Digital Scholarship in the Humanities, 2022, **00**, 1–15
https://doi.org/10.1093/llc/fqac063
Full Paper

Challenging stylometry: The authorship of the baroque play *La Segunda Celestina*

Laura Hernández-Lorenzo [1,*] **, Joanna Byszuk** [2]

[1]Departamento de Literatura española e hispanoamericana, Universidad de Sevilla, Spain
[2]Pracownia Metodologiczna, Instytut Języka Polskiego (Polska Akademia Nauk), Poland

*Correspondence: Laura Hernández-Lorenzo, Universidad de Sevilla. C. Palos de la Frontera, s/n, 41004 Seville, Spain. E-mail: lhernandez1@us.es

Abstract

The aim of this study was to verify the possibility of Sor Juana Inés de la Cruz authoring the anonymous part of the baroque play *La Segunda Celestina*, commissioned to Agustín de Salazar, and left unfinished after his death. This is a first systematic stylometric study on this problem and a baroque hispanoamerican text. In our study, we faced building a balanced corpus from few available resources, and took extensive evaluation measures to deal with unclear stylometric signals. We use a variety of established attribution and verification methods, and introduce a novel evaluation procedure of examining historic texts with scarce corpora. The results support Sor Juana's authorship, and unravel new connections between her and other authors of the time, showing, still undermined, powerful impact of her works on the epoch. The solutions adopted in solving methodological problems of such a complex task show how stylometry can overcome similar challenges.

1 Introduction

In November 1675, Spanish writer Agustín de Salazar (1642–75) died leaving unfinished the play *La Segunda Celestina*, which he was writing on commission for the birthday celebrations of the Spanish Queen Mariana de Austria (1634–96), widow of Philip IV (1605–65), and mother of Charles II (1661–1700), to be held at the royal palace. Salazar had written the 'loa', the first and second 'jornada', and the beginning of the third, leaving this last one mostly unfinished (Sabat de Rivers, 1992; Schmidhuber de la Mora, 2016). Given the proximity of the Queen's birthday, taking place on 22 December, the unfortunate event of his death forced the organizers of the Queen's celebrations to replace Salazar's play with the theatrical performance of *Faetón* by Calderón de la Barca (1600–81).

However, *La Segunda Celestina* was finished by an anonymous writer and performed the following year. For centuries, both the identity of this author and the document containing the version of the play with the anonymous ending were to be discovered (Sabat de Rivers, 1992). In 1990, Schmidhuber de la Mora published a newly discovered 'suelta'[1] of *La Segunda Celestina* with the anonymous ending and claimed it had been written by Sor Juana Inés de la Cruz, a prominent Hispanoamerican writer of the time, whom he also thought to have made significant changes to the original (Schmidhuber de la Mora and Peña Doria, 1990). A long and heated controversy followed this declaration, and eminent scholars such as Octavio Paz and Antonio Alatorre participated in the debate. In spite of their arguments, the authorship problem remains far from being solved, and the hypothesis of Sor Juana's writing the ending, although considered favourably by some scholars, is still to be confirmed.

Given significant advances in the computational methods of authorship attribution and greater availability of digitized resources, in this article, we seek to provide a more scrutinous analysis of possible contribution of Sor Juana to the preserved text of *La Segunda Celestina* with the anonymous ending.

For this purpose, this article is structured as follows: after this introduction (Section 1), we sum up the debate on Sor Juana's authorship and the main adduced arguments for and against her authorship (Section 2). Then, we present the dataset used in our study (Section 3) and discuss the methodology used (Section 4). After that, we present the analysis we carried out, as well as the obtained results (Section 5). Finally, we draw some conclusions (Section 6). Additionally, there is an Appendix at the end of our article with complementary materials.

2 The Debate on Sor Juana's Authorship

Different authors, editors, and scholars were involved in the controversy around *La Segunda Celestina*. Agustín de Salazar wrote the first draft of the play minus the ending, of which Sor Juana Inés de la Cruz is considered by some scholars to be a probable author. Two editors of Sor Juana's literary works supported her authorship: Castorena y Úrsua and Alberto G. Salceda. This claim was then supported or rejected by various literary scholars, most notably Schmidhuber de la Mora, Octavio Paz, Antonio Alatorre, Alfonso Sánchez Arteche, José Pascual Buxó, Georgina Sabat de Rivers, and Thomas O'Connor.

The hypothesis of Sor Juana as the author of the anonymous ending had actually started some years ago, when the 'suelta' containing it was still to be discovered. This idea came into being in 1700, when the editor of *Fama y obras posthumas*—a posthumous printed edition of Sor Juana's works—Castorena y Úrsua (1677–33), mentioned that she finished and improved a literary text by Salazar:

> A poem left unfinished by Don Agustín de Salazar, and perfected by the poet with gracious sense, whose original is held in the discrete esteem of Don Francisco de las Heras, a gentleman of the order of Santiago, a ruler of this town, and because it belongs to the first volume, I do not give it a stamp in this book, and it is being printed to represent for their majesties[2] (de la Cruz, 1700, s.f.).

Years later, when the last volume of Sor Juana's entire works was published in 1957, one of the editors, Alberto G. Salceda (Méndez Plancarte and Salceda, 1957), thought that the collaborative poem mentioned by Castorena was actually a comedy in verse written by Salazar and Sor Juana, and that most probably it was *La Segunda Celestina*. He then connected Castorena's mention with Sor Juana referencing a comedy about Celestina in her own play *Los empeños de una casa*[3]:

> Amigo, mejor era Celestina,
> en cuanto a ser comedia ultramarina:
> que siempre las de España son mejores,
> y para digerirles los humores,
> son ligeras; que nunca son pesadas
> las cosas que por agua están pasadas.
> Pero la Celestina que esta risa os causó, era mestiza
> y acabada a retazos,
> y si le faltó traza, tuvo trazos,
> y con diverso genio
> se formó de un trapiche y de un ingenio.
> Y en fin, en su poesía,
> por lo bueno, lo malo se suplía… (de la Cruz, 1692, f. 502, t. II).

From Salceda's point of view, with this mention to a Celestina's comedy, Sor Juana is referring to her participation in *La Segunda Celestina*, which she playfully criticizes because of the role she played in finishing the text. In this sense, she qualifies the play as mestiza, a work that was written by a Spanish author (Salazar) and a Hispanoamerican one (herself). Since she had to hurry finishing the text created by another writer, it was a patchwork; which made the play not entirely symmetrical, and left traces of different author's hands; however, she humbly declares that the brilliance of Salazar's writing makes out for her own flaws.[4]

In December 1989, Schmidhuber de la Mora found a suelta of *La Segunda Celestina* at the University of Pennsylvania and, after showing the text to Octavio Paz, published it adding Sor Juana's name along with Salazar.[5] At the same time, Antonio Alatorre had found another 'suelta' of the play at the Spanish National Library and was also preparing a printed edition (Sabat de Rivers, 1992, p. 499).

The discovery led to a discussion between Sor Juana's experts on the attribution of the play. On the one hand, Octavio Paz (1990) and Schmidhuber de la Mora (1991, 2016) pointed to Sor Juana as the author of the anonymous ending and argued that she made significant changes in the rest of the play. On the other hand, Antonio Alatorre (1990) rejected the attribution for various reasons, the most important being the following ones:

Firstly, the problem of navigation times, which would not have allowed the play to go from Spain to America and the other way around in time to be presented at the Royal Palace in Madrid in 1676. We know for sure Sor Juana never left America, and Salazar's unfinished manuscript was in Madrid when he died. Thirteen months were not enough for the text of the play to travel overseas and still give Sor Juana the necessary time to finish it.

