Clustering the Corpus of Seneca: A Lexical-Based Approach

Gabriele Cantaluppi and Marco Passarotti

Abstract

We present a lexical-based investigation into the corpus of the *opera omnia* of Seneca. By applying a number of statistical techniques to textual data we aim to automatically collect similar texts into closely related groups. We demonstrate that our objective and unsupervised method is able to distinguish the texts by work and genre.

Keywords

Hierarchical Clustering Analysis • Contribution Biplots • Principal Component Analysis • Latin • Seneca

1 Introduction

We present a lexical-based investigation into the corpus of the *opera omnia* of Seneca. By applying a number of statistical techniques to textual data, we aim to automatically organize the texts in such a way that those works that share a relevant amount of lexical items are considered to be very similar to each other and get automatically collected into closely related groups.

G. Cantaluppi (✉)
Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore,
Milano, Italy
e-mail: gabriele.cantaluppi@unicatt.it

M. Passarotti
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione,
Università Cattolica del Sacro Cuore, Milano, Italy
e-mail: marco.passarotti@unicatt.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical
and Applied Statistics, DOI 10.1007/10104_2014_6, Published online: 12 November 2014

In order to detail the lexical similarities and differences between the texts, we apply a technique that is able to highlight the words that mostly characterize one or more texts in comparison to the others.

The paper is organized as follows. Section 2 presents the data that we used. Section 3 details our method for clustering the data and performing principal component analysis. Section 4 shows and evaluates the results. Section 5 reports a number of conclusions and introduces our future work.

2 Data

Lucius Anneus Seneca (4 BC–65 AD) was a Roman Stoic philosopher, statesman and gramatis. He is considered to be among the most important authors of the Classical era of Latin literature. His tragedies (most of which are based on Greek original texts) are the only complete Latin tragedies extant. The corpus of the opera omnia of Seneca is quite diverse in terms of both literary genres featured and topics addressed. This motivates its clustering analysis, aimed to collect together those texts that feature a similar lexicon, checking if the results are consistent with differences in literary genre and topics.

The corpus featuring the *opera omnia* of Seneca is taken from the lexicon of the Stoics provided by [15]. The corpus comprises 23 works, among which are eight tragedies, ten dialogues and the full text of *Apocolocyntosis*, *Epistulae morales*, *Naturales quaestiones*, *De clementia* and *De beneficiis* (divided into seven books). Two tragedies of disputed attribution (*Hercules Oetaeus* and *Octavia*) are provided as well. The size of the corpus is approximately 364,000 words. All texts come from authoritative editions. For more details, see [15], XV–XVI. The corpus is fully lemmatized.

3 Method

We applied two statistical techniques to textual data, namely clustering and principal component analysis.

All the experiments were performed with the R statistical software [14]. In particular, we used the “tm” package to build and analyze the document-term matrices that are employed for clustering [4, 5]. Distance and similarity measures provided by the package “proxy” were used as well [12].

3.1 Clustering

Clustering methods can be applied to several different kinds of data, among which are textual data, whose “objects” are occurrences of words in texts. As far as word sense disambiguation is concerned, clustering lies on the theoretical assumption stated by Harris’ Distributional Hypothesis, according to which words that are used in similar contexts tend to have the same or related meanings [9]. This basic assumption is well summarised by the famous quotation of Firth [6]:

You shall know a word by the company it keeps.

In this work, we apply hierarchical agglomerative clustering in order to compute and graphically present similarity/dissimilarity between texts. As we deal with texts instead of occurrences of words, this led us to slightly modify the two basic theoretical assumptions mentioned above. Thus, here we assume that

1. texts that feature a similar (distribution of) lexicon tend to address the same or related topics (Harris-revised);
2. you shall know a text by the words it keeps (Firth-revised).

