Anca Dinu / Madalina Chitez /
Liviu Dinu / Mihnea Dobre (eds.)

Recent Advances in Digital Humanities

Romance Language Applications

PETER LANG

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available online at http://dnb.d-nb.de.

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Cover illustration:
© Cristina Onet, member of CODHUS Research Centre.

ISBN 978-3-631-81147-4 (Print)
E-ISBN 978-3-631-87773-9 (E-PDF)
E-ISBN 978-3-631-88384-6 (EPUB)
DOI 10.3726/b19920

This publication has been peer reviewed.

www.peterlang.com

Table of Contents

Miguel Cavadas Docampo and Pablo Gamallo Otero

Chapter 11 Automatic Authorship Attribution in the Work of Tirso de Molina

Abstract: Automatic Authorship Attribution (AAA) is the result of applying tools and techniques from Digital Humanities to authorship attribution studies. Through a quantitative and statistical approach this discipline can draw further conclusions about renowned authorship issues which traditional critics have been dealing with for centuries, opening a new door to style comparison. The aim of this paper is to prove the potential of these tools and techniques by testing the authorship of five comedies traditionally attributed to Spanish playwright Tirso de Molina (1579–1648): *La ninfa del cielo*, *El burlador de Sevilla*, *Tan largo me lo fiáis*, *La mujer por fuerza* and *El condenado por desconfiado*. To accomplish this purpose some experiments concerning clustering analysis by Stylo package from R and four distance measures are carried out on a corpus built with plays by Tirso, Andrés de Claramonte (c. 1560–1626), Antonio Mira de Amescua (1577–1644) and Luis Vélez de Guevara (1579–1644). The results obtained point to the denial of all the attributions to Tirso except for the case of *La mujer por fuerza*.

Keywords: authorship attribution, Spanish literature, clustering

1. Introduction

The objective of this article is to highlight the role that certain tools coming from computational linguistics may have in an authorship study and, consequently, to point out the value of AAA as a profitable convergence between disciplines. In order to illustrate this, we will focus specifically on the Tirsian debate with the aim to draw conclusions that may serve as relevant arguments in order to reinforce some of the most supported critical positions. More in particular, we aim to find the most probable authors of five plays from the Golden Age that were traditionally attributed to Tirso de Molina, including the two that introduce the character that will initiate the Don Juan myth. All these plays have been surrounded, especially in the last decades, by great controversy regarding their alleged authorship. Tab. 1 depicts their current situation.

Tab. 1. Texts under discussion and their corresponding possible authors

Text	Possible Authors
El burlador de Sevilla	Claramonte / Tirso
El condenado por desconfiado	Claramonte / Guevara / Mira / Tirso / Collaboration
La mujer por fuerza	Tirso / Other (no name proposed)
La ninfa del Cielo	Guevara / Tirso
Tan largo me lo fiáis	Claramonte / Tirso

Therefore, the authors involved in this study are, apart from Tirso: Andrés de Claramonte, Mira de Amescua and Luis Vélez de Guevara. We have collected and pre-processed several *comedias* by all of them from both the Miguel de Cervantes Virtual Library and the online library of the Association for Hispanic Classical Theater.

2. Method

AAA relies on different statistical measures that take into account the distribution of function words. The selection of the most appropriate measure depends on the available corpus and the objective of the study. However, according to Grieve (2005), the best approach to quantitative authorship attribution is one that is based on the values of as many textual measurements as possible. This is because, in general, small variations in the configurations can produce very large changes in the results. For this reason, five different strategies are used in this work. The five strategies are divided in two opposing approaches: instance and profile-based.

2.1. Instance-based approach

Individual texts are the basic items of the procedure. Following this approach, we use clustering analysis on individual texts so as to group them on the basis of stylistic similarities, which is equivalent, if the adjustments are correct, to group the texts by author. The resulting grouping of individual texts is displayed in a dendogram. The clustering process is based on the Delta measure, created by John Burrows in 2002 specifically for stylometric purposes. This method was implemented by Stylo, a flexible

R package for the high-level stylistic analysis of text collections (Eder, Rybicki and Kestemont 2016). The Delta measure normalizes frequencies by means of z-score to reduce the influence of very frequent words. For fi(D) being the frequency of n-grami in document D, μi the mean frequency of the n-gram in the corpus, and σi its standard deviation, then z-score is defined as follows:

$$z\left(fi\left(D\right)\right)=\left(fi\left(D\right)-\mu i\right)/\sigma i \tag{1}$$

