in the individual texts are retrieved and put into an initial matrix of words (rows) by individual texts (columns): each cell will contain a single word's frequency in a single text.[Eder *et al.* , 2017, 7] On these matrices various statistical procedures are performed. In this paper, Cluster Analysis (CA), Multidimensional Scaling (MDS), and Principal Components

[8] *The R Project for Statistical Computing* (http://www.r-project.org).

[9] For more research done with PCA, see Grieve (2002) 32-43.

[10] Although it must be noted that the list mainly consists of function words, as these often have the highest frequency.

Analysis are used.[11]

The Cluster Analysis produces a dendrogram showing hierarchical clustering of analysed texts.[Eder *et al.* , 2017, 15] This option is used to show similarity between the examined tragedies. The option for Multidimensional Scaling visualises the level of similarity of individual cases of a dataset and maps these cases in a two-dimensional plot. In addition to this, the level of similarity is also visualised using the Principal Component Analysis option. For this last procedure, the covariance matrix was chosen to create the chart, as it covers a greater percentage of the dataset when compared to the correlation matrix.[Eder *et al.* , 2017, 13]

The *Culling* parameter mentioned in these graphs refers to the automatic manipulation of the word list: its value specifies which words will be considered for the analysis: for example, a value of ten means that only words appearing in at least 10% of the given texts will be added to the word list.[Eder *et al.* , 2017, 13] This paper used a culling setting of zero, meaning that no words will be removed.

[11]For more information on these procedures, take a look at the NCSS Statistical Software Manual, Chapter 445: *Hierarchical Clustering/Dendrograms*, J. B. Kruskal and M. Wish (1978) *Multidimensional Scaling* and J. Shlens (2014) *A Tutorial on Principal Component Analysis.*

3 Data and Implementation

3.1 Source Text

The tragedies used in this paper have been retrieved from the Perseus Digital Library, from which almost all Latin and Greek texts from the Classical period can be downloaded. These texts are encoded in XML, which have been converted to plain text using the ElementTree XML package for Python.[12] These ten text files, with an average of six thousand Latin words per tragedy, have been used for all the methods in this thesis. In total, the used corpus consists of 11,658 lines made up from 63,700 words. Additionally, the works of Horace, Ovid and Virgil have been downloaded from the Latin Library[13] to be compared to Seneca and to serve as a test of robustness in some of the methods.

3.2 Methods

The average Word-length in Section 4.1 has been computed with Voyant-Tools[14] and the Natural Language Toolkit (NLTK) for Python.[15] The Syllable Distribution from Section 4.2 was calculated using the pyphen package[16] and an Italian dictionary, as no option for the Latin language exists (the used dictionary seems to work remarkably well with splitting Latin words). The average sentence-length in Section 4.3 was calculated by dividing the total number of words (defined as a continuous string of graphemes) of an individual tragedy by the total number of sentences of said tragedy (defined by a full stop, question mark or exclamation mark) using a custom-made Python script. The Hapax Legomena (Section 4.4) were retrieved using a Python dictionary and simply counting words. The Type/Token Ratio in Section 4.5 was also retrieved via Voyant-Tools. In order to use the Zipmethod (Section 4.6), xarchiver[17] was used to create a bzip2 archive, its size being calculated via the filemanager.

The data for the word frequency graphs in Section 4.7 is provided by the Perseus Digital Library (more specifically its Vocabulary Tool),[Perseus, 2007] and the *stylo*-script created by Eder, Rybicki and Kestemont. For every text Perseus provides an XML file with all the occurring words of that text. In it the weighted frequency of a word is stored, which is the approximate count of that word in the tragedy.[18] For example: (*habeo*: 84) means that the word *habeo* and its derivatives occur approximately 84 times in the text. This number is then turned into a relative frequency by dividing it by the total number of words of the tragedy and multiplying it by one hundred, effectively turning it into a percentage (for example, *habeo* constituting for 1.5% of all words in text X), which is more informative. Take for instance a look at the Greek verb pempô. In all the texts of Plutarch, the verb occurs 146 times. This is rather unimpressive when compared to Xenophon, who uses the word 350 times. However, the corpus of Plutarch consists of 107,000 words, which is a third the size of Xenophon's 312,000. The relative frequency of pempô in Plutarch is thus 13.67,

[12] Available via (https://docs.python.org/2/library/xml.etree.elementtree.html).
[13] Available via (http://www.thelatinlibrary.com/).
[14] Available via (https://voyant-tools.org/).
[15] Available via (https://www.nltk.org/).
[16] Available via (https://pyphen.org/).
[17] Available via (http://xarchiver.sourceforge.net/).
[18] Approximate, as the Vocabulary Tool itself is not flawless when assigning a word to a word stem.

compared with Xenophon's maximum of 11.21. If we want to compare these authors in a useful way, the relative frequency of words should be used, as the size of the corpora vary.[Perseus, 2007]

ag	herf	oct	thy
et	et	et	et
non	est	non	est
est	in	est	in
in	non	in	non
quid	ad	te	quid
iam	quid	per	me
te	qui	me	iam
cum	sed	quid	quis
nec	aut	qui	nec
ut	nec	si	si

Table 1: The ten most frequent words for four of the tragedies.

From the Perseus Vocabulary Tool tables like Table 1 were generated, showing the most frequent words in descending order per tragedy. As seen in Listing 1, this data together with the aforementioned percentages were inserted in the JSON format for easy computing using Python. For the Wilcoxon and Kendall's Tau statistics for example this JSON file was used to fill the two arrays needed for the tests (one being the average scores over the authentic tragedies, the other the scores per individual tragedy). As the dataset is rather small, the top hundred words are consistently used for all the methods relying on word frequency. This seems sufficient, as tragedy specific words like *Medea* and *Hercules* start appearing after these hundred words. These words are not very informative for the other tragedies and the characterisation of Seneca and are therefore omitted.

```
{
    average
        et : 2.5,
        est : 1.0,
        non : 1.0,
    ag
        et : 2.6,
        est : 1.2,
        non : 0.5,
}
```

Listing 1: Excerpt of the JSON file used.

For the Principal Component Analysis of Section 4.8, using the *stylo*-script created by Eder, Rybicki and Kestemont, again the top hundred most frequent words (MFW) have been used to

generate the graphs. This is in contrast to Deneire's 2018 paper, in which he used a MFW of 100, 300 and 500.[Deneire, 2018] This seems to me to be cherry picking one's data/graph to fit the conclusion.

All the bar charts have been created with Libreoffice Calc, the tables and excerpts with Latex and the PCA charts with the *stylo*-script.

4 Analysis and Results

4.1 Average Word-length

As seen in Table 2, the average word-length of the Senecan tragedies is the same, so according to Mendenhall, no real conclusion can be drawn from this. It is however safe to say that the tragedies are similar in this regard.

ag	herf	hero	med	oct	oed	phae	phoe	thy	tro
5	5	5	5	5	5	5	5	5	5

Table 2: Average length of a word per tragedy.

4.2 Syllable Distribution

Measuring the percentage of one-, two-, three-, four-, five-syllable words renders the result as seen in Figure 2. All texts have the approximate same distribution of different syllables per word: only *Phoenicians* prefers words with one syllable above those with three (with a difference of 0.3%). In this regard the texts are also remarkably similar.

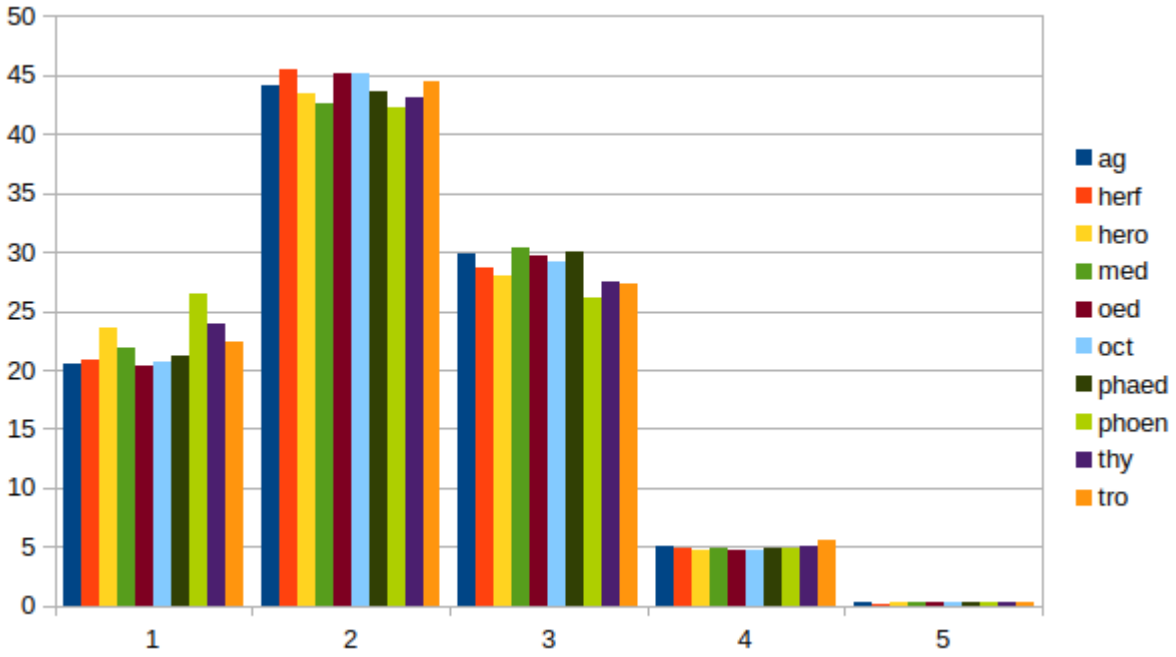

Figure 2: The percentages of the number of syllables per word over the tragedies.