Secondly, Sor Juana was at the time under the influence and control of a very strict confessor, padre Núñez. He would not have allowed her to participate in the writing of a comedy, especially one about a picaresque character such as Celestina.

Alatorre concluded that the ending of *La Segunda Celestina* written by Sor Juana is not the one Schmidhuber de la Mora found and published, but another version of the text which has not been found yet, and was probably written later than 1676. He was joined in the rejection of Sor Juana's authorship by other scholars, Sánchez Arteche (1991) and Pascual Buxó (1991).

To counter this position, Schmidhuber de la Mora (1991) tried to prove the attribution making historical, linguistic, and even simple stylometric arguments. However, Alatorre and Pascual Buxó persisted in their view.

Another scholar,[6] Georgina Sabat de Rivers, conducted a comprehensive study and research in this topic (Sabat de Rivers, 1992). She concludes that, although there is not enough evidence to make it a fact, it is highly probable that Sor Juana indeed wrote the ending, but not that she edited the whole text. Countering Alatorre's arguments, she points out that ships with court mail to and from America would depart every three months. As a result, Sor Juana would indeed have had little time to finish the comedy, but enough to do so given that comedy is a genre with fixed patterns and rules. Also, Sabat de Rivers supports the previous idea first proposed by Schmidhuber de la Mora and Paz that the Marquis of Mancera[7] would be the one to ask Sor Juana to finish the play and send her the text. In that case, the nun would likely have chosen to favour Mancera's request, even if it was against her confessor's approval. Once the finished text arrived in Spain, it would immediately be printed, as the court controlled the printed presses. Nevertheless, she considered it to be highly improbable that Sor Juana would have made changes in the rest of the play as there is another version of the play with a different ending, written by Juan de Vera Tassis.[8] Sabat de Rivers declares that even if Vera Tassis ending was written later than Sor Juana's one, given that the rest of the play is the same, excluding minor changes, it is obvious that it was an original text by Salazar.

Another scholar, O'Connor (1992) provides some more historical data, based on which he agrees with Sabat de Rivers in the probable Sor Juana's authorship of the ending as well as the improbability of Sor Juana editing the rest of the text.[9]

Taking into account this complex authorial problem on *La Segunda Celestina* and the debate on Sor Juana's authorship, in this study, we aim to answer two main research questions through the use of stylometric methods: firstly, to determine if Sor Juana was the author of the controversial anonymous ending and secondly, to find out if this anonymous writer made changes to the rest of the play.

3 Dataset

In the course of our study, we found that the availability of digitized Spanish texts, especially historic ones, poses a great problem due to few resources and repositories, as well as poor state of digitization: for Spanish works, it mostly means scanning images of early editions (especially manuscripts or old prints). The typographic variance as well as poor state of preservation of some editions, makes for the fact that OCR, be it the one already available in the files or obtained anew, is not very useful.

As a result, our corpus was composed based on various sources. The text of *La Segunda Celestina* was extracted from a digital edition (Schmidhuber de la Mora, 2016), and converted into plain text. Other dramatic works by Sor Juana were extracted from the Cervantes Virtual Library (Bia and Pedreño, 2001).[10] Salazar's plays were much more difficult to find, as there are no proper digital editions,[11] which forced us to extract the texts from the image digitization of his texts offered by the 'Biblioteca Digital Hispánica'.[12] The OCR provided by the library and our software (ABBYY-FineReader 12) produced so many irregular errors that we decided to manually transcribe *El amor más desgraciado* and *Más triunfa el amor rendido*. While we are aware that OCR or HTR solutions based on machine learning, such as Transkribus (Kahle *et al.*, 2017; Muehlberger *et al.*, 2019),[13] could perhaps help us obtain slightly better results, we decided against trying it. Our decision was guided by three reasons: the scarcity of the data that could be used as a training set, high irregularity of expected errors as well as the poor state of print on many pages of these editions which was unlikely to be correctly recognized. Preparing a training set and post-OCR corrections would be more time-consuming than transcribing and proofreading two plays. To ensure as few typos as possible we both proofread each of the transcriptions checking them against the prints available through the library. We also share our transcriptions on the Github page of the project (see the Appendix) to allow future researchers to use them in their studies and perhaps also in training OCR tools for Golden Age Spanish prints.

To place the problem in a broader perspective and literary context of its time, we used the Canon-60 corpus (Oleza Simó, 2014), a collection of digitized Spanish Golden Age plays that includes canonical baroque works. However, we stayed alert that this corpus is imbalanced for quantitative studies: some authors are over-represented, whereas others, less famous or relevant for literary history, are represented with singular plays.

Our final corpus combined the Canon-60, Sor Juana's and Salazar's texts, and lacked balance in terms of genre (a mix of secular and religious plays, as well as comedies and tragedies), gender (only two women: Sor Juana and María de Zayas, which shows that the relevance of other Spanish baroque female writers needs further studies and creating editions of their works[14]), and, finally, nationality: all the authors are Spanish-born, except for Sor Juana who was born in 'Nueva España' (i.e. Mexico). To account for as many factors as possible, we limited our corpus to only one genre: 'la comedia de capa y espada' (Gregg, 1977), and further restricted it in the next steps of the study.

4 Methodology

This section discusses classification methods applied in our study.

Apart from method selection, an important part of the study is selection of features to be used as a material for building classifying profiles and as features to compare. Unfortunately, there are few systematic comparisons of different features types—and even fewer considering Spanish Literature—with the evidence pointing to single most frequent words (MFWs) as the most reliable style carrier (Eder, 2011; Cafiero and Camps, 2019), which is why we focus on this kind of features and they are used consistently across different kinds of analyses. Other examined but not discussed features included character n-grams and PoS tags, which were outperformed by single words.

4.1 Applications of stylometry to authorship attribution

In the following study we apply a number of stylometric methods to examine possible authorship of *La Segunda Celestina*, and especially its last part, tentatively attributed to Sor Juana Inés de la Cruz. We start with the assumption that Schmidhuber de la Mora's hypothesis of her authorship of the anonymous part and some editing to the whole text is possible, and seek to verify it.

All stylometric methods are rooted in the observation that the frequential pattern of use of some MFWs, largely function words bearing little semantic meaning, carries an information about 'a stylistic signal' of an author, which can be used in creating predictive models allowing for distinguishing between writers (Mosteller and Wallace, 1964), and to some extent other stylistic factors such as genre or chronology (Lutosławski, 1897; Stamou, 2008; Calvo Tello *et al.*, 2017; Calvo Tello, 2019). While the earliest studies with this approach go as far as Lorenzo Valla in the 15th century and Augustus de Morgan, Thomas C. Mendenhall, and Wincenty Lutosławski in the 19th, only the development of more powerful personal computers allowed for a more robust development of the discipline, with famous study by Mosteller and Wallace on *The Federalist Papers* in 1964 and John Burrows's work on Jane Austen (Burrows, 1987). Burrows's stylometric research was particularly important for the field, as he introduced a measure allowing for more reliable comparison of not only individual words but whole texts: the so-called Burrows's Delta (Burrows, 2002). This distance measure is one of the best performing ones according to state-of-the-art research, along with the so-called Cosine Delta (Smith and Aldridge, 2011; Evert *et al.*, 2017).

To address the varying length of texts in our corpora, we applied sampling to the corpus, examining chunks of texts rather than their full versions. The sampling procedure is popular in stylometry and authorship attribution studies, as it allows to (a) account for different lengths of texts and perform the classification or even-sized pieces of them, (b) examine the stylistic profile of smaller excerpts of texts, thus detecting possible shifts between various parts in more detail, or to put simply—see if the recognized author is consistent across the whole text or there are more than one influences, (c) account for possible stylistic dominance of certain parts of the text over others in the whole scope, for example, a strong authorship signal in the first part of the text but not in the latter. We used the sequential sampling, that is, the texts in our corpus were cut into parts of the same length in the order they come, so, for example, in the 2000 word scenario, Salazar_Triunfa_1 will cover the words 1–2,000, Salazar_Triunfa_2 2,001–4,000, etc.