The difference between a set of training texts written by the same author and an unknown text is the mean of the absolute differences between the z-scores (Stamatos 2009). Given the normalized document vectors, the Burrows's Delta is just the Manhattan distance by using normalized frequencies with z-scores. Given documents D1 and D2, distance Delta Δ is computed as follows:

$$\Delta=\sum_{i=}^{n} 1\left|\left(fi\left(D1\right)\right)-z\left(fi\left(D2\right)\right)\right| \tag{2}$$

The lower the Delta value the higher the similarity between the texts studied.

2.2. Profile-based approach

In this approach, the known texts belonging to one author are merged into one single document (profile of the author) and then, a distance measure is computed between the profile of the author and the profile of an unknown text. Four different distance measures were designed and implemented: Kullback-Leibler divergence, Perplexity, Ranking-based distance, and Distributional similarity. These measures represent four different corpus-based strategies to compare texts. They were not originally designed to serve the purposes of AAA, but are commonly employed in other computational tasks such as language identification, language distance, information retrieval and data mining.

Kullback-Leibler. Kullback-Leibler divergence compares two distributions, more precisely, is a measure of how one probability distribution (for instance, the profile of an unseen document) is different from a second, reference probability distribution (the profile of the author). In

Iriarte et al. (2018) it was used to measure the distance between texts written by different social groups of individuals: men / women, university / non-university people, and so on. Given a test or unknown text (T) and the known texts of an author (A), the Kullback–Leibler divergence KL of the distributions T and A is defined as follows:

$$\mathrm{KL}(\mathrm{A},\mathrm{T}) = \sum A(\mathrm{ngr}_i) \log \frac{A(\mathrm{ngr}_i)}{T(\mathrm{ngr}_i)} \tag{3}$$

Equation 3 allows computing how far the T distribution is from the A distribution, taking into account the probabilities of the n-grams (of words or characters) in each compared text corpus, either T or A. The divergence (which is an asymmetric measure) was converted into a symmetric one (i.e., into a distance) by computing the mean of the two complementary comparisons: divergence of X from Y, and divergence of Y from X. In our experiment, we applied Kullback-Leibler divergence on distributions of the most frequent word unigrams.

Perplexity. Perplexity is frequently used as a quality measure for language models built with n-grams extracted from text corpora, and can be used to measure how well a model (for instance, the profile of an author) fits the test data (the profile of an unseen document). More formally, perplexity is the normalized inverse probability of an input test. It can be used to compare a test text (T) with the author model (A). The perplexity PP of T given the author model A is defined by the following equation:

$$\mathrm{PP}(\mathrm{A},\mathrm{T}) = 2 - \sum T(\mathrm{ngr}i) \log 2M(\mathrm{ngr}i) \tag{4}$$

where ngri is a n-gram shared by both T and A. Equation 4 can be used to set the divergence between a test set and the author model. The lower is the perplexity of T given A, the lower is the distance between the two compared objects. Texts may be modelledwith n-grams of either words or characters. In our experiments, we applied PP distance to texts lwith 7-grams of characters. In other pieces of work, PP was also used to compare the linguistic distance between 40 European languages (Gamallo, Pichel and Alegria 2017), as well as to compute the distance between diachronic varieties of the same language (Pichel, Gamallo and Simões 2018).

Rank-Based. The rank-based distance between two languages is obtained by comparing the ranked lists of the two languages. It takes two-word profiles (the author and the unseen document) and calculates a simple rank-order statistic based on an "out-of-place" measure. This measure determines how far out of place an n-gram in one profile is from its place in the other profile (Cavnar and Trenkle 1994). This measure is often used to compute language identification (Gamallo et al. 2014). More formally, given the ranked lists RankT and RankA of the test text (T) and the texts of a given author (A), respectively, the rank-based distance, R, is computed as follows:

$$R(\mathrm{A},\mathrm{T}) = \sum_{\mathrm{i}=1}^{K} \left| \mathrm{Rank}_A\left(\mathrm{ngr}_i\right) - \mathrm{Rank}_T\left(\mathrm{ngram}_i\right) \right| \tag{5}$$

where K stands for the number of the most frequent n-grams, RankA(ngri) is the rank of a specific n-gram, ngri, in A, and RankT(ngri) is the rank of the same n-gram in T. In our experiments, we applied this distance on lists of word unigrams.