4.3 Average Sentence-length

Despite the criticisms to this method, the average sentence-length for Seneca's tragedies as seen in Figure 3 is rather uniform, hovering around 26 words/sentence. As shown in orange, the disputed tragedies *Hercules Oetaeus* and *Octavia* both fall within the extremes of the genuine tragedies, so no real conclusion against their authenticity can be drawn from this graph. It is however too hasty to conclude the opposite, although the indication of similarity is striking.

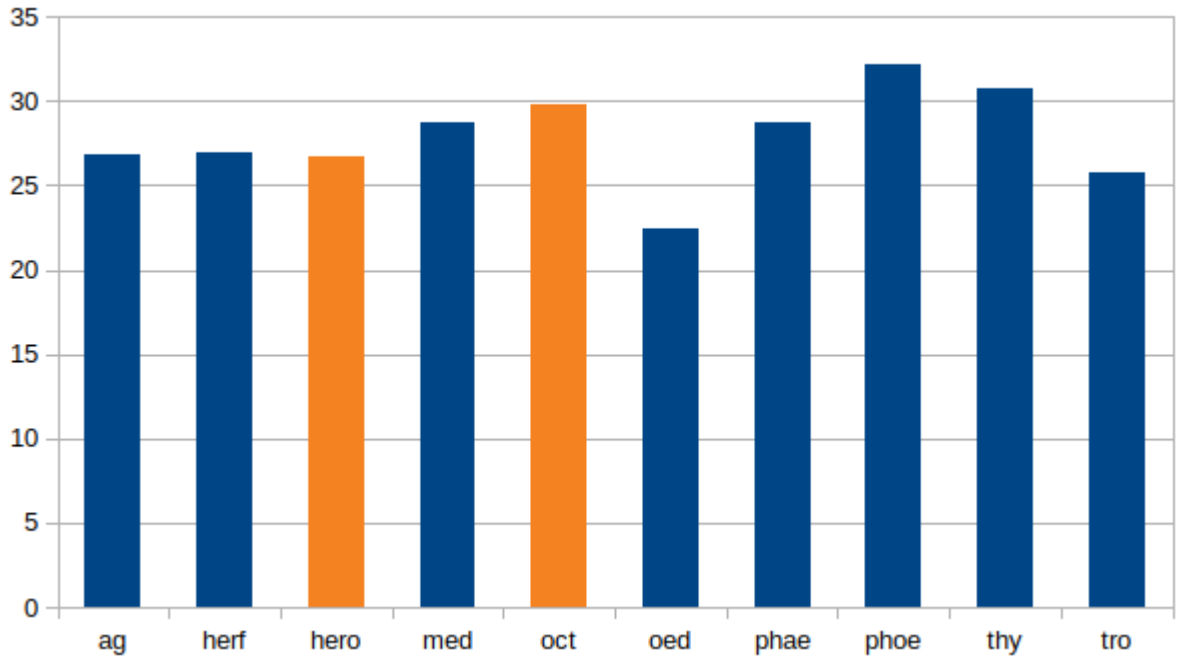

Figure 3: The average sentence-length per tragedy as found by Voyant-tools.

4.4 Hapax Legomena

As seen in Figure 4, $1,5\%$ of Seneca's words are on average a hapax. Three things are remarkable. Firstly, *Hercules Oetaeus* does not use any hapax legomena, which corresponds with the idea that a bad imitator does not use such rare words. Secondly, *Octavia* has an evidently lower usage of this word-type, which is interesting to say the least. However, and this is the third point, *Hercules Furens* does not contain any hapax either. Since this is considered an authentic piece, any real conclusion from this graph would not be justified. Of course, *einmal ist keinmal*, but drawing the conclusion that *Octavia* is unlike Seneca with regard to hapax legomena would necessarily point *Hercules Furens* in the same direction (which is not supported by literature).

4.5 Type/Token Ratio

The Type/Token Ratio (TTR) is, as described in Section 2.2.8, influenced by the length of a text. As seen in Table 3, only *Hercules Oetaeus* has a clearly deviating length, with twelve thousand

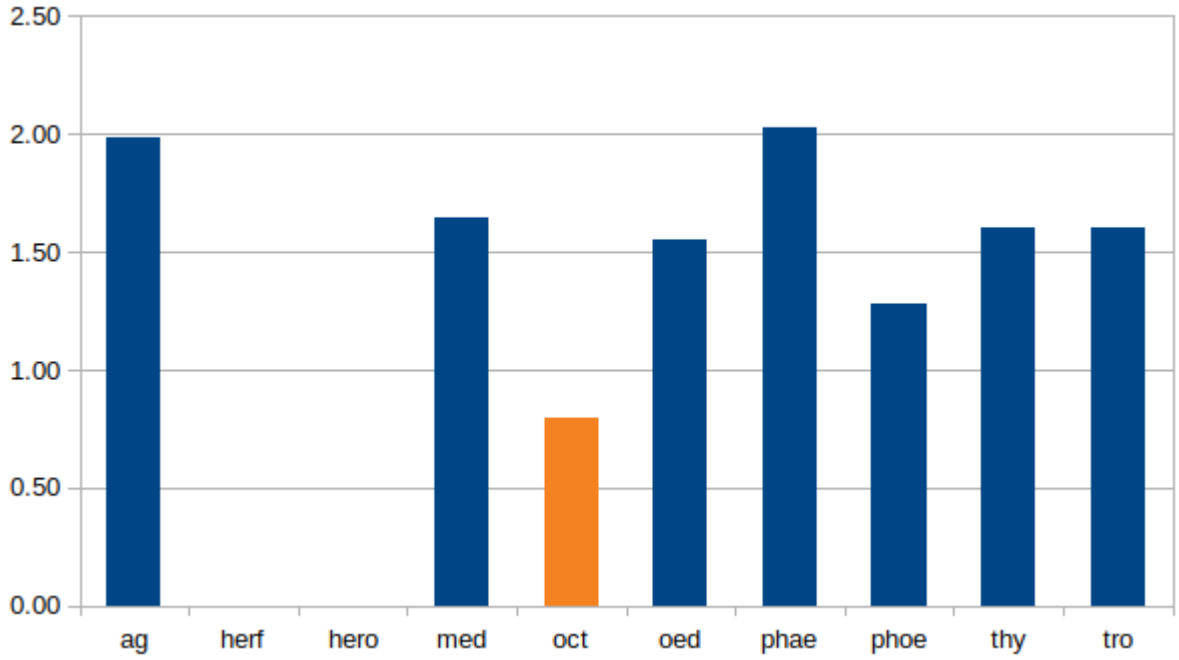

Figure 4: The percentage of words being a hapax legomenon per tragedy.

words when compared to the other texts, which hover around the six thousand words (*Phoenicians* being the exception with 4066 words).

ag	herf	hero	med	oct	oed	phae	phoe	thy	tro
5398	7382	11078	5462	5036	5666	7007	4066	5990	6615

Table 3: Length in words of the ten tragedies.

As the number of word-types can be higher for larger texts, a higher score for such a text should be taken with a grain of salt: the ratio is after all calculated by dividing the number of word-types by the number of word-tokens. For a larger text, a lower or similar score denotes an even lower richness than the other (shorter) tragedies. The result of this method as shown in Figure 5 is fascinating, as it rather clearly agrees with the scholarly consensus in literature. As is visible, *Hercules Oetaeus* has a much lower score than the other tragedies, supporting the idea of a clumsy imitator.[Walde, 1991, 1] The lexical richness of this text is, according to the results, not as big as the genuine tragedies. Although *Hercules Oetaeus* seems very Senecan according to the average word/sentence-length, the TTR does not agree with this idea. One must however observe that this score can not be interpreted in the usual way. Usually, words are counted by their word stem (*tree* and *trees* are of the same word-type), which is not easy to do for Latin texts and their grammatical cases. This problem will be further explored in the next chapters, but for now a score of for example 0.5 does not mean half of the words used are from unique word stems: they are however unique in their used form.

Octavia yields an interesting result too, as it falls below all the other tragedies in score, but

does not deviate that much from the genuine tragedies. This thus seems to be the first real indication to the authenticity of both *Hercules Oetaeus* and *Octavia*: the former in the direction of a misattribution, the latter of a genuine piece.