In our study, we chose to use 'stylo', an R package for computational text analysis (Eder *et al.*, 2016) for all conducted analyses. Our choice was guided by the wide selection of options offered within the tool, allowing for conducting various types of stylometric studies and good adjusting of parameters, as well as the relative user-friendliness and popularity of the tool in the field, which allows any willing researchers to replicate and verify our results.

4.2 Network analysis

The first step in our analysis was a network analysis of relations between authors and texts in our corpus. Network analysis of big corpora before conducting actual classification experiments is particularly useful in understanding the outline and potential biases present in a dataset, resulting, for example, from weak authorial signals, which would be exhibited by no discernible groups within the network; or dominance of some authors, which would be evidenced by detecting hubs—nodes with a number of connections significantly larger than the average.

Our stylometric networks were constructed following the implementation by Eder (2017b) in 'stylo', which creates an automatic table of textual connections for every cluster analysis or consensus tree produced. To calculate stylistic differences between texts, we used bootstrap consensus tree algorithm, a method relying on a series of hierarchical cluster analyses to determine which texts are repeatedly recognized as each others' strongest connections[15] (Eder, 2013), and, should this number of repetitions cross a desired and determined threshold, on this basis considered their neighbours. One particular advantage of this method is that it verifies similarity of sample texts over a range of frequency settings, and another is that it preserves information about a number of times two nodes were considered neighbouring, which then translates into the weight of particular edges, thus offering quite fine-grained information about strong influences between texts.

As 'stylo' produces numerical representation of the networks (with some, but limited, support for visualization), we used Gephi (Bastian *et al.*, 2009) to visualize it and to further examine it with Louvain community detection algorithm (Blondel *et al.*, 2008). Community detection is a method derived from social networks analysis, used for automatic discovery of particularly densely connected nodes, in our case—texts. Louvain algorithm has been previously described as fit for the task of detecting textual communities (cf. *inter alia* Newman (2006) or Traag *et al.* (2019)) and examined more closely for stylometric uses in Ochab *et al.* (2019) and Ochab and Essler (2019) which deemed it as providing more accurate and granulated results for authorship and general groupings attributions than usual clustering methods. While Ochab and Essler argue slightly better performance of modularity optimization algorithm over Louvain modularity, both methods are considered by them as highly valuable, and Louvain had to us a benefit of being already implemented in Gephi, which allowed us to use fewer tools without compromising results.

4.3 Authorship attribution and verification

There are two distinguishable types of authorship examination—attribution and verification (Koppel *et al.*, 2009). Authorship attribution includes determining which of the candidate authors in the examined dataset is the most likely to be the author of the text in question. In turn, authorship verification deals with checking whether any of the candidate authors is at all likely to author the examined text. Therefore, the two approaches differ in the sense that authorship attribution is a close class problem, whereas authorship verification is known as open-set attribution, as it acknowledges the possibility that the real author of the disputed text is not among the candidates.

While both have been applied in numerous stylometric investigations, attribution obtains significantly more attention in applications (starting with Mosteller and Wallace, 1964) and review of methods (Grieve, 2007; Stamatatos, 2009) than verification, with the latter approach developing more intensely only in the last decade or so (e.g. Koppel *et al.*, 2009; Koppel and Winter, 2014; Kestemont *et al.*, 2016a).

We believe that, like most studies concerning historical texts, the particular context of *La Segunda Celestina* calls for including both types of the analysis. Given the importance of the event for which the play was commissioned, it is reasonable to assume that only a writer known to the court as experienced and competent to fulfill the task would be trusted with it. However, some doubt must be reserved to accommodate for the chance that the actual author could be someone not preserved in the canon as we know or for whom, despite our best efforts, it was not possible to obtain digitized texts at this point.

For authorship attribution, in addition to already mentioned unsupervised classification with Cosine Delta, we use supervised machine learning classification methods, such as support vector machines (SVM), Delta method, and nearest shrunken centroids (NSC). In this type of analysis, one of the above-mentioned classifiers 'learns' the style of each of the authors, based on which knowledge it is able to point which of them the text of disputed authorship is the most similar to. Importantly, this type of classification also enables more control over an experiment and facilitates performing cross-validation, that is, a procedure of evaluating the results.

Stylometry unfortunately lacks a proper benchmark comparison of machine learning methods for authorship attribution, with the most notable, but slightly outdated, study of Jockers and Witten (2010) proposing NSC and regularized discriminant analysis as best performing. Despite their criticism of SVM, we largely focus on this method in discussing results as it was proven to perform with a more reliable stability when dealing with high dimensional and sparse data (Stamatatos, 2013; Franzini *et al.*, 2018).

For authorship verification, we use the General Imposters method (Koppel and Winter, 2014; Kestemont *et al.*, 2016b). The main difference from the approach described above is that, in this case, rather than try to guess the author, the algorithm compares pairs of texts against the others to see whether any of them is significantly more similar to one another than to the rest of the dataset.

4.4 Classification of particular parts of *La Segunda Celestina*

Finally, as one of our goals is to find out if there are different authorial takeovers in our texts, we used a stylometric sequential method, Rolling Classify. Proposed by Eder (2016), this procedure is designed to assess mixed authorship, dividing the text studied in fragments or samples, and analysing each of them for their stylistic consistency.

Given their good performance in previous studies—and applied to Sor Juana's plays in particular, that is, *Amor es más laberinto* (Hernández-Lorenzo, 2019)—and in order to test consistency and stability of our results, we decided to apply Rolling Classify in its three different flavours: SVM, NSC, and Delta.

5 Analysis

This section of the article presents the results of our study, starting from the broad perspective of Spanish and New Spanish baroque drama to the focus on the specific question of the authorship of *La Segunda Celestina*.

5.1 Works from New Spain in the Spanish Baroque perspective

We approached the issue of verifying Sor Juana's authorship in a multi-step study, starting with a distant look at the literary context of *La Segunda Celestina*. With the primary network analysis (see Fig. 1) conducted on the large corpus (Canon-60 + Sor Juana + Salazar) with Bootstrap Consensus algorithm for 300–1,000 MFWs, as implemented in the stylo package (Eder *et al.*, 2016), we determined optimal settings granting stable results. We decided against using culling, a procedure of removing words used predominantly only in singular or few works within a corpus from examined features, as it completely distorted any authorial signal. We also chose to rely on the Cosine Delta distance measure which, again, offered the most stable results, confirming earlier studies showing its greater reliability than Classic Delta (Jannidis *et al.*, 2015; Ochab *et al.*, 2019).

The application of Modularity algorithm, in our case Louvain community detection algorithm as implemented in Gephi with resolution equal 1, to the resulting network leads to discovery of thirteen different communities, mostly author or play related, although over-represented authors are present in different communities (e.g. Lope or Tirso) and some communities contain more than one author. Of particular interest for our study is the fact that *La Segunda Celestina* seems to be strongly connected and in the same community as several samples of Sor Juana's *Los empeños de una casa*. The second sample of *La Segunda Celestina* also presents a connection with *El Amor al uso* by Antonio de Solís (1610–86), although they are not in the same community (see Fig. 2).