These two assumptions are reflected in our clustering method, which compares the texts by computing their distance in terms of similarity as follows:

Data Cleaning. We remove punctuations and function words from input data. All characters were translated to lower case. In particular, we remove all (both coordinative and subordinative) conjunctions, prepositions, pronouns and those adverbs that cannot be reduced to another lemma (like *diu*, *nimis* and *semper*);

Hierarchical Agglomerative Clustering Analysis: Distance. Clustering analysis is run on document-term matrices by using the cosine distance $\mathrm{d}(i,i') = 1 - \cos\{(x_{i1}, x_{i2}, \ldots, x_{ik}), (x_{i'1}, x_{i'2}, \ldots, x_{i'k})\}$. The arguments of the cosine function in the preceding relationship are two rows, i and i', in a document-term matrix; x_{ij} and $x_{i'j}$ provide the number of occurrences of word j $(j = 1, \ldots, k)$ in the two texts corresponding to rows i and i' (profiles).

Zero distance between two documents holds when two documents with the same profile are concerned (i.e. they have the same relative conditional distributions of terms). In the opposite case, if two texts do not share any word, the corresponding profiles have distance 1;

Hierarchical Agglomerative Clustering Analysis: Clustering. We run a complete linkage agglomeration method. While building clusters by agglomeration, at each stage the distance (similarity) between clusters is determined by the distance (similarity) between the two elements, one from each cluster, that are most distant. Thus, complete linkage ensures that all items in a cluster are within some maximum distance (or minimum similarity) of each other.

Roughly speaking, according to our clustering method, works that share a high number of lemmas with similar distribution are considered to have a high degree of similarity and, thus, fall into the same or related clusters.

3.2 Principal Component Analysis

While clustering computes and represents the degree of similarity/dissimilarity between texts by clusters, it does not inform about which features distinguish one text from the other. These features are those properties that make two texts similar or dissimilar to each other.

As our method is highly lexical-based, the features that we consider are words (either lemmas or forms). In order to know which words distinguish one or more texts from the others, we apply the principal component analysis technique.

Principal component analysis is a method used to retrieve a structure built according to one or more latent dimensions. This structure can be defined by using different features: in our case, the features are words, which are used as bag-of-words representations of texts. Such representations of texts get mapped into a vector space that is assumed to reflect the latent dimension structure.

We follow the Principal Component Analysis (PCA) presentation (described e.g. by [11]) and produce contribution biplots that graphically represent a vector space [8, p. 67]. Starting from an $I \times J$ term-document matrix $\mathbf{Y}$ (whose values were previously standardized by column, in order to overcome the size differences between texts), a reduction of the column (document) space can be achieved by using principal component analysis and considering dimensions which relate texts that show high correlation in their term distributions.

A singular value decomposition (SVD) of $\mathbf{Y}/(IJ)^{1/2}$ is then performed

$$\mathbf{S} = \mathbf{Y}/(IJ)^{1/2} = \mathbf{U}\mathbf{D}_\beta\mathbf{V}'$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices containing respectively the left and the right singular vectors and $\mathbf{D}_\beta$ is a diagonal matrix containing the singular values in decreasing order.

The SVD allows the calculation of coordinates $\mathbf{U}$ for terms and $\mathbf{G} = J^{1/2}\mathbf{V}\mathbf{D}_\beta$ for documents. By considering the first two columns of $\mathbf{U}$ and $\mathbf{G}$, we have the coordinates with respect to the first two principal components.

The squares of the elements in $\mathbf{D}_\beta$ divided by their total inform about the amount of variance explained by the principal components. By considering the squared values of the coordinates of terms we obtain their contribution to principal axes.

4 Results and Evaluation

The results on Seneca's works are reported by a genre-based order: first the dialogues, then the tragedies and, finally, the *opera omnia*.

4.1 Dialogues

Figure 1 presents the clustering plot for the ten dialogues of Seneca.

According to agglomerative hierarchical clustering, each text starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy by an always lower degree of similarity. Clustering ends once all the texts are collected into one common cluster, in this case showing that the dialogues of Seneca are dissimilar at the height of 0.20 (i.e. similar at 0.80).