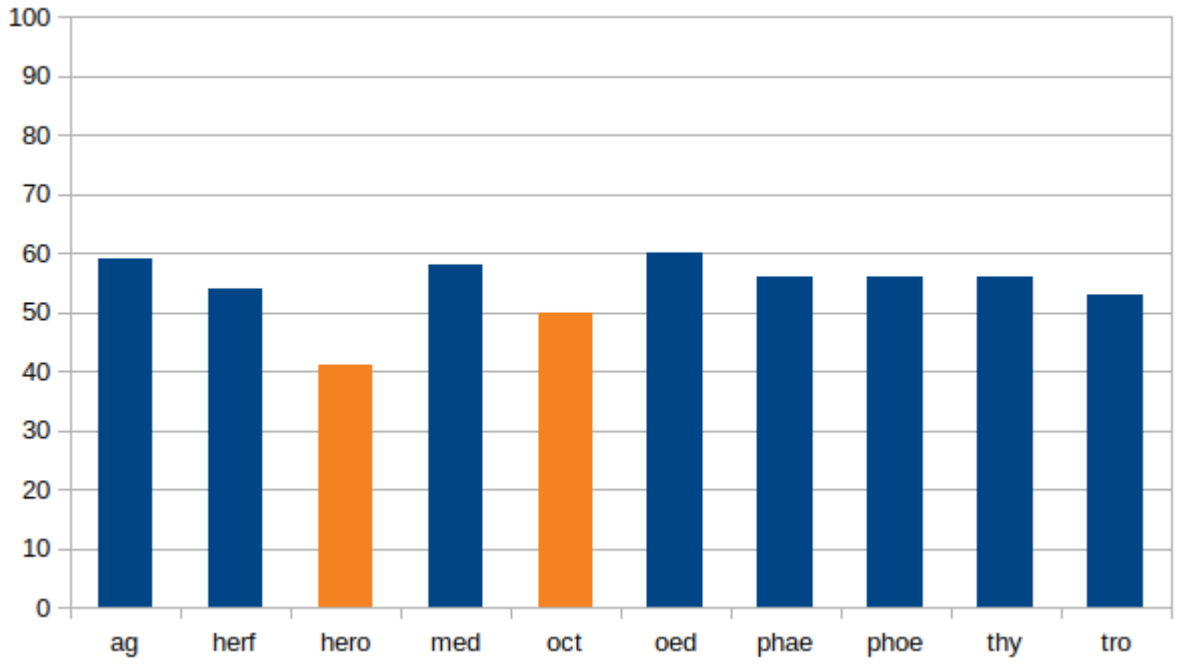

Figure 5: The Type/Token Ratio as found by Voyant-tools.

4.6 Zipmethod

Figure 6 shows that *Agamemnon* receives a score of 0.11 when compressed with an archive consisting of the other seven genuine tragedies. This number only has a meaning when compared to the other scores. For the other authentic texts, these all lie within the interval of $(0.075, 0.142)$, with *Octavia* scoring within this interval and *Hercules Oetaeus* outside it. According to this method, *Hercules Oetaeus* is the least similar when compared to the other eight Senecan tragedies, whilst *Octavia* seems to be more similar to these tragedies than most of the authentic tragedies themselves. Again a method in favour of the authenticity of *Octavia* and misattribution of *Hercules Oetaeus*.

4.7 Word Frequency

For the following two statistics, the hundred most frequent words (MFW) over the eight authentic tragedies have been computed by adding the relative frequencies per word per tragedy together and dividing each result by eight. I will call this the Senecan average. Listing 1 shows a snippet of the result in the JSON format. At the top the average frequencies are shown over the authentic tragedies. So on average, 2.5% of all words used by Seneca in his tragedies is the word *et*. Below this Senecan average the same words per tragedy are shown. So in *Agamemnon*, the word *est* has a slightly higher frequency than average.

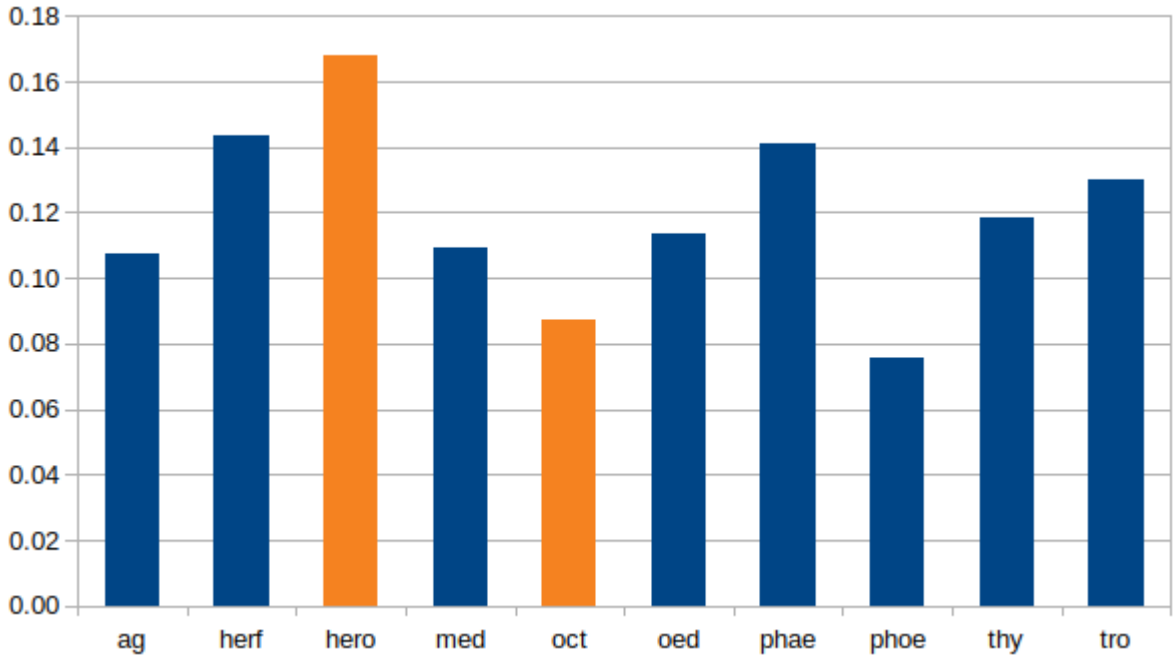

Figure 6: Results of the Zipmethod. A lower score denotes similarity.

The first statistic used on this data is Kendall's Tau. For this method, the words from Listing 1 are sorted on frequency for the Senecan average and for each individual tragedy. Next, the similarity of the Senecan average ranking of words is compared with that of an individual tragedy. Lastly, a correlation is calculated. The results are shown in Figure 7. The authentic tragedies score between 0.002 and 0.1, which indicates no correlation.[19] The tragedies in question score within this interval. No real conclusions regarding the authenticity of the pieces can be drawn from this statistic. It seems that for ranking of the hundred most frequent words, no correlation exists between the Senecan average and each individual piece.

The second statistic used is the Wilcoxon signed rank test. From Listing 1, the score for *Agamemnon* is calculated as $|2.6-2.5|$ for *et*, $|1.0-1.2|$ for *est* and $|1.0-0.5|$ for *non*. This equals to a mark of 0.8, which only has meaning when compared to the other scores. These are listed in Figure 8. The higher the score, the more deviating a tragedy is from the Senecan average. Interestingly *Octavia* scores as deviating as *Phoenicians*: since mostly function words are being compared, this might direct the latter in the direction of a misattribution, or the former in the direction of a genuine piece (as it does not deviate from the authentic *Phoenissae*). *Hercules Oetaeus* seems to score closer to the other pieces, which seems to be in contradiction with the results from the Zipmethod and Type/Token Ratio, but in accordance with the methods for sentence/word-length.

4.8 Principal Component Analysis

The first two graphs (Figure 9 and 10) show that clustering on the hundred most frequent words is indeed able to distinguish Seneca from other authors like Horace, Ovid and Virgil. Using

[19] Note that a correlation has the interval $(-1, 1)$ with 0 being no correlation.

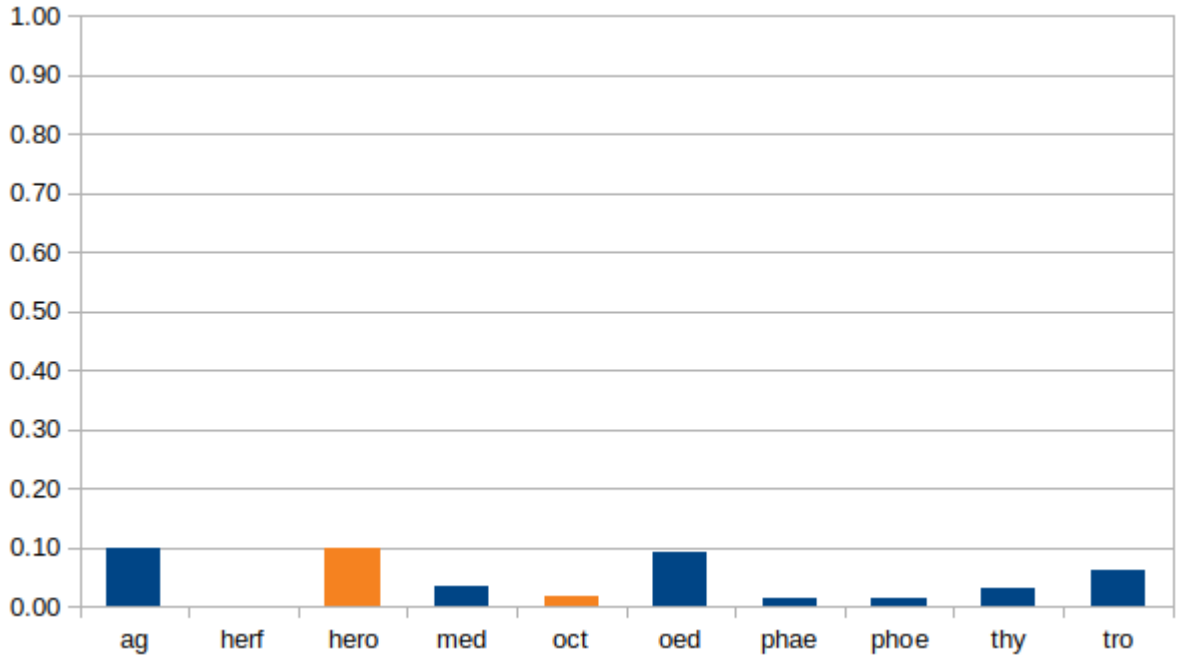

Figure 7: Correlation of each tragedy with the Senecan average according to Kendall's Tau.