5.2 Strength of authorial signal and determining authorship

Preliminary authorship attribution and verification tests showed very unstable classification results. We performed them using the same methods as described in detail below in relation to smaller, more refined, corpus, but using the full corpus of Spanish Baroque Drama. In the cross-validation with SVM, Delta, and NSC, and verification with the so-called General

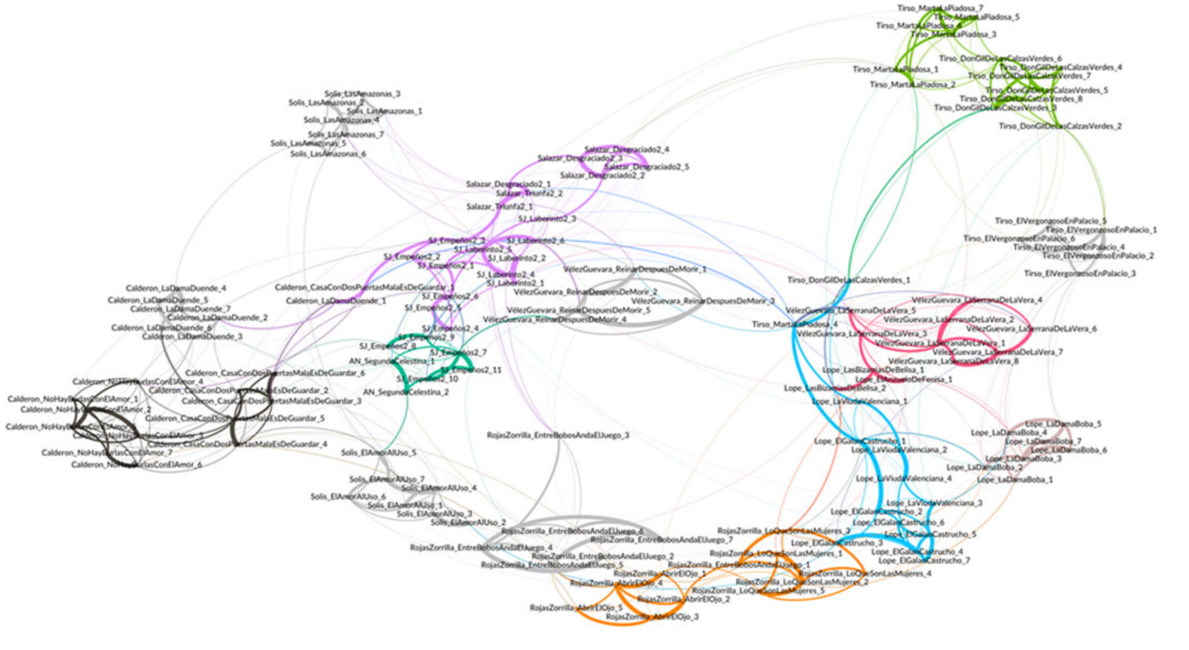

Tirso_MartaLaPiadosa_7
Tirso_MartaLaPiadosa_5
Tirso_MartaLaPiadosa_6
Tirso_MartaLaPiadosa_3
Tirso_DonGilDeLasCalzasVerdes_6
Tirso_MartaLaPiadosa_1
Tirso_DonGilDeLasCalzasVerdes_4
Tirso_DonGilDeLasCalzasVerdes_7
Tirso_MartaLaPiadosa_2
Tirso_DonGilDeLasCalzasVerdes_5
Tirso_DonGilDeLasCalzasVerdes_8
Tirso_DonGilDeLasCalzasVerdes_3
Tirso_DonGilDeLasCalzasVerdes_2
Solis_LasAmazonas_3
Solis_LasAmazonas_2
Solis_LasAmazonas_1
Solis_LasAmazonas_4
Solis_LasAmazonas_7
Solis_LasAmazonas_5
Solis_LasAmazonas_6
Salazar_Desgraciado2_4
Salazar_Desgraciado2_3
Salazar_Desgraciado2_5
Salazar_Desgraciado2_2
Salazar_Desgraciado2_1
Salazar_Triunfa2_2
Salazar_Triunfa2_1
SJ_Laberinto2_3
SJ_Empeños2_3
SJ_Laberinto2_6
SJ_Laberinto2_5
SJ_Empeños2_2
SJ_Laberinto2_2
SJ_Empeños2_1
SJ_Laberinto2_4
SJ_Laberinto2_1
VélezGuevara_ReinarDespuesDeMorir_1
Tirso_ElVergonzosoEnPalacio_5
Tirso_ElVergonzosoEnPalacio_1
Tirso_ElVergonzosoEnPalacio_6
Tirso_ElVergonzosoEnPalacio_4
Tirso_ElVergonzosoEnPalacio_2
Tirso_ElVergonzosoEnPalacio_3
Calderon_LaDamaDuende_4
Calderon_LaDamaDuende_5
Calderon_LaDamaDuende_7
Calderon_LaDamaDuende_2
Calderon_LaDamaDuende_6
Calderon_LaDamaDuende_3
Calderon_CasaConDosPuertasMalaEsDeGuardar_1
Calderon_LaDamaDuende_1
SJ_Empeños2_6
SJ_Empeños2_5
VélezGuevara_ReinarDespuesDeMorir_2
VélezGuevara_ReinarDespuesDeMorir_3
VélezGuevara_ReinarDespuesDeMorir_5
VélezGuevara_ReinarDespuesDeMorir_4
Tirso_DonGilDeLasCalzasVerdes_1
VélezGuevara_LaSerranaDeLaVera_4
VélezGuevara_LaSerranaDeLaVera_5
VélezGuevara_LaSerranaDeLaVera_2
VélezGuevara_LaSerranaDeLaVera_6
Tirso_MartaLaPiadosa_4
VélezGuevara_LaSerranaDeLaVera_3
VélezGuevara_LaSerranaDeLaVera_1
VélezGuevara_LaSerranaDeLaVera_7
VélezGuevara_LaSerranaDeLaVera_8
SJ_Empeños2_4
SJ_Empeños2_9
SJ_Empeños2_8
SJ_Empeños2_7
AN_SegundaCelestina_1
SJ_Empeños2_11
SJ_Empeños2_10
AN_SegundaCelestina_2
Calderon_CasaConDosPuertasMalaEsDeGuardar_6
Calderon_NoHayBurlasConElAmor_4
Calderon_CasaConDosPuertasMalaEsDeGuardar_2
Calderon_NoHayBurlasConElAmor_1
Calderon_CasaConDosPuertasMalaEsDeGuardar_3
Calderon_NoHayBurlasConElAmor_2
Calderon_CasaConDosPuertasMalaEsDeGuardar_5
Calderon_NoHayBurlasConElAmor_5
Calderon_NoHayBurlasConElAmor_3
Calderon_CasaConDosPuertasMalaEsDeGuardar_4
Calderon_NoHayBurlasConElAmor_7
Calderon_NoHayBurlasConElAmor_6
Lope_LasBizarriasDeBelisa_1
Lope_ElAnzueloDeFenisa_1
Lope_LasBizarriasDeBelisa_2
Lope_LaViudaValenciana_1
Lope_LaDamaBoba_5
Lope_LaDamaBoba_4
Lope_LaDamaBoba_7
Lope_LaDamaBoba_6
Lope_ElGalanCastrucho_1
Lope_LaViudaValenciana_2
Lope_LaDamaBoba_3
Lope_LaDamaBoba_2
Lope_LaDamaBoba_1
RojasZorrilla_EntreBobosAndaElJuego_3
Lope_LaViudaValenciana_4
Solis_ElAmorAlUso_5
Solis_ElAmorAlUso_7
Solis_ElAmorAlUso_4
Solis_ElAmorAlUso_6
Solis_ElAmorAlUso_1
Solis_ElAmorAlUso_3
Solis_ElAmorAlUso_2
Lope_LaViudaValenciana_3
Lope_ElGalanCastrucho_2
Lope_ElGalanCastrucho_6
RojasZorrilla_EntreBobosAndaElJuego_6
RojasZorrilla_EntreBobosAndaElJuego_7
RojasZorrilla_LoQueSonLasMujeres_3
Lope_ElGalanCastrucho_5
Lope_ElGalanCastrucho_3
RojasZorrilla_EntreBobosAndaElJuego_4
RojasZorrilla_LoQueSonLasMujeres_1
Lope_ElGalanCastrucho_4
Lope_ElGalanCastrucho_7
RojasZorrilla_EntreBobosAndaElJuego_2
RojasZorrilla_EntreBobosAndaElJuego_5
RojasZorrilla_EntreBobosAndaElJuego_1
RojasZorrilla_AbrirElOjo_1
RojasZorrilla_LoQueSonLasMujeres_4
RojasZorrilla_AbrirElOjo_4
RojasZorrilla_LoQueSonLasMujeres_2
RojasZorrilla_LoQueSonLasMujeres_5
RojasZorrilla_AbrirElOjo_2
RojasZorrilla_AbrirElOjo_5
RojasZorrilla_AbrirElOjo_3