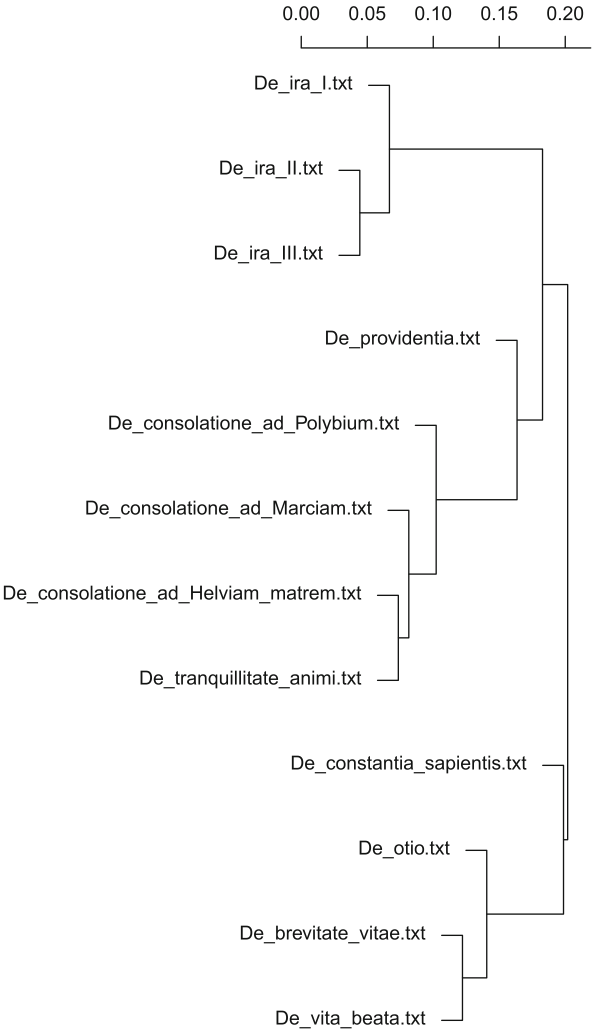

Fig. 1 Clustering the dialogues

For instance, the three books of *De ira* (which are clustered together) are dissimilar from *De providentia*, from the three *Consolationes* and from *De tranquillitate animi* at the height of 0.18 (i.e. similar at 0.82), while they are dissimilar from each other at the height of 0.07 (i.e. similar at 0.93). Among the three books of *De ira*, the second and the third are closer to each other than to the first one.

In Fig. 1 we can see that the three books of *De ira* are clustered apart from the other dialogues. Principal component analysis is able to answer the question about what makes *De ira* different from the other dialogues. As our method is lexical-based, this question concerns the words (in this case, the lemmas) that distinguish *De ira* from the other dialogues.

Figure 2 is a contribution biplot that presents the results of the principal component analysis performed on the term-document matrix of the dialogues of Seneca. In particular, the biplot represents the rows and the columns of the term-document matrix through a graph whose axes are the two first principal components, as we observed that these are able to explain over the 90 % of the total variance among texts.[1]

The first principal component gets graphically represented on the horizontal axis of the contribution biplot and it is able to explain alone most of the variance (0.876). As all the dialogues polarize in the same direction (the rightside of the biplot), the first principal component describes a dimension that is common to all the texts involved.

The second principal component is reported on the vertical axis of the biplot and it explains the 0.026 of the variance among texts. This component describes a dimension that is able to detail what mostly characterize one or more texts in comparison to the others.