Multidimensional Scaling, Figure 9 shows a clear cluster of Seneca's tragedies on the left, with the other poetry scattered around the plot. This is sensible, as these texts are love poems, epic, letters *et cetera*. Interestingly, *Phoenicians*, is an outlier in the Senecan cluster. Figure 10, using Cluster Analysis, shows a clear division between Seneca and the others, with *Octavia* being the most distinct within Seneca's tragedies. Remarkably *Phoenicians* appears again as a rather distinct tragedy in both figures.

When taking a more in-depth look at the tragedies alone (Figures 11 and 12), five things stand out.

- Both figures show *Octavia* as the most distinct tragedy. This might seem plausible, as its style is different from the others and content words are allowed in this test. However, the used MFW list does not contain a single content word. This adds a serious doubt to the authenticity of the piece.

- Both figures show *Phoenicians* as an outlier, be it less extreme than *Octavia*. If we would like to conclude from this graph that *Octavia* is a misattribution, the position of *Phoenicians* as an authentic piece should be looked into, since both show quite a deviation from the main cluster. Both have extreme positions in the graph: we can not conclude *Octavia* a misattribution because literature suggests it without pointing *Phoenicians* in the same direction.

- According to MDS *Hercules Oetaeus* is less distinct from the other pieces than according to PCA. From MDS a dissimilarity like *Phoenicians* and *Octavia* might be concluded, whilst PCA does not support this in the slightest.

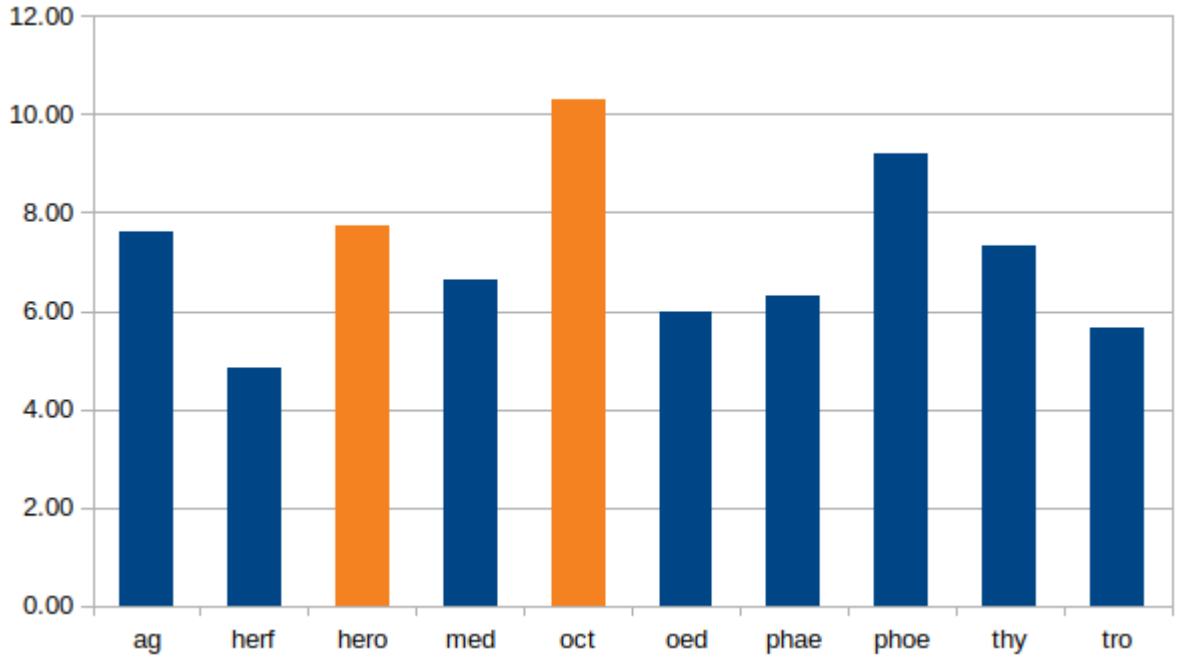

Figure 8: The Wilcoxon statistic on the hundred most frequent words, denoting the absolute difference between the Senecan average and each individual tragedy.

- *Thyestes* and (to a lesser extent) *Medea* lie apart from the main cluster. This idea is however not supported by the dendogram of Figure 10.
- When viewed from the bottom up, Figure 12 agrees with J.G. Fitch's dating of the tragedies. [Kohn, 2003, 273] The early plays would be *Phaedra*, *Agamemnon* and *Oedipus*; the middle plays *Troades*, *Hercules Furens*, and *Medea*; and the late plays *Thyestes* and *Phoenissae* (and *Octavia*, which Fitch does not consider genuine). Further research in the potential of adding dates to texts using PCA might be interesting.

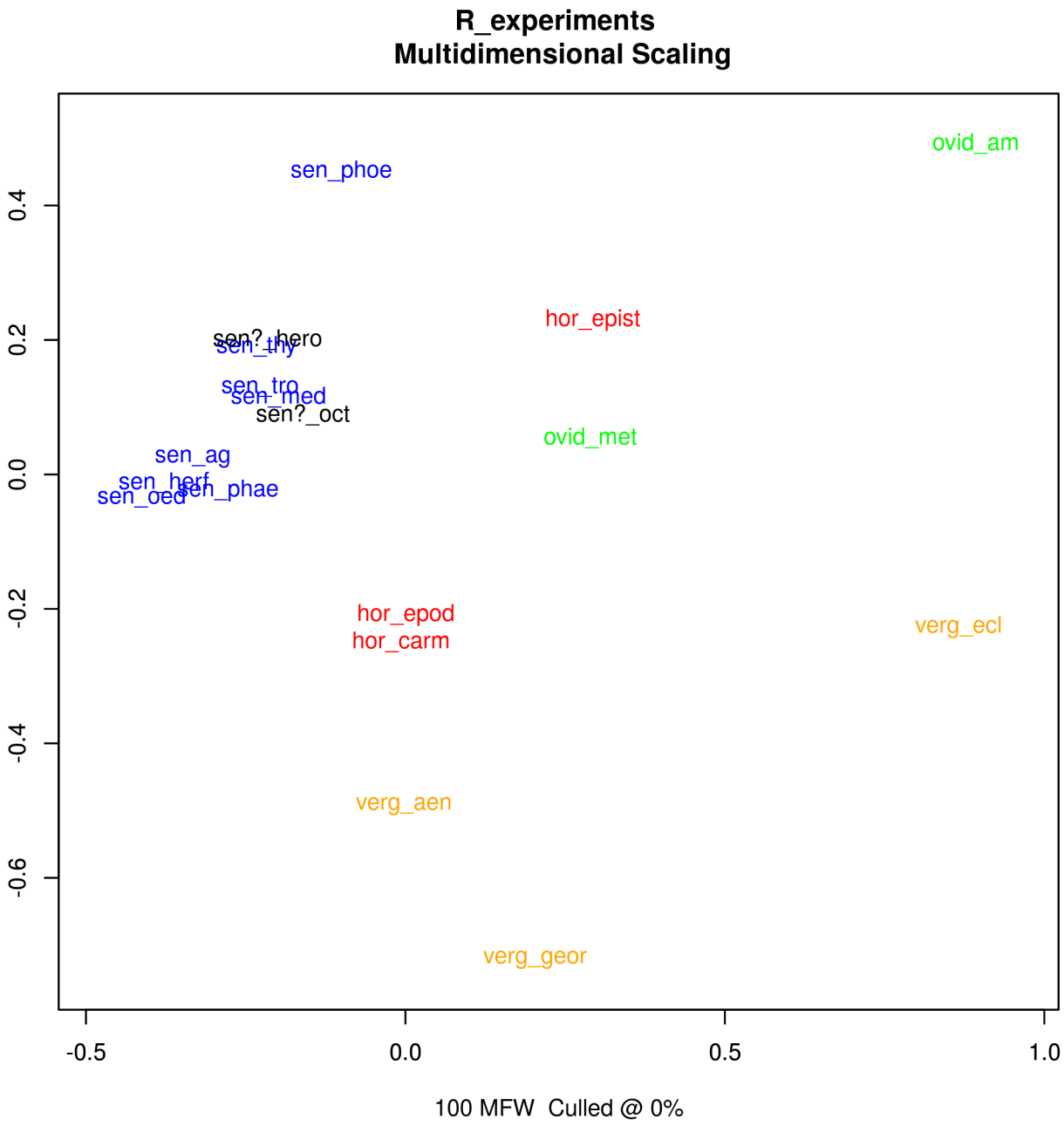

Figure 9: Multidimensional Scaling including Horace, Ovid and Virgil.