Figure 1. Network of the *comedia de capa y espada* works in the corpus. Each work divided into 2,000 word normal samples, 300–1,000 MFW used as features (We decided to use these parameters following the most reliable results obtained in the next section), applied Louvain community detection algorithm of resolution 1, as implemented in Gephi, the colours match the most distinguishable division into clusters

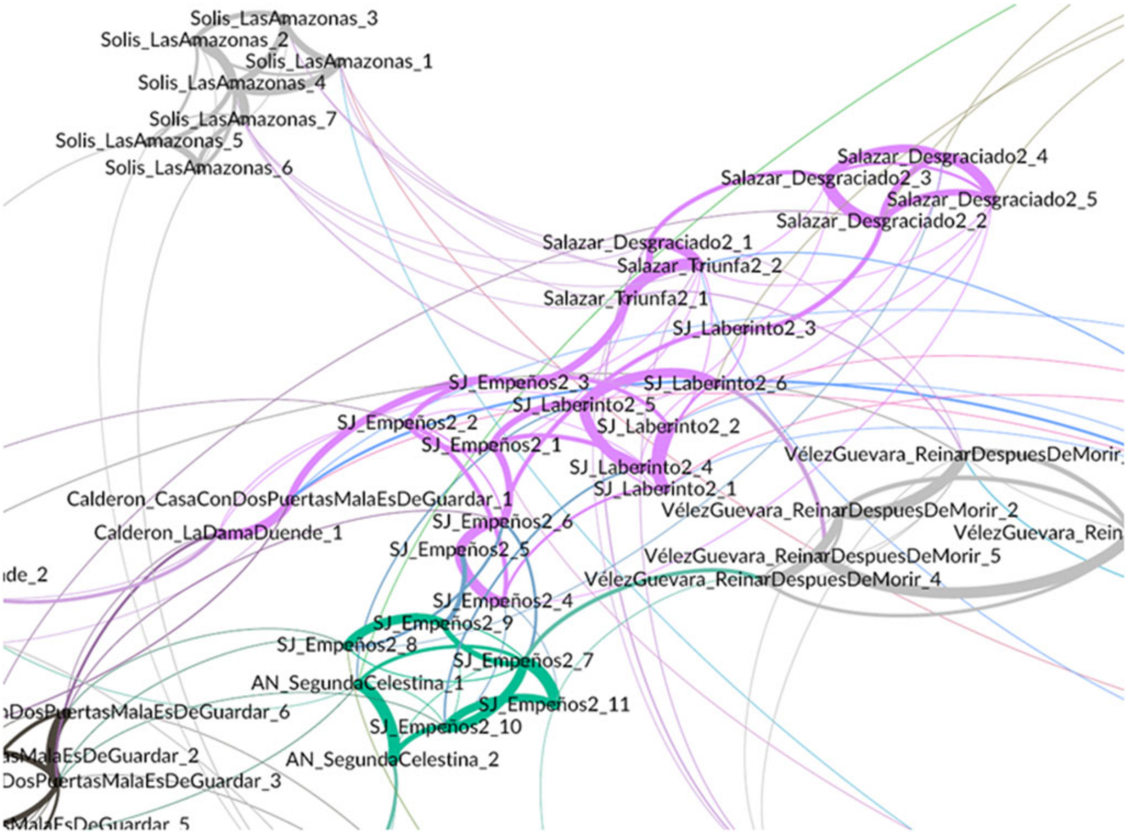


Figure 2. Zoom-in area of the previous network with discussed works

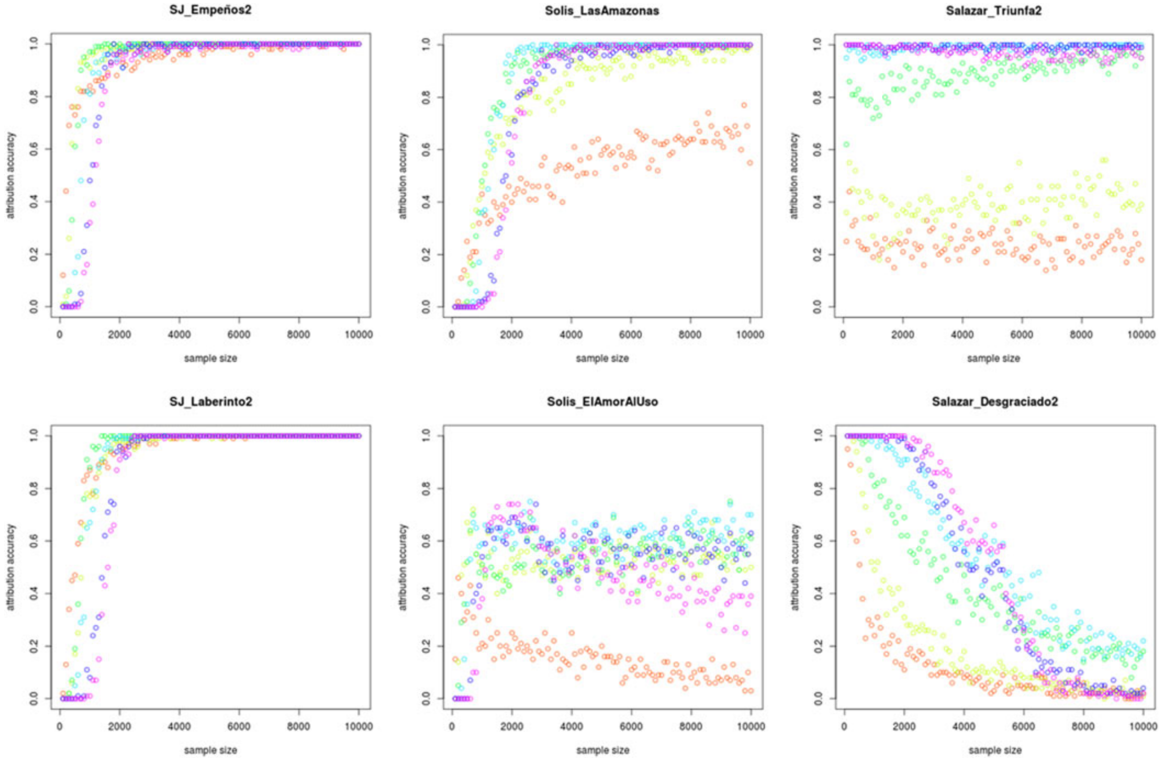
Imposters method (Koppel and Winter, 2014; Kestemont *et al.*, 2016b), varying on settings, a number of candidates were recognized as the author of the anonymous part—from Calderón and Moreto to Sor Juana, Solís, and de Vera Tassis. Since authors such as Calderón could not have possibly authored the text in question, we recognized that the size of the corpus and stylistic dominance of some authors led to significant noise hiding possible real authorships.