Distributional similarity. As a fourth measure, we use vector space models, which is one of the most popular representations of document vocabulary, to compute distributional similarity between documents. In particular, this strategy compares the word vectors extracted from the author's profile with those extracted from the unseen document. Given that it is a measure of similarity, the higher its value, the higher will be the similarity between the texts, being the maximum value 1. As it works in an inverse way to the other three, which are distance measures, we turn it into a distance by subtracting the values from 1. In our experiments, we used Distributional similarity by considering the comparative study made by Afzali and Kumar (2017) on different metrics, namely Cosine, Jaccard and Dice, to evaluate their performance in finding the similarity of two text documents. Cosine outperformed the other metrics in a significant way. Given A and T, Cosine (as a distance) is defined in this way

$$\mathrm{Cosine}(\mathrm{A},\mathrm{T}) = 1 - \frac{\sum A\left(\mathrm{ngr}_i\right) T\left(\mathrm{ngr}_i\right)}{\sqrt{\sum A\left(\mathrm{ngr}_i\right)^2} \sqrt{\sum T\left(\mathrm{ngr}_i\right)^2}} \tag{6}$$

Mean of measures. Finally, the four measures defined above were merged by averaging their values. In order to compute the mean of the values obtained by the four distance measures, their final scores were normalized. The four measures and its mean combination have been implemented by the authors into an executable written in PERL language (see *Autoria*, Gamallo and Cavadas n.d.).

It must be stressed that the use of four different profile-based strategies, in addition to the instance-based one defined above, allows us to reach more solid and reliable conclusions about authorship. These are not different configurations of a single strategy, as is usually done in most Stylo-based work, but five complementary methods of covering diverse aspects of the same problem.

3. Experiments

3.1. Clustering analysis

To configure the Stylo tool (version 0.6.9) properly we carried out a set of preliminary tests with a reduced version of the corpus (eliminating works of unknown authorship). This way we checked which values offered the most consistent results. We concluded that the most appropriate value for the MFW parameter (most frequent words) was 250 for both maximum and minimum, in order to build one single dendogram. This is the essential point of the experiment, as variations on that figure may alter the results substantially. The features we decided to extract from the texts were unigrams of words, that is, tokens, and we did not consider necessary to perform any culling. The dendogram resulting from the experiment is depicted in Fig. 1.