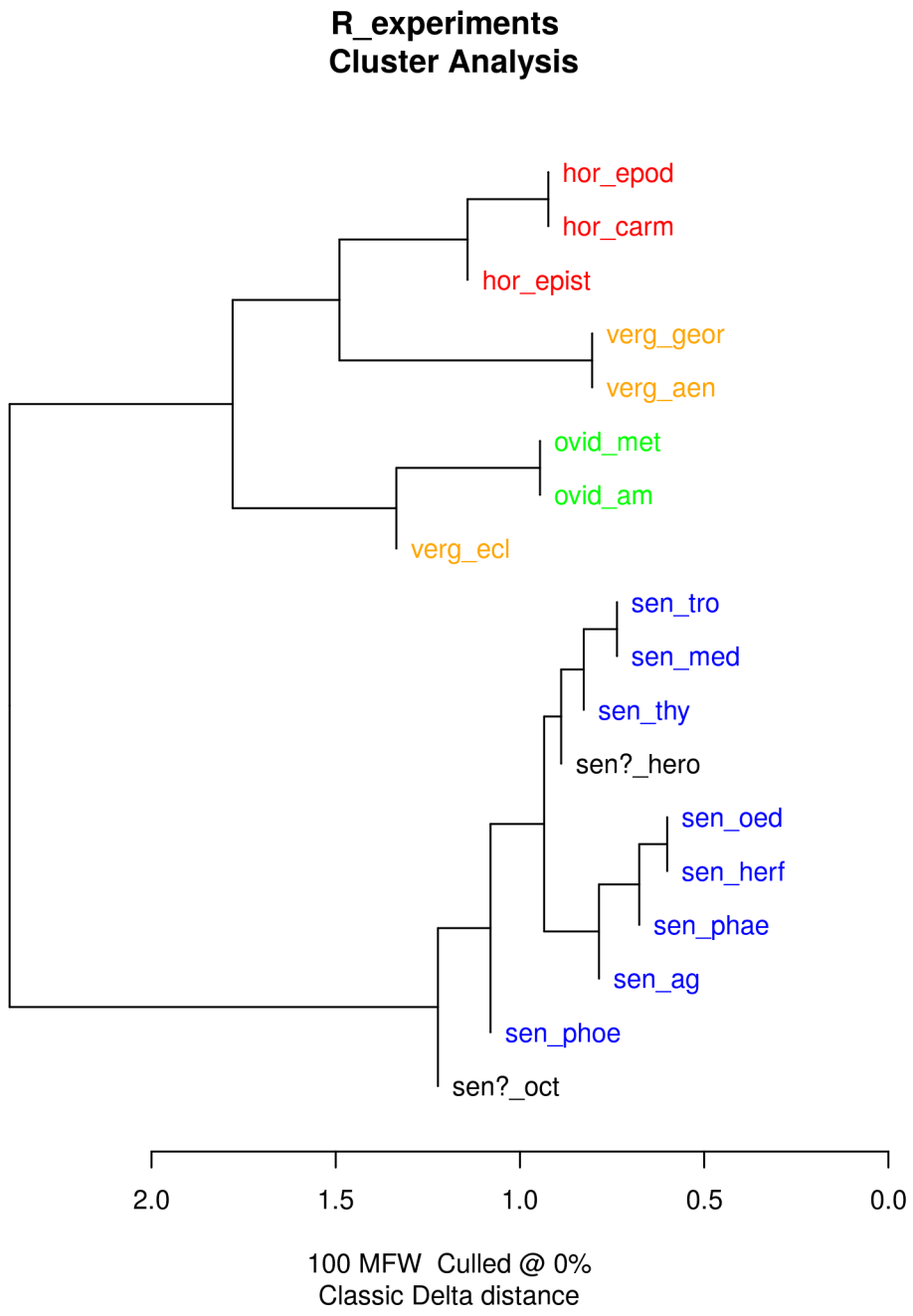

Figure 10: Cluster Analysis including Horace, Ovid and Virgil visualised by a dendogram.

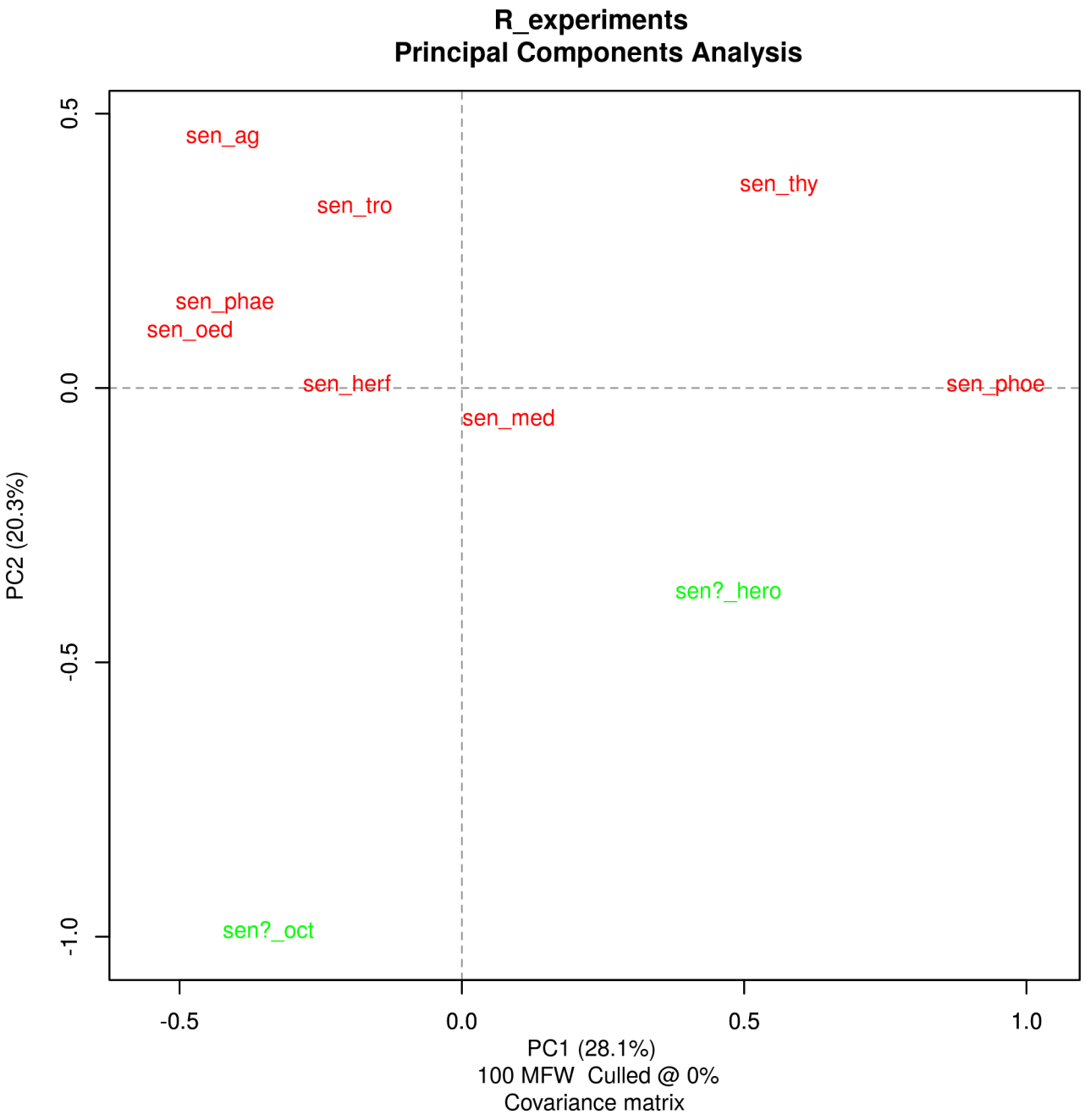

Figure 11: PCA using the covariance matrix on Seneca alone.