We thus decided to examine the strength of particular authorial signals in our corpus, which led us to exclude those who could not author the anonymous part for objective reasons such as the time of its creating (e.g. Lope) or being hub authors—strongly connected to every text in the corpus (Moreto). Inspired by Eder's (2017a) evaluation of authorial signal in short samples and thanks to his courtesy in making the script from the study available to us, we conducted a series of evaluation tests on our corpus until we were left with two authors beside Salazar: Sor Juana and Solís.

The method proposed by Eder (2017a) performs a series of leave-one-out classifications using increasing number of features and length of samples. Therefore, it allows to precisely examine the performance of a particular feature vector for varying amounts of data, and observe which settings are most likely to provide an accurate and reliable classification.

Of the three considered authors, Sor Juana had the most stable signal, showing very strong authorial signal for all examined MFWs feature vectors even in samples as short as 2,000 words or slightly less (see Figs 3–8). Solís had a clearer signal than Salazar, but not as strong as Sor Juana, with poor accuracy for 100 MFW (around 60%) and best results for 300–1,000 MFW from 2,000 word samples. Salazar exhibits a very chaotic signal, which performs fairly badly no matter how big the sample, and provides acceptable accuracy results only for *Más triunfa el Amor rendido*, with the use of 500–1,000 MFW features (see Fig. 5).

In the final part of our examination, we once again performed cross-validated classification and verification on the small corpus consisting of one-genre works by the mentioned three authors against the anonymous part of the text. As the anonymous part is only 4,863 words and some of the texts in our corpus are longer, we opted for using sequential sampling of 1,000 and



Figures 3–8. Accuracy of recognition of particular authors (Sor Juana, Solís and Salazar) by classification algorithm. Colours represent the results of different vectors of most frequent words: 100 (red), 200 (yellow), 300 (green), 500 (cyan), 750 (blue) and 1000 (violet).

2,000 word samples scenarios to account for the representativeness of MFW distribution in each sample, although we also performed the classification for the whole unsampled texts. While we again performed the classification with three algorithms, SVM, NSC, and Delta, we found the SVM results most reliable and consistent, and given also the general good performance of this method in stylometry proved in other studies such as Koppel and Schler (2004), Luyckx and Daelemans (2008), and Koppel *et al.* (2009), we focus on reporting these results in detail.

We run classification tests considering one text by each of the three candidate authors in the training set, verifying that, as anticipated, the texts determined in the previous step as best reflecting the author's individual style proved to perform best as the training material. In fact, as can be examined in Supplemental Materials in detail (see the Appendix), the selection of less representative texts to include in the training corpus led to classifier misrecognizing even author's own plays and significantly changed the accuracy of classification (e.g. for samples of 1,000 words and classification with SVM algorithm, the results of differently setup training sets provide accuracy of 62.9–68.4% for the texts with less clear authorial signal, and 81.9% for the texts recognized as having the highest classification power, that is being most representative of the author).

Across all the scenarios, Sor Juana was the main author candidate for the authorship of the anonymous part of *La Segunda Celestina*, with Solís overtaking in some of the samples depending on the number of features used in classification. The general attributive success with SVM varied from 66.7% for non-sampled texts to 81.9% for samples of 1,000 words and 90% of samples of 2,000 words, as presented in Table 1. NSC performed similarly, with 78.6% accuracy for 1,000 word samples and 89% accuracy for 2,000 word samples, while Delta overperforms in non-sampled texts (accuracy of 100%), underperforms in 1,000 word samples (73.3%), and similar performance in 2,000 word samples (92%). Importantly, while there are slight differences in who is recognized as the author of particular samples across methods and features used, all of them show Sor Juana's stylistic dominance over the first and Solís's dominance over the second part of the studied anonymous part, which strengthens our trust in the reliability of this result (see the Appendix).

As presented in Table 1, the highest accuracy of classification is obtained through the use of 2,000 word

Table 1. Results of the classification tests performed with 100–500 MFW range as features across different classifiers and sampling. All tests were conducted in stylo

Method	Sampling	General attributive success across the corpus (%)
SVM	Non-sampled	66.7
	Samples of 1,000 words	81.9
	Samples of 2,000 words	90
Delta	Non-sampled	100
	Samples of 1,000 words	73.3
	Samples of 2,000 words	92
NSC	Non-sampled	–
	Samples of 1,000 words	78.6
	Samples of 2,000 words	89

samples, and the results are most consistent for the 300–500 MFW range, across all sampling scenarios, in agreement with the ranges of MFW more reliable for detecting Salazar's or Solís' authorial signal. The fact that best results are achieved with 2,000 word samples seems to indicate that the total size or length of the plays influences the results, since non-sampled texts perform worse than sampled ones. To this respect, the best performance of 2,000 word samples agrees with Eder's (2017a) results in his last study about sampling and authorship attribution.

The obtained results point strongly towards the high likelihood of Sor Juana's authorship of the anonymous part, although the influence of Solís on the second half of the anonymous part, especially the 3,001–4,000 frame seems quite interesting as well. Interestingly, parts of works by other authors were consistently misclassified as Sor Juana, which might indicate either/both her domineering style or her taking inspiration from either of the authors, of whose works she must have been aware. Our results also indicate that Salazar has a terribly weak authorial signal, extremely difficult to identify, even in a smaller corpus, and he frequently gets misclassified.

While we initially performed similar classification inquiry for our bigger corpus, we found that results were very bad for sampled texts, with accuracy of 40–50% and strong over-representation of Sor Juana's and Tirso's influence on some of the authors—which led us to staying with this smaller corpus of authors that logistically seemed more likely to have authored the text.

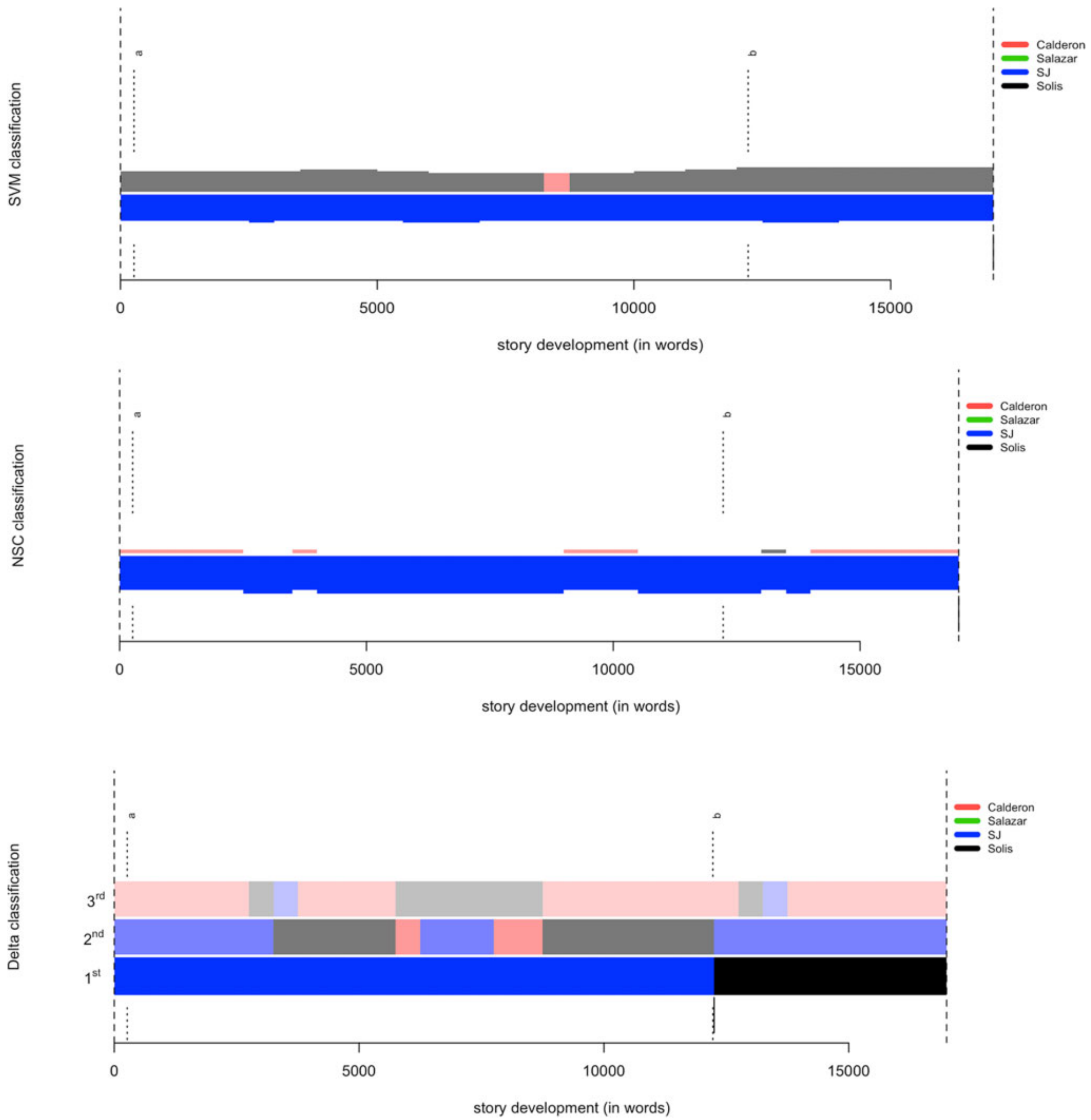
5.3 Editorial influence in the non-anonymous part and the ending