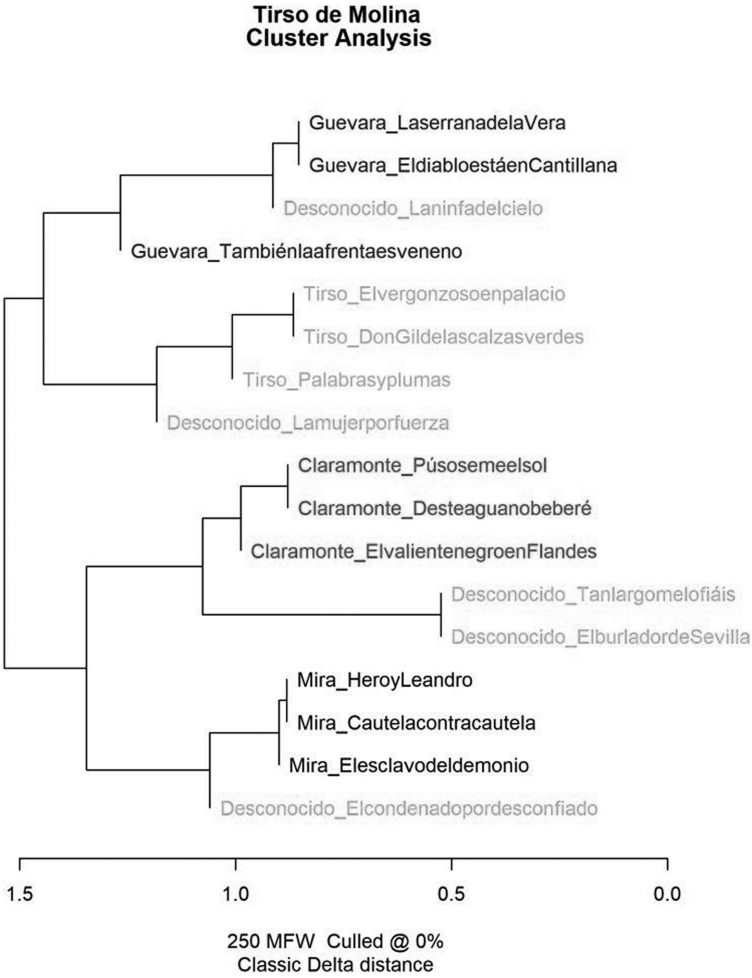

Fig. 1. Dendrogram of the clustering analysis using Delta Burrows distance

As was to be expected, the smallest distance is that which separates *El burlador de Sevilla* and *Tan largo me lo fiáis*, since they are, in a high percentage, the same text. They are undoubtedly grouped together with Claramonte's production. Mira seems to be the author with the most homogenous style by virtue of the small distance between his three comedies of assured authorship, while Guevara presents the most unstable style. The pairs Guevara-Tirso and Claramonte-Mira are the ones that share the most stylistic similarities. As for the attributions of the plays of unknown authorship, the results obtained are categorical: *el condenado por desconfiado* is associated with the plays of Mira, *La mujer por fuerza* is agglutinated with the plays of Tirso, and *La ninfa del cielo* joins Guevara, the latter being the most categorical attribution of all.

3.2. Distance measures

We implemented an AAA free software tool to use the four distance measures defined in the previous section: Perplexity, Kullback-Leibler divergence, Ranking-based measure, and Distributional similarity. Tab. 2 shows the normalized average scores of the four measures.

Tab. 2. Mean of the results obtained comparing doubtful authorship plays with authors under discussion. Authors abbreviations: Mira = Mira de Amescua, Guev = Luis Vélez de Guevara, Tirso = Tirso de Molina, Clar = Andrés de Claramonte

La ninfa del Cielo	*El burlador de Sevilla*	*Tan largo me lo fiáis*	*La mujer por Fuerza*	*El condenado por desconfiado*
0.000 Mira	*0.000 Clar*	*0.000 Clar*	*0.043 Tirso*	*0.123 Mira*
0.596 Guev	*0.571 Tirso*	*0.575 Tirso*	*0.193 Mira*	*0.192 Clar*
0.631 Tirso	*0.685 Mira*	*0.620 Mira*	*0.546 Clar*	*0.587 Tirso*
0.720 Clar	*0.830 Guev*	*0.718 Guev*	*0.911 Guev*	*0.961 Guev*

La ninfa del cielo, whose authorship for Guevara was established as quite sure by experts, does not seem to come so close to the style of this author, but rather presents more confluences with Mira de Amescua. The results for *El burlador de Sevilla* and *Tan largo me lo fiáis* are sufficiently categorical and similar across all measures to reject Tirso de Molina's supposed authorship. All the measures aim to support the position of critics

who defend the authorship of Claramonte. *La mujer por fuerza* and *El condenado por desconfiado* are the plays that leave more room for doubt and alternative hypotheses, since the four distances differ in small values from each other. In fact, the distance between the first and second authors is less than or equal to 0.15 in all cases. Mira and Tirso, respectively, are the most likely authors of these two comedies.

4. Conclusions

It is now up to us to carry out a joint comparison of the various results obtained by all the strategies. This will allow us to draw conclusions in order to confirm or not the authorship of the texts. Authorships proposed by the different methods, including traditional philological studies, are illustrated in Tab. 3.

Tab. 3. Authorship attributions of the plays under discussion proposed by the strategies

	Philological studies	**Clustering Analysis**	**Distance Measures**
La ninfa del Cielo	Tirso / Guevara	Guevara	Mira
El burlador de Sevilla	Tirso / Claramonte	Claramonte	Claramonte
Tan largo me lo fiáis	Tirso / Claramonte	Claramonte	Claramonte
La mujer por fuerza	Tirso / Other	Tirso	Tirso
El condenado por des- confiado	Tirso / Claramonte / Guevara / Mira / Collaboration	Mira	Mira

In short, Tirso has been attributed a considerable number of works based on conjectures and critical arguments lacking in solidity and documentary proof. So far, a high percentage of the production plays traditionally assigned to Tirso do not actually belong to this author. In the 17th century comedies were published under the name of famous authors in order to increase sales; it seems that subsequent literary critics have been carried away by this personalist tendency by favoring attributions to renowned authors. The curious thing about Tirso de Molina's case is that these false attributions have been elaborated one on top of the other, in such a way that questioning one implies questioning them all. It is important to point

out that the controversial plays are precisely those on which the fame of Tirso among the public and his excellent critical appraisal are based. So perhaps the place occupied by this playwright in the history of Spanish literature should begin to be reconsidered. It is more urgent, however, to draw attention to the traditional studies of authorship attribution, which on many occasions have not respected the basic principles of scientific rigor that should govern any kind of humanistic research.