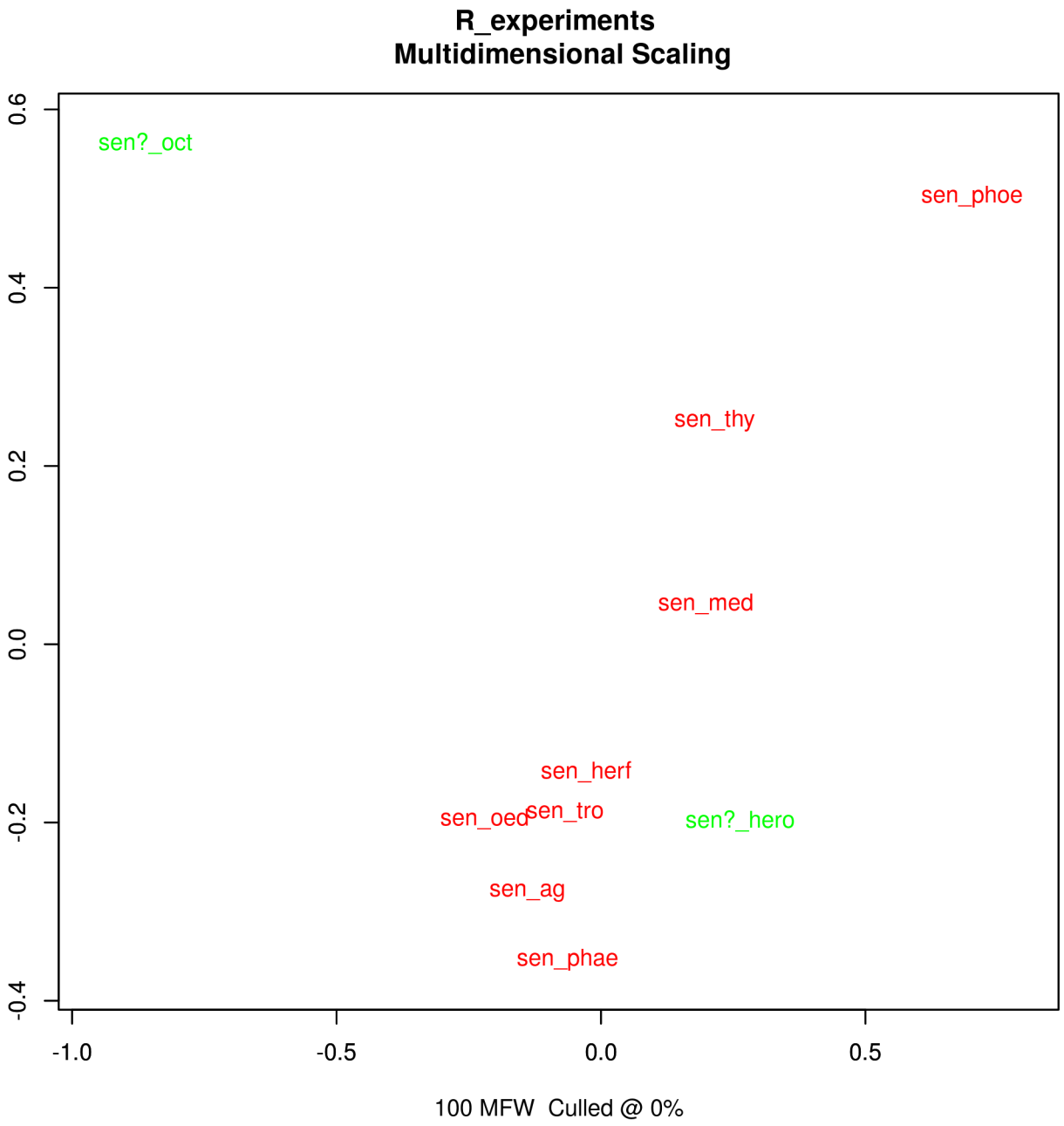

Figure 12: Multidimensional Scaling on Seneca alone.

5 Discussion

According to Grieve, *when an attribution study comes to a conclusion that contradicts its original assumption, it may be that it is the method and not the assumption that is flawed.*[Grieve, 2002, 16] In Section 4.6 we have seen that the Zipmethod indicated that *Hercules Oetaeus* was the least similar to Seneca. This is however in contradiction with the results from Section 4.7, in which the word frequency method showed that *Oetaeus* was rather similar to the Senecan average. Interestingly however, the results from PCA puts *Octavia* and *Phoenicians* as the outliers, a conclusion which is supported by Wilcoxon.

To test the robustness of the used methods, figures 9 and 10 added three different authors. This was done to test if the *stylo*-script was able to detect the difference between Seneca, Horace, Ovid and Virgil. Applying the same idea to the Zipmethod yields a rather interesting result. In Table 4 the scores of Octavia compared to the four authors is shown. Keep in mind that a lower score means, according to the Zipmethod, similarity.

Horace	Ovid	Seneca	Virgil
0.17	0.06	0.09	0.07

Table 4: Compressing *Octavia* with four different authors.

Although the differences are small, the Zipmethod indicates that *Octavia* is more similar to Ovid and Virgil than it is to Seneca. This really undermines the credibility of this method and its results as seen in Figure 6. I think it is safe to conclude that the Zipmethod as used in this paper is not very suitable for such a small dataset[20] and its results should be viewed with extreme caution. However, when viewed on its own with the Senecan corpus only, the graph does seem to show the similarity of *Octavia* with the rest and the difference of *Hercules Oetaeus* to the other pieces (which might be a wanted result).

Testing the robustness of the Wilcoxon statistic on word frequencies by adding Ovids *Methamorphoses* and Virgils *Aeneid* generates Figure 13, which shows the same tendency. Both new authors lie within the interval of the Senecan average. Virgil's *Aeneid* would be, according to this graph, similar to *Medea*.

Eder's *stylo*-script does not natively support the option to mine on all existing function words. However, a word list is generated with words to be used in the metrics. I replaced the words in this list with all the function words of the Latin language and ran the script again, which resulted in a rather interesting result. As this method is not officially supported, I did not include the results in the main part of this paper, but the outcome is too interesting not to describe in this section. According to the dendogram and the MDS (Figure 14 and 15), *Octavia* and *Hercules Oetaeus* are rather extreme outliers. The PCA using the covariance matrix agrees with this idea, but also shows *Phoenissae* and *Thyestes* as outliers (which is not at all supported by the dendogram and the MDS). Using only function words as argued in literature seems indeed to support the idea of *Octavia* and *Oetaeus* being misattributions. However, and this is important, instead of the hundred MFW, all function words — however obscure — were used, which might have skewed the result. Although it does agree with a conclusion we might seek, the method is not officially supported and

[20] The sizes are as follows: Horace (105KiB), Ovid (560KiB), Seneca (325KiB) and Virgil (450KiB).

its usage should be tested and debated, so for now only cautious preliminary conclusions might be drawn in my opinion.

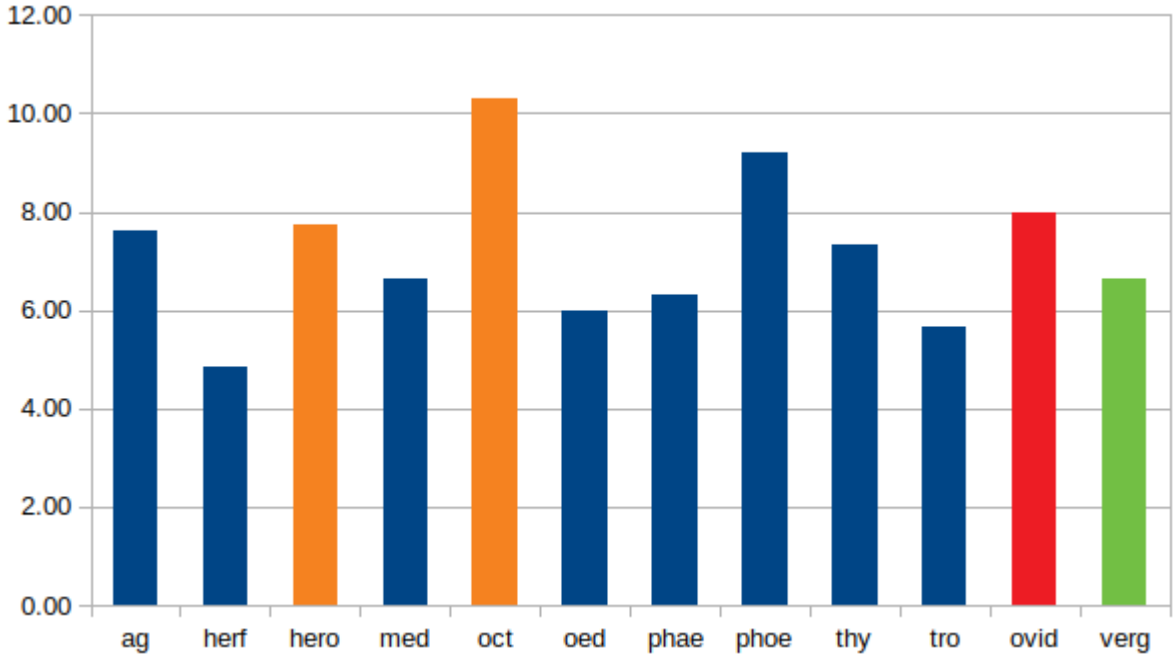

Figure 13: Wilcoxon on Seneca with Ovid *(Metamorphoses)* and Virgil *(Aeneid)* added.

5.1 Limitations

As seen in Section 2, all methods have their advantages and disadvantages and all come with their fair share of praise and criticism. In combination with the small dataset used, drawing any definite conclusions from these graphs would be absurd. However, as discussed above, some statistics seem to perform better than others (this however needs to be proven, not deduced by our desires to a certain outcome). The conclusion in Section 6 will therefore limit itself to a cautious one.

For calculating the word frequencies, both the Perseus Vocabulary Tool and the *stylo*-script have been used. Both returned similar results, though be it not the exact same. The reason seems to be the extreme difficulty for a computer to determine the word-stem of a Latin word. However, this problem only exists for content words, as function words mostly have one conjugation. So although these graphs are not completely accurate, patterns are still rather trustworthy.

In this paper the differences in text editions were not taken into account. As there are many different manuscript traditions, many subtle but important differences exist in the tragedies. Interesting would be to conduct all the statistics again on the other available Latin versions to see if these change the result.