This problem of authorship requires detecting multiple authorial voices as we know for sure Salazar wrote a significant part of the play before another person finished it, which is why we apply Rolling Classify (Eder, 2016) to detect authorial takeovers. This allows to discover both who authored the ending, and if this author made significant changes to the rest of the play. We marked two important points in the whole play. Mark 'b' represents the place at which, according to Vera Tassis, Salazar left the play unfinished. As it can be observed in Figs 9–11, the ending is attributed to Sor Juana in SVM and NSC, and to Solís in Delta, in line with results in previous sections. The most surprising thing is that Salazar's signal is not detected at all, which may be related to the weakness of his signal detected in previous analyses, and Sor Juana seems to dominate the rest of the play. Could it be that, if it was Sor Juana who finished the play, she altered the rest of the text to the extent that we are not able to see Salazar anymore? This would confirm the claims of some of the scholars who defend her authorship (Paz, 1990; Schmidhuber de la Mora, 2016).

We also marked a scene in the beginning which portrays the first encounter between protagonists: doña Beatriz and don Juan, with the mark 'a'. It is retold and alluded to several times in the play, something unusual in the Golden Age theatre. In this first encounter, doña Beatriz is hunting in the forest when don Juan, a stranger, starts following her. She reacts by threatening to shoot him if he doesn't leave her alone, and disregards his excuses of having fallen in love with her at first sight. This is a very feminist confrontation for a 17th century text, in which we see a strong and resolute female character who confronts a male one. The actual first encounter is later in the text belittled in don Juan's retelling, who just says that doña Beatriz ran to the safety of her servants. It seems to betray a female writer which is why we examined it. As it can be observed in Figs 9–11, in all tests this fragment is attributed to Sor Juana. This is even more interesting if we consider that Sor Juana could be reflecting a biographical experience, as she had to reject many admirers in her youth, when she was serving the Viceroy's wife at the palace. She probably felt identified with doña Beatriz's situation and gave her the attitude and strength of character she herself used in that situation.

6 Conclusions

In this article, we examined the authorship of an anonymous part of *La Segunda Celestina*, a long disputed problem in Spanish Baroque Drama studies, with stylometric methods. We have determined that like the literary evidence, quantitative analysis strongly points to the author being Sor Juana Inés de la Cruz, and have found traces of her stylistic influence across the play, supporting the theory of her being the author of the anonymous part and editor of the whole text. While we consider this theory confirmed, we also observe the complex and blurry position of this text within the



Figures 9–11. Rolling SVM, NSC and Delta on *La Segunda Celestina*. 500 MFW and 5000 words per slice.

Spanish Baroque Drama corpus, as the broader perspective points to the importance of the text within it and numerous links to other authors. This said, it must be explained that for similarly dated literature it is impossible to obtain a perfectly clean and balanced dataset and any definite answers, while we can hope that it will sometime be possible to verify our findings on a more detailed canvas, realistically it is unlikely that a significantly more precise perspective of the text within the Spanish Baroque Drama can be produced. In this sense, our study shows the difficulties in detecting even candidates and collaborative authorship because of the unclear situation at the time and the rather poor historical data.

The second best candidate, whose influence is present well above chance level especially in the second half of the anonymous part, Solís, is a new discovery, and his possible relation to *La Segunda Celestina* and Sor Juana, in terms of influence or themes, or even as a possible co-author, collaborator or editor, may also be of interest to future studies. Although Solís had left the court and was devoted to writing the *History of the conquest of Mexico* in the years the play was written, historical data do not preclude a possible intervention by this playwright in the text. Therefore, his emergence in our results will need to be analyzed by Golden Age Spanish Drama experts.

One of the biggest challenges of the project was creating a representative corpus of possible authors, and adjusting it to the specific problem. Our experience emphasizes the need for and usefulness of taking corpus evaluation steps in all analyses, and especially in the case of historic works, for which it is impossible to create a truly balanced corpus. In this study, we applied a novel method of determining the strength of authorial signal in the works of specific authors, as developed by Eder in 2017, which, together with other, more traditional, steps of the analysis helped us narrow down the corpus and, especially, set up the training set for the classification to include most representative and clear voices of the candidate authors. This stage is of crucial importance for authorship studies, since the selection of less representative texts led to errors in the recognition of undoubted texts. We found the method highly useful, as it allowed us to make our training set choices more reliably, and to show that between the worst and best setup the difference in accuracy of classification amounted to 19%. On top of that, the application of Rolling Classify and sequential analysis to assess different samples of the entire text was of crucial importance for determining the extent of Sor Juana's intervention, with results showing that there is little of Salazar's original redaction left.

Notes

1. A 'suelta' is a printed version of the play for the public. The suelta found by Schmidhuber was probably printed for the public accompanying the performance of the comedy at the Royal Palace in 1676 (Sabat de Rivers, 1992, p. 493).
2. This is an English translation of the original Spanish text: 'Un poema que dejó sin acabar Don Agustín de Salazar, y perficionó con graciosa propriedad la poetisa, cuyo original guarda la estimación discreta de Don Francisco de las Heras, caballero del orden de Santiago, regidor de esta villa, y por ser proprio del primer tomo, no le doy estampa en este libro, y se está imprimiendo para representarse a sus majestades'.
3. This play was performed for the first time in October 1683 in Mexico. The aforementioned passage is a dialogue between Muñiz and Arias in the second one-act farce—one of the minor performances intercalated in the main play.
4. Additionally, there is another fragment in the same passage which includes mentions to the very young writer of the play:

   Diósela un estudiante
   que en las comedias es tan principiante,
   y en la poesía tan mozo,
   que le apuntan los versos como el bozo.