For instance, the verb *sum* is placed right in the center of the vector (approximately at height 0.0 on the vertical axis). This means that *sum* is a kind of a "neuter" lemma, which is common to all the texts and does not characterize any of them in

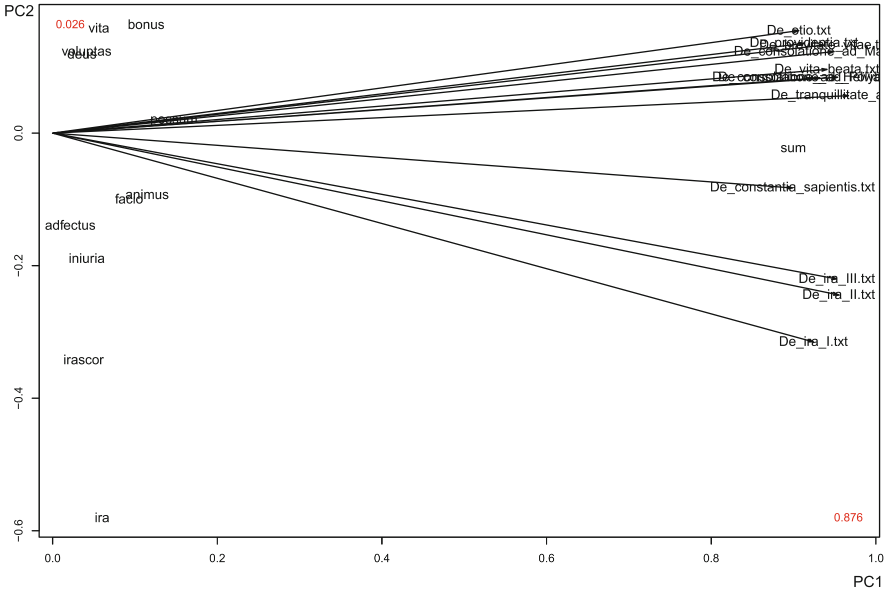

Fig. 2 Principal component analysis of the dialogues

[1] In more detail, the first two principal components explain the 0.902 of the variance, this proportion resulting from the sum of the explaining power of each of the two components (respectively, 0.876 and 0.026).

comparison to the others. Instead, the lemmas *iniuria*, *ira* and *irascor* are moved from the center and characterize the three books of *De ira*, which are all set apart from the other dialogues in the biplot.

Although the second principal component explains just the 0.026 of the total variance among texts, it is still able to report meaningful differences, which allow to recognize the specific lexical features that distinguish *De ira* from the other dialogues.

4.2 Tragedies

Figure 3 reports the clustering plot for the eight tragedies of Seneca plus the two ones of disputed attribution.