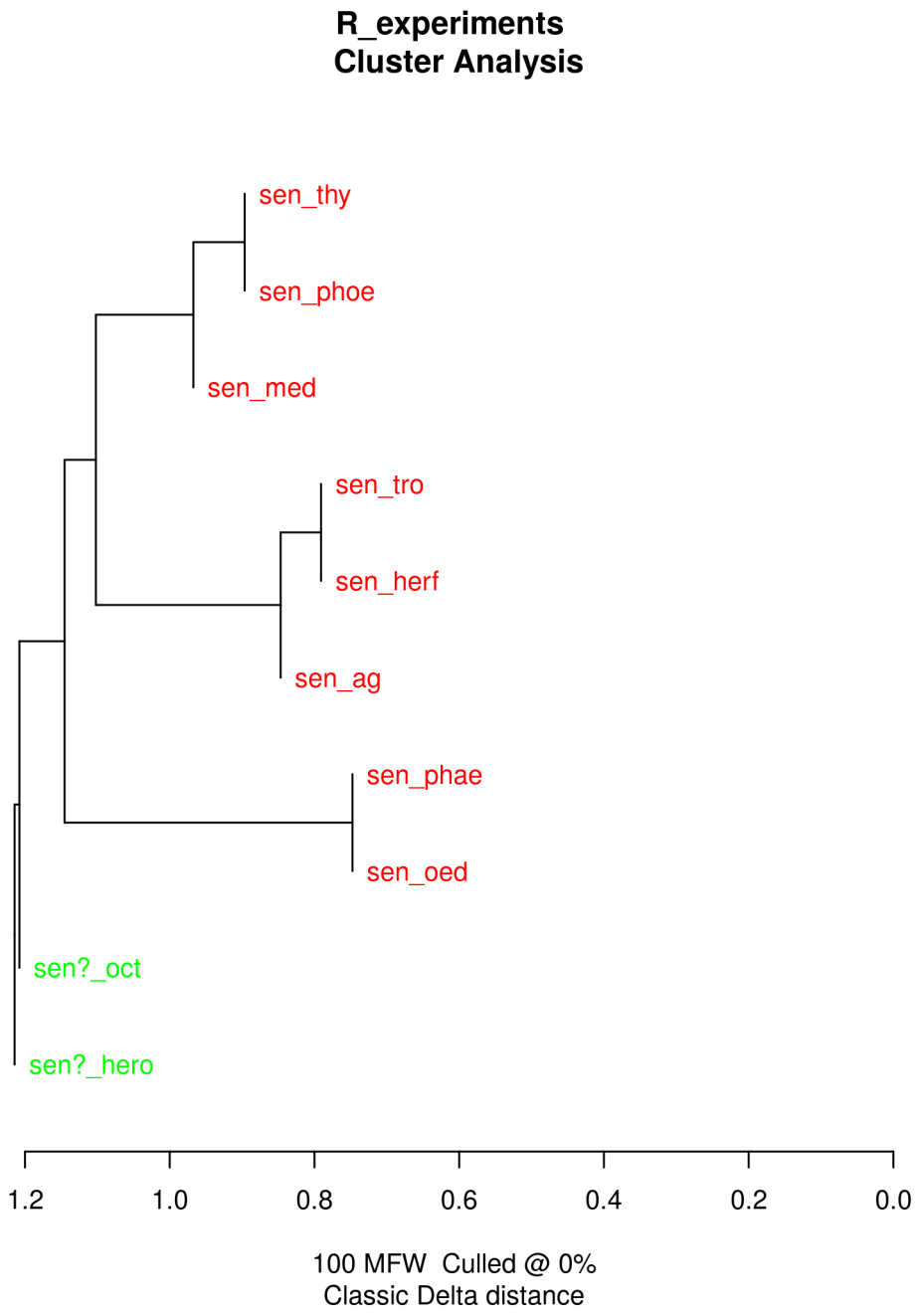

Figure 14: Dendogram on the Senecan tragedies only using all function words.

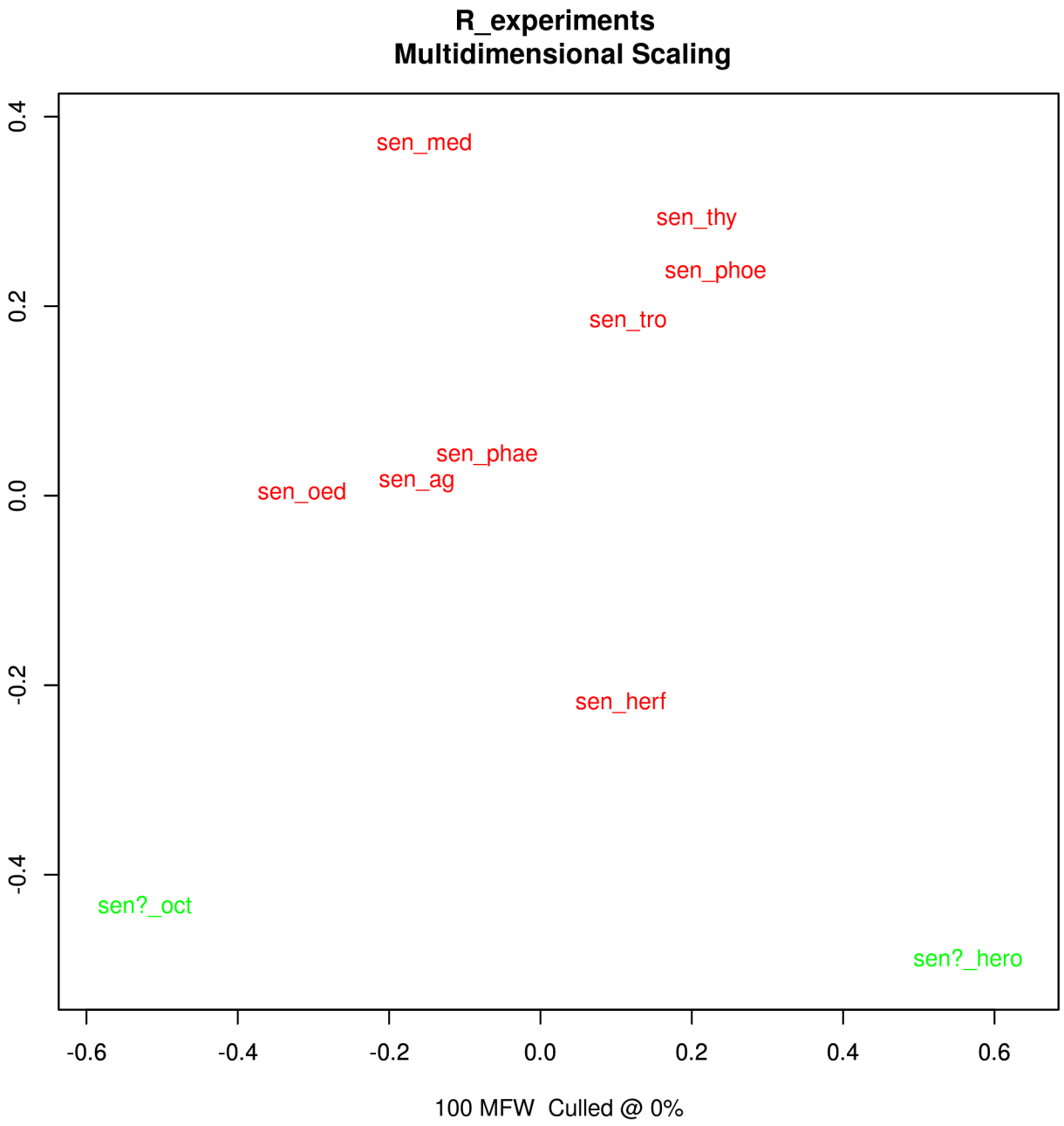

Figure 15: MDS on the Senecan tragedies only using all function words.

6 Conclusions and Further Research

Let us take another look at all the graphs, focusing mainly on *Octavia* and *Hercules Oetaeus*, but also taking note of any other peculiarities. The average word-length, syllable distribution and average sentence-length all showed a striking similarity between the ten tragedies. Dissimilarities started to show with the Type/Token Ratio, which clearly indicated a lower vocabulary richness of *Hercules Oetaeus*, with *Octavia* scoring within the margin of error. Although found not really fit for this dataset in the discussion of Section 5, the Zipmethod indicated a dissimilarity for *Oetaeus* too. Seemingly against the conclusion of the TTR, it showed *Octavia* to be very similar to the genuine pieces. Wilcoxon showed *Phoenissae* as an outlier together with *Octavia*, however, when put in the same graph with Ovid and Virgil, the trustworthiness of this method was shaken. Interestingly though, this position of *Phoenissae* and *Octavia* is partly acknowledged by the Multidimensional Scaling and the Cluster Analysis (when applied on Seneca, Horace, Ovid and Virgil), as it puts the former (again) as an outlier, the latter as clear part of the cluster. When focusing the Principal Component Analysis only on the tragedies, MDS indicates the dissimilarity of *Octavia* and *Phoenissae*, whilst PCA adds *Hercules Oetaeus* to this list (a conclusion not at all supported by MDS). One must not forget however that directing *Octavia* to being a misattribution on the basis of these graphs, one must almost necessarily point *Phoenissae* in the same direction. Its position is therefore interesting and invites further research, as this tragedy as a misattribution has been discussed by A. Paul[Paul, 1953] and E. Fantham[Fantham, 1983].

It thus seems that the conclusion would be two-fold. Firstly, the quantitative statistics used show indeed a great similarity between the ten tragedies and seem to function fine on such a small Latin dataset. Secondly, *Octavia* and *Hercules Oetaeus* do indeed, as literature suggests, feel like strangers within this set, as they frequently (but not always!) score above or below the Senecan average (and outside its interval). However, a strong conclusion cannot be drawn, as the differences are often rather small.

A more qualitative approach as done with the Hapax Legomena, which focuses on the Latin text in a more traditional way (though aided by computational power), might be interesting for further research, as the graph clearly showed *Oetaeus* and *Octavia* as rather extreme outliers. Examples of possible methods are finding *ut* and *et* at the beginning of a verse, finding multiple words ending with *-m* and searching for vowels at the end of verses, which were often thought to be not very elegant in the art of writing.

For now, based on the results of the TTR, I would conclude that *Hercules Oetaeus* is indeed misattributed, as this score is extremely low compared to the other pieces. However, when approached in the abstract and quantitative way as done in this paper, the other methods show that the play is very similar to Seneca's tragedies (which is logical for an imitation). Regarding *Octavia*, further research is necessary to really draw any useful conclusions. Although most of the methods advocate a misattribution, its position in the PCA, MDS and CA graphs seems to go hand in hand with the fate of *Phoenissae*. To really conclude anything about *Octavia*, *Phoenissae* should be studied as well. If the latter could indeed be a misattribution, *Octavia* would surely be one according to the Principal Component Analysis. If it remains to be seen as an authentic tragedy however, the claim to *Octavia* being a misattribution would be weaker, as we then could not rely on the PCA, MDS and CA graphs.