   Sor Juana would be twenty-four or twenty-seven when the ending was written and if she was the author of the ending, it would be her first play.
5. The original 'suelta' displays only the name of Salazar as author. It starts with a short theatrical laudatory piece known as 'loa', dated in 1675, and written by Salazar. The comedy is dated in 1676.
6. After Sabat de Rivers' study, other papers on *La Segunda Celestina* have appeared, such as the one by Schmidhuber de la Mora (1994), but no progress was achieved in this polemic attribution problem.
7. Antonio de Toledo y Salazar (1622–1715), Marquis of Mancera, was the viceroy of New Spain from 1664 to 1673. After that, he returned to the Spanish court to assist Queen Mariana de Austria as her senior butler, a position he had in the years when *La Segunda Celestina* was written, completed, and performed at the Royal Palace for the Queen. He probably was in charge of organizing the Queen's birthday celebrations. At his time as viceroy of New Spain, a young Sor Juana was assisting at the court and she became very close to him, and even more so to his wife, Leonor de Carreto, marchioness of Mancera. From that moment on, the marquises, highly impressed with Sor Juana's intelligence, became her protectors and mecenas. Sabat de Rivers then considers it very probable that when the marquis of Mancera found himself without a play to present at the Queen's birthday, due to Salazar's death, he thought of Sor Juana to finish it (Sabat de Rivers, 1992).
8. This other version was published under the title *El encanto es la hermosura o el hechizo sin hechizo* as part of the complete works by Salazar, *Cítara de Apolo* (Madrid, 1681), edited by Vera Tassis. In his prologue, Vera Tassis declares that *El encanto es la hermosura* was the title chosen by Salazar for the comedy and specifies the exact line where Salazar left the play unfinished (l. 2508) and that the rest of the play was written by himself.
9. O'Connor (1992) does not agree with previous scholars on when the play was performed. He defends that the play performed at the Queen's birthday in 1676 was *El encanto es la hermosura*, that is, the play by Salazar with the ending written by Vera Tassis. *La Segunda Celestina*, with the anonymous ending presumably by Sor Juana, would have been performed at a later date both in Madrid and New Spain. O'Connor (1994) defends this hypothesis also in his edition of *El encanto es la hermosura* and *La Segunda Celestina*—that is, the play with the two different endings. Sabat de Rivers, however, replied to him in a new paper (Sabat de Rivers, 1997) in which she insists that there is no documentation to prove O'Connor's suspicion, and instead, the loa of *La Segunda Celestina* had the printed date of 1675, and the comedy of 1676. She considers that this proves that *La Segunda Celestina* is the first version of the play. Thus, Vera Tassis' participation in the comedy would be at a later date.
10. http://www.cervantesvirtual.com/portales/sor_juana_ines_de_la_cruz/ (accessed 5 February 2021).
11. We follow Sahle's (2016) observation that digitized print edition is not a strict digital edition, as explained here: https://www.digitale-edition.de/exist/apps/editions-browser/about.html (accessed 8 February 2021).

12. http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/index.html (accessed 5 February 2021).
13. https://transkribus.eu/Transkribus/ (accessed 5 February 2021)
14. Important advances in this direction can be found at BIESES project, which provides a database of Spanish women writers: https://www.bieses.net/ (accessed 5 February 2021).
15. The exact percent of this agreement in the bootstrap consensus tree should be determined by the researcher, and, in the lack of state-of-the-art recommendations, we followed general practice recommended by Eder (2017b) and our intuition in using 50% threshold. However, the resulting stylometric network will retain not only these strongest connections, but also other ones, with the aim of producing a general overview of stylistic relations between texts.

Acknowledgments

The authors would like to thank professors Maciej Eder (Computational Stylistics Group, Institute of Polish Language) and Juan Montero (University of Seville) for their inspiring comments and suggestions on this project, professor Gema Areta (University of Seville) for her insights on Latin American Literature and Sor Juana's hand and style, and José Calvo (State and University Library of Göttingen) for sending us his clean plain text-converted version of the CANON-60 corpus.

Funding

L.H.-L. was supported by the University of Seville (VIPPIT-2018-I.3), and the Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics project (SONATA-BIS 2017/26/E/HS2/01019) funded by the Polish National Science Center (NCN), for whose support she is very grateful. In last stages of this article, she was supported by a Margarita Salas' postdoctoral fellowship at the University of Seville, funded by the European Union (NextGenerationEU), and the projects *Hacia la Institucionalización Literaria: Polémicas y Debates Historiográficos (1500–1844)* (RTI2018-095664-B-C22), and *Comunidades femeninas y escritura en la España de la primera edad moderna* (PID2019-106471GB-I00), both funded by the Spanish National Science Center (*Ministerio de Ciencia, Innovación y Universidades*). J.B. was supported by the Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics project (SONATA-BIS 2017/26/E/HS2/01019) funded by the Polish National Science Center (NCN), for whose support she is very grateful.

Appendix

See our corpus, list of plays and complementary materials in our GitHub (DOI: 10.5281/zenodo.5879010): https://github.com/JoannaBy/La-Segunda-Celestina.

Table A1. Results of the classification tests performed with 100–500 MFW range as features across different classifiers and non-sampled texts. All tests were conducted in stylo

Recognized author of the anonymous part				
Number of sample	MFW used	Author recognized with SVM (accuracy 66.7%)	Author recognized with Delta (accuracy 100%)	NSC
Text	100 MFW	Salazar	Salazar	–
	200 MFW	Solís	Solís	–
	300 MFW	SJ	Solís	–
	400 MFW	SJ	Solís	–
	500 MFW	SJ	Solís	–

Table A2. Results of the classification tests performed with 100–500 MFW range as features across different classifiers and samples of 1,000 words. All tests were conducted in stylo

Recognized author of the anonymous part				
Number of sample	MFW used	Author recognized with SVM (accuracy 81.9%)	Author recognized with Delta (accuracy 73.3%)	Author recognized with NSC (accuracy 78.6%)
Sample 1	100 MFW	Solís	Sor Juana	Sor Juana
	200 MFW	Solís	Sor Juana	Sor Juana
	300 MFW	Sor Juana	Sor Juana	Sor Juana
	400 MFW	Sor Juana	Sor Juana	Sor Juana
	500 MFW	Sor Juana	Sor Juana	Sor Juana
Sample 2	100 MFW	Sor Juana	Sor Juana	Sor Juana
	200 MFW	Sor Juana	Sor Juana	Sor Juana
	300 MFW	Sor Juana	Sor Juana	Sor Juana
	400 MFW	Sor Juana	Sor Juana	Sor Juana
	500 MFW	Sor Juana	Sor Juana	Sor Juana
Sample 3	100 MFW	Solís	Solís	Solís
	200 MFW	Solís	Solís	Solís
	300 MFW	Solís	Solís	Solís
	400 MFW	Solís	Solís	Solís
	500 MFW	Sor Juana	Solís	Solís
Sample 4	100 MFW	Solís	Sor Juana	Solís
	200 MFW	Solís	Sor Juana	Solís
	300 MFW	Solís	Solís	Solís
	400 MFW	Sor Juana	Solís	Solís
	500 MFW	Sor Juana	Solís	Solís

Table A3. Results of the classification tests performed with 100–500 MFW range as features across different classifiers and samples of 2,000 words. All tests were conducted in stylo

Recognized author of the anonymous part				
Number of sample	MFW used	Author recognized with SVM (accuracy 90%)	Author recognized with Delta (accuracy 92%)	Author recognized with NSC (accuracy 89%)
Sample 1	100 MFW	Sor Juana	Sor Juana	Sor Juana
	200 MFW	Sor Juana	Sor Juana	Sor Juana
	300 MFW	Sor Juana	Sor Juana	Sor Juana
	400 MFW	Sor Juana	Sor Juana	Sor Juana
	500 MFW	Sor Juana	Sor Juana	Sor Juana