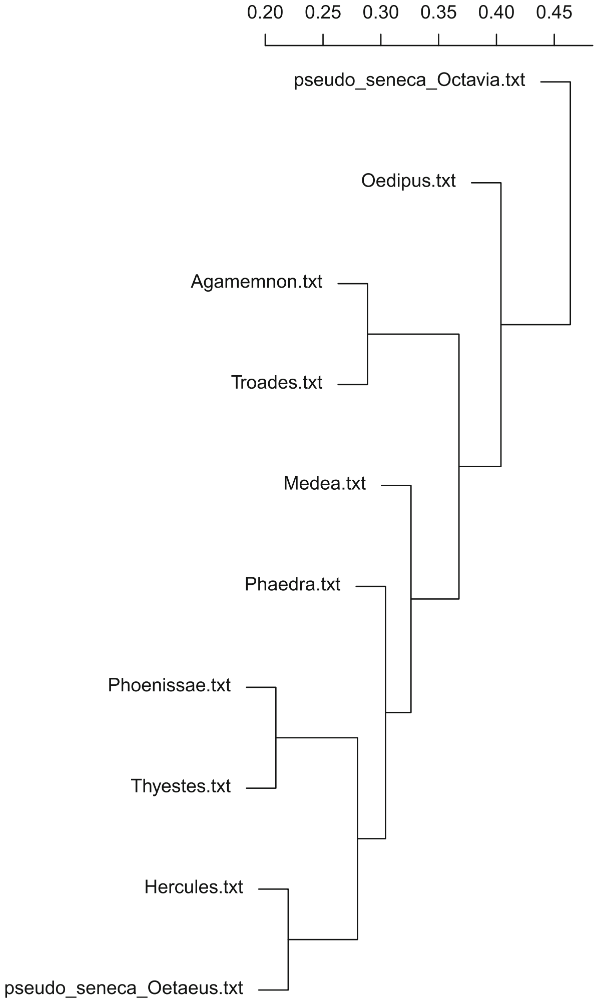

Fig. 3 Clustering the tragedies

The long lasting debate about the attribution to Seneca of the *Hercules Oetaeus* and the *Octavia* has led to the generally assumed conclusion that the *Hercules Oetaeus* is much probably original, while the *Octavia* is an imitation of the tragic style of Seneca.[2] This is reflected by our results too. Indeed, the *Hercules Oetaeus* is collected into the cluster of the original tragedies and, in particular, it is clustered together with the *Hercules* (which is not surprising, because these two tragedies cover a much similar topic). Conversely, the *Octavia* is clustered apart from the other tragedies (like the *Oedipus*).

Figure 4 shows the results of the principal component analysis performed on the tragedies. The lemmas that characterize the *Octavia* in comparison to the other tragedies are *coniunx*, *nero*, *nutrix*, *octavia*, and *seneca*. These words summarize well the contents of this *fabula praetexta* that tells the story of Octavia, who was the first wife (*coniunx*) of the emperor Nero. Further, one of the main arguments in favour of considering the *Octavia* a not original tragedy of Seneca is that one of the characters of the story is named Seneca, which is again reflected by our principal component analysis.

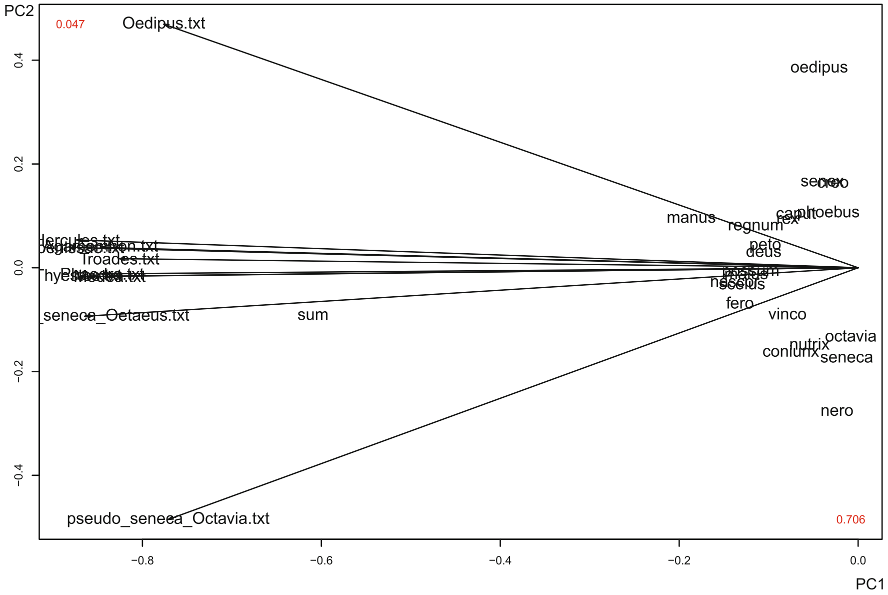

Fig. 4 Principal component analysis of the tragedies

[2] About the attribution of the *Hercules Oetaeus* see [3, 10, 13]. On the *Octavia* see [1, 2, 7, 17]. Among the contributions in favour of the authenticity of both the tragedies see [16].

4.3 Opera Omnia

Figure 5 presents the results of clustering the *opera omnia* of Seneca.

Texts get organized into two main clusters, one including the tragedies and the other featuring the dialogues and the other writings. Within the latter, the *Apocolocyntosis* is clustered apart from the other works and all the seven books of *De beneficiis* get clustered together.

The *Apocolocyntosis* is a menippean satyre (a kind of mixture of prose and poetry) and it is indeed a text quite different from the others of Seneca: it is, thus, not surprising that it belongs to a separate cluster.