So in short, is Text Mining able to find Seneca in Seneca? It does seem so, but (as always) a more in-depth approach is needed to more conclusively answer this question. The alleged misattributions

seem to hide themselves extremely well behind their Senecan masks: a heavier sledgehammer is needed to convince these pieces to take these off.

Further Research

Six things would be interesting for further research. Firstly, different text editions should be used to see if any clear changes occur in the methods applied. This would show the importance or triviality of these different editions. Secondly, the usefulness of the *stylo*-script when exclusively applied on function words should be debated and studied, as literature seems to agree that these words work the best when attributing authorship. If my method as discussed in Section 5 is indeed sound, evidence against *Oetaeus* and *Octavia* as genuine pieces would strengthen. Thirdly, as the used methods seem to work rather well on this dataset (no extraordinary outliers have been seen yet), it would be interesting to see if it could attribute these tragedies, which are, as mentioned in the introduction, said to be written by *Marcus Lucius Anneus Seneca*, to the Stoic rhetor and philosopher *Lucius Annaeus Seneca*. This could help solve the duality of Seneca, but would also be difficult as the texts differ in topic and style. Fourthly, any confirmations of the suspicion that the first seven hundred lines of *Hercules Oetaeus* are genuine might change the results of the methods used in this thesis. Treating the two halves as two separate tragedies might therefor be useful in further research if clear divisions are found. Fifthly, the recurring outlier position of *Phoenissae* is extremely interesting. Not only is the piece unfinished in two locations, it is also the shortest tragedy by quite a margin. Further research might support the idea of a misattribution as discussed in A. Paul[Paul, 1953] and E. Fantham.[Fantham, 1983] The conclusion about *Phoenissae* might help the conclusion for *Octavia*. Lastly, the ability of adding dates to the texts using PCA as possibly seen by Figure 12 would be extremely useful and should see some more exploration, as dating a text is often very difficult.

As Seneca would say: *Longam viam ingressus es.*

References

[Axelson, 1967] Axelson, B. 1967. *Korruptelenkult, Studien zur Textkritik der unechten Seneca-Tragödie Hercules Oetaeus.*

[Bamman & Smith, 2012] Bamman, D., & Smith, D. 2012. Extracting Two Thousand Years of Latin from a Million Book Library. *Journal on Computing and Cultural Heritage*, **5**, 1–13.

[Birt, 1911] Birt, Th. 1911. Was hat Seneca mit seinen Tragödien gewollt? *NJbb*, **27**, 336–364.

[Boyle, 2013] Boyle, A.J. 2013. *Tragic Seneca: An Essay in the Theatrical Tradition.*

[Burrows & Hassall, 1988] Burrows, J.F., & Hassall, A.J. 1988. Anna Boleyn and the Authenticity of Fielding's Feminine Narratives. *Eighteenth-Century Studies*, **21**, 427–453.

[Carlsson, 1926] Carlsson, G. 1926. *Die Überlieferung der Seneca-Tragödien.*

[Christopher, 2015] Christopher, S. 2015. Roman Tragedy and Philosophy. *Pages 238–259 of:* Harrison, G.W.M. (ed), *Brill's Companion to Roman Tragedy.* Leiden, Boston: Brill.

[Cilibrasi & Vitanyi, 2005] Cilibrasi, R., & Vitanyi, P. 2005. Clustering by compression. *IEEE Trans. Inf. Theory*, **51**, 1523–1545.

[Deneire, 2018] Deneire, T. 2018. Filelfo, Cicero and epistolary style : a computational study. *Brill's studies in intellectual history*, **289**, 239–270.

[Eddy, 1887] Eddy, H.T. 1887. The Characteristic Curves of Composition. *Science*, **March 25**, 297.

[Eder *et al.* , 2017] Eder, M., Rybicki, J., & Kestemont, M. 2017. *'Stylo': a package for stylometric analysis.*

[Edert, 1909] Edert, O. 1909. *Uber Senecas Herakles und den Herakles auf dem Oetaen.*

[Ellegård, 1962a] Ellegård, A. 1962a. *A Statistical Method for Determining Authorship: 1769-1772.*

[Fantham, 1983] Fantham, E. 1983. Nihil iam iura naturae valent: incest and fratricide in Seneca's Phoenissae. *Ramus*, **12**, 61–76.

[Ferri, 2003] Ferri, R. 2003. *Octavia: A Play Attributed to Seneca.*

[Forsyth *et al.* , 1999] Forsyth, R.S., Holmes, D., & Tse, E. 1999. Cicero, Sigonio and Burrows: Investigating the Authenticity of the Consolatio. *Literary and Linguistic Computing*, **14**, 375–400.

[Grieve, 2002] Grieve, J.W. 2002. Quantitative Authorship Attribution: A History and an Evaluation of Techniques.

[Herdan, 1960] Herdan, G. 1960. *Type Token Mathematics.*

[Holmes, 1985] Holmes, D. 1985. The Analysis of Literary Style - A Review. *The Journal of the Royal Statistical Society A*, **148**, 328–341.

[Holmes, 1998] Holmes, D. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, **13**, 111–117.

[Kestemont *et al.* , 2014] Kestemont, M., Moens, S., & Deploige, J. 2014. Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. *DSH*, **30**, 199–224.

[Kohn, 2003] Kohn, T.D. 2003. Who Wrote Seneca's Plays? *The Classical World*, **96(3)**, 271–280.

[Kukushkina *et al.* , 2003] Kukushkina, O.V., Polikarpov, A.A., & Khmelev, D.V. 2003. Using literal and grammatical statistics for authorship attribution. *Probl. Inf. Trans.*, **37**, 172–184.

[Levèvre, 1985] Levèvre, E. 1985. Die philosophische Bedeutung der Seneca-Tragödie am Beispiel des 'Thyestes'. *Pages 1263–1283 of:* H., Temporini (ed), *Aufstieg und Niedergang der römischen Welt: Geschichte und Kultur Roms im Spiegel der neueren Forschung*. Berlin: Walter de Gruyter.

[Madigan *et al.* , 2005] Madigan, D., Genkin, A., Lewis, D.D., Argamon, S., Fradkin, D., & Ye, L. 2005. Author Identification on the Large Scale. *Joint Annual Meeting of the Interface and the Classification Society of North America*, **50**, 3250–3264.

[Malyutov, 2005] Malyutov, M. 2005. Authorship attribution of texts: a review, Electron. *Notes Discrete Math*, **21**, 353–357.

[Marshall, 2013] Marshall, C. W. 2013. The works of Seneca the Younger and their Dates. *Pages 33–44 of:* Heil, A., & Damschen, G. (eds), *Brill's Companion to Seneca: Philosopher and Dramatist*. Leiden, Boston: Brill.

[Mascol, 1888] Mascol, C. (a.k.a. W.B. Smith). 1888. Curves of Pauline and Pseudo-Pauline Style III. *Unitarian Review*, **30**, 452–460 and 539–546.

[Mimno, 2012] Mimno, D. 2012. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage*, **5**, 3:1–3:19.

[Morton, 1986] Morton, A.Q. 1986. Once. A test of authorship based on words which are not repeated in the sample. *Literary and Linguistic Computing*, **1**, 1–8.

[Oliveira *et al.* , 2013] Oliveira, W., Justino, E., & L.S., Oliveira. 2013. Comparing compression models for authorship attribution. *Forensic Science International*, **228**, 100–104.

[Paul, 1953] Paul, A. 1953. *Untersuchungen zur Eigenart von Senecas Phoenissen*.

[Perseus, 2007] Perseus, Digital Library. 2007. *Perseus Vocabulary Tool*.

[Sherman, 1893] Sherman, L.A. 1893. *Analytics of Literature*.

[Sluiter, 1949] Sluiter, Th. H. 1949. *Octavia Fabula Praetexta*.

[Smith, 1987] Smith, M.W.A. 1987. Hapax legomena in prescribed positions: An investigation of recent proposals to resolve problems of authorship. *Literary and Linguistic Computing*, **2**, 145–152.

[Tarrant, 2017] Tarrant, R.J. 2017. Custode rerum Caesare: Horatian Civic Engagement and the Senecan Tragic Chorus. *Pages 93–112 of:* Stöckinger, M.C., Winter, K., & Zanker, A.T. (eds), *Horace and Seneca: Interactions, Intertexts, Interpretations.* Berlin, Boston: Walter de Gruyter.

[van Dalen-Oskam & van Zundert, 2005] van Dalen-Oskam, K., & van Zundert, J. 2005. De list van het lexicon. Auteursonderscheiding met behulp van computer-ondersteunde woordenschatanalyse. *Nederlandse Letterkunde.*, **10**, 212–231.

[Wake, 1957] Wake, W.C. 1957. Sentence-length Distributions of Greek authors. *Journal of the Royal Statistical Society A*, **120**, 331–346.

[Walde, 1991] Walde, C. 1991. *Herculeus Labor, Studien zum pseudosenecanischen Hercules Oetaeus.*

[Yule, 1939] Yule, G.U. 1939. On Sentence-length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship. *Biometrika*, **31**, 356–361.

[Yule, 1944] Yule, G.U. 1944. *The Statistical Study of Literary Vocabulary.*