The rotundity of the results from the non-traditional studies, along with the most popular philological hypothesis, force us to position ourselves in favour of those theories that propose less famous authors who have been relegated to a second place in the panorama of Golden Age theatre. Among these less-famous authors, Mira de Amescua, Vélez de Guevara and, especially, Andrés de Claramonte should be pointed out. The conclusive proof that Claramonte was the author of the first don Juan seems to be closer than ever. An act of justice would be to vindicate his work, starting by editing it. Yet, in spite of the evidence shown by our study, in order to be even more certain of the results, it will be still necessary to approach this problem with new studies that use more texts, more authors and more advanced NLP strategies such as those based on distributional semantics and contextualized word embeddings (Gamallo et al. 2019).

In future work, the comedy *El condenado por desconfiado* would deserve a separate study, as its authorship hypotheses are too varied and confusing to fit into our current work. For this purpose, it will be necessary to employ tools that study pieces of texts separately so as to determine if it is a work composed in collaboration or if one of the proposed authors is indeed the authentic one. The development and improvement of the AAA tools is, in fact, another step to follow in order to continue deepening the AAA studies. This is the most promising point of our work, since we explored computational strategies that are useful and efficient in this task, even though their original functionality was not the quantification of style for authorship attribution. We think that, in the future, it will be possible to find other techniques and strategies that fit well within the AAA studies. In any case, the most urgent initiative that should be encouraged is the edition and digitization of Golden Age plays, so that it is possible to replicate our experiment on the work of Tirso on a large scale.

References

Azfali, M. and Kumar, S. (2017) "Comparative Analysis of Various Similarity Measures for Finding Similarity of Two Documents". *International Journal of Database Theory and Application* 10(2), 23–30.

Cavnar, W.B. and Trenkle, J.M. (1994) "N-gram-based Text Categorization". *Proceedings of the Third Symposium on Document Analysis and Information Retrieval*. 11–13 April, Las Vegas. Las Vegas: University of Nevada.

Eder, M., Rybicki, J. and Kestemont, M. (2016) "Stylometry with R: A Package for Computational Text Analysis". *R Journal* 8(1), 107–121.

Gamallo, P. and Cavadas, M. (n.d.) *Autoria: Authorship Attribution* [online] available from https://github.com/gamallo/Autoria. (Accessed: 23 March 2022).

Gamallo, P., Garcia, M., Sotelo, S. and Pichel, J.R. (2014) "Comparing Ranking-based and I Bayes Approaches to Language Detection on Tweets". *Proceedings of XXX Congreso de la Sociedad Española de Procesamiento de lenguaje natural*, 18–20 September, Girona. Girona: SEPLN.

Gamallo, P., Pichel, J.R. and Alegria, I. (2017) "From Language Identification to Language Distance". *Physica A* 484, 162–172.

Gamallo, P., Sotelo, S., Pichel, J.R. and Artexte, M. (2019) "Contextualized Translations of Phrasal Verbs with Distributional Compositional Semantics and Monolingual Corpora". *Journal of Computational Linguistics* 45(3), 395–421.

Grieve, J. (2005) *Quantitative Authorship Attribution: A History and an Evaluation of Techniques*. Burnaby: Simon Fraser University.

Iriarte, Á., Gamallo, P. and Simões, A. (2018) "Estratégias Lexicométricas para Detetar Especificidades Textuais". *Linguamática* 10(1), 19–26.

Pichel, J., Gamallo, P. and Alegría, I. (2019) "Measuring Diachronic Language Distance Using Perplexity: Application to English, Portuguese, and Spanish". *Natural Language Engineering*, 1(1), 1–22 (2019).

Stamatos, E. (2009) "A Survey of Modern Authorship Attribution Methods". *Journal of the American Society for Information Science and Technology* 60(3), 538–556.