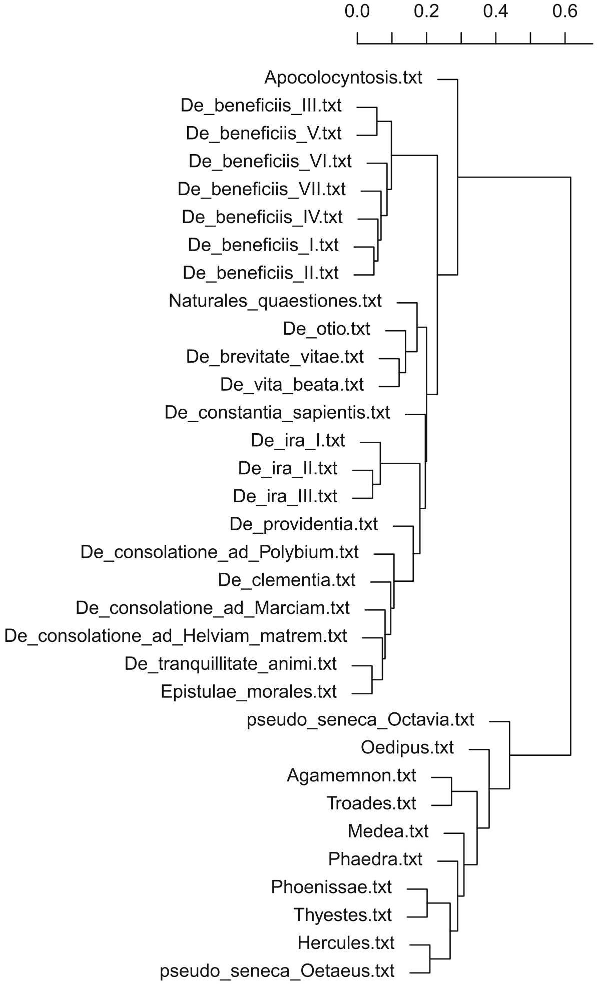

Fig. 5 Clustering the *opera omnia*

The fact that the tragedies are set apart from the other works and that all the books of *De beneficiis* and all those of *De ira* are clustered together shows that our method is able to distinguish texts not only by genre but also by single work.

Figure 6 presents the results of the principal component analysis performed on the *opera omnia* of Seneca. The seven books of *De beneficiis* deviate from the other works and are characterized by the following lemmas: *beneficium*, *gratia*, *gratus*, *ingratus* and *reddo*.

Along all our experiments we observed that *De consolatione ad Polybium* was always clustered separately from the other two *consolationes*. Figure 7 reports the contribution biplot that highlights the lexical features of these three texts, showing that *De consolatione ad Polybium* is characterized by the lemmas *bonus*, *caesar*, *dolor*, *fortuna* and *frater*, while *De consolatione ad Marciam* and *De consolatione ad Helviam matrem* are distinguished by *filius*, *locus*, *mater*, *vir* and *vivo*. The three texts share an high average relative frequency of lemmas like *animus*, *homo* and *natura*.

The biplot reported in Fig. 7 looks different from those presented so far, as it features a massive central black area formed by those lemmas that are shared by the three *consolationes*. Although such an area was present also in all the biplots reported above, it was always removed for presentation purposes. In this case, we left the black area in on purpose, in order to show how big is the number of lemmas with similar relative frequency that are shared by these three texts which, indeed, appear as clustered very close to each other in Fig. 5.

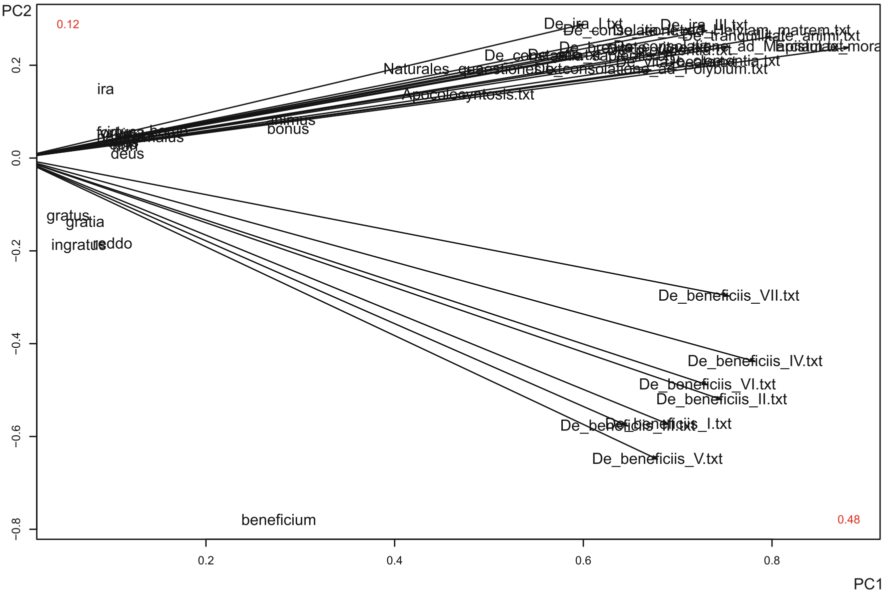

Fig. 6 Principal component analysis of the *opera omnia*

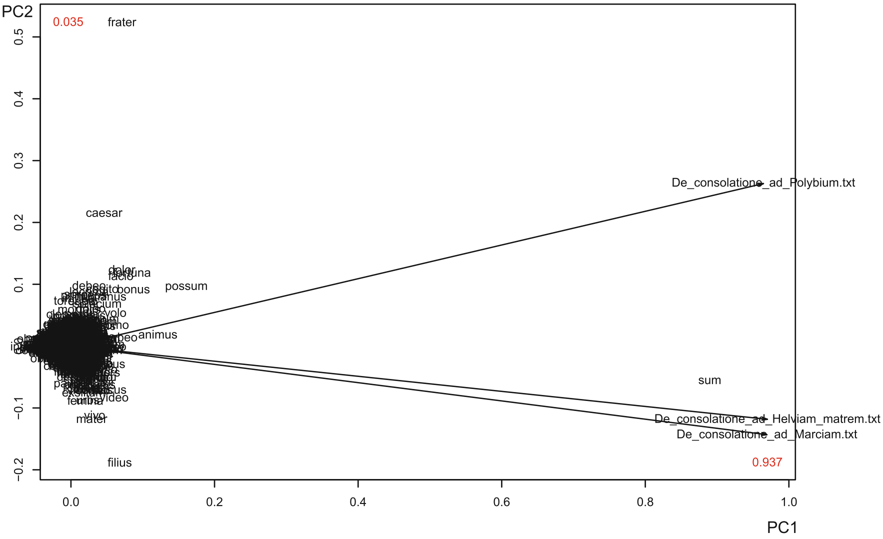

Fig. 7 Principal component analysis of the three *Consolationes*

5 Discussion and Future Work

The main feature of our method is that it provides an objective and unsupervised analysis of textual data. Our results can be replicated and the method is open to be refined in order to achieve better results.

The R software allows an efficient managing of big amounts of data. This gave us the opportunity to perform always full-text analyses, instead of grounding our experiments on manually selected excerpts built in a subjective fashion.

At first, our method wants to quantify and verify (or not) previous intuition-based assumptions on the evidence provided by textual data. For instance, our results do not just show that the tragedies of Seneca are different from his dialogues (which is indeed neither a new nor a really interesting fact), but they report objectively how much different the tragedies are from the dialogues and how much different they are from each other according to their lexical features.

Then, our fully data-driven approach does not only add empirical evidence to subjective intuitions about texts, but it allows also to bring to light previously overlooked relations between texts, like in the case of the relation between *De consolatione ad Polybium* and the other two *consolationes*.

Further, such a method can also be used for authorship attribution purposes, like in the case of the *Octavia*. However, this may lead to promising results just in those cases where the works of disputed attribution differ from the original ones by lexical features. If differences concern other linguistic properties of the texts (ranging from syntax to semantics and, more generally, to literary style), a lexical-based approach is not the best fitting one.

As mentioned above, all our results were driven by the assumption that “you shall know a text by the words it keeps”. This entails that the texts involved in our experiments get clustered according to their lexical properties. In this context, thus, saying that one work is close to another means that they share a relevant amount of (non-function) words showing a similar distribution. In light of the results achieved, such a basic assumption seems to be working, as the organization of texts that automatically results from applying our method corresponds to the different works and genres involved in the several experiments performed.

In the near future, we want to refine our method both by providing a more fine-grained subdivision of data and by exploiting higher layers of linguistic annotation of texts.

As the former is concerned, we shall organize the data according to the sub-parts of the texts (books, chapters etc.): for instance, we should provide one separate file for each letter of the *Epistulae morales*.

As for the latter, we first want to compare the texts by distribution of Parts of Speech (PoS) and colligations (i.e. co-occurrences of PoS). At the higher level, we have to exploit syntactically annotated data (produced by parsers, or made available in treebanks) in order to compare the texts by phrases and/or chuncks instead of single words. And finally, we can use second-order features as well (like semantic descriptions of lexical items provided by Latin WordNet) to enhance the information provided by words.

References

1. Beck, J.W: «Octavia» Anonymi: zeitnahe «praetexta» oder zeitlose «tragoedia»?: mit einem Anhang zur Struktur des Dramas, Duehrkohp und Radicke, Göttingen (Göttinger Forum für Altertumswissenschaft. Beihefte 15) (2004)
2. Bruckner, F.: Interpretationen zur Pseudo-Seneca-Tragödie Octavia, Offsetdruckerei Hogl, Erlangen (1976)
3. del Río Sanz, E.: Problemas de autenticidad del Hercules Oetaeus. Estado de la cuestión, Cuadernos de Investigación Filológica, XII-XIII, 147–153 (1987)
4. Feinerer, I., Hornik, K.: tm: Text Mining Package. R package version 0.5-9.1. http://CRAN.R-project.org/package=tm (2013)
5. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in R. J. Stat. Softw. **25**(5), 1–54. http://www.jstatsoft.org/v25/i05/ (2008)
6. Firth, J.R.: Papers in Linguistics 1934–1951. London University Press, London (1957)
7. Giancotti, F.: L’Octavia attribuita a Seneca. Loescher-Chiantore, Torino (1954)
8. Greenacre, M.: Biplots in Practice. Fundación BBVA, Madrid (2010)
9. Harris, Z.S.: Distributional structure. Word **10**, 146–162 (1954)
10. Iorio, V.: L’autenticità della tragedia Hercules Oetaeus di Seneca. Rivista Indo-Greca-Italica di filologia, lingua, antichità, 20, 1–59 (1936)
11. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River (2002)
12. Meyer, D., Buchta, C.: proxy: Distance and Similarity Measures. R package version 0.4-10. http://cran.r-project.org/web/packages/proxy/index.html (2013)
13. Paratore, E.: Lo Hercules Oetaeus è di Seneca ed è anteriore al Furens, in Acta classica: verhandelinge van die Klassieke Vereniging van Suid-Afrika = Proceedings of the Classical Association of South Africa, I, 72–79 (1958)

14. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/ (2012)
15. Radice, R., Bombacigno, R.: Lexicon IV: Stoics. Biblia, Milano (2007)
16. Ruiz de Elvira, A.: La «Octauia» y el «Hercules Oetaeus»: tragedias auténticas de Séneca, in Urbs aeterna: actas y colaboraciones del Coloquio Internacional Roma entre la Literatura y la Historia: homenaje a la profesora Carmen Castillo, Ediciones Universidad de Navarra, Pamplona, pp. 909–919 (2003)
17. Runchina, G.: Sulla pretesta Octavia e le tragedie di Seneca. In: Rivista di cultura classica e medioevale, vol. 6, pp. 47–63 (1964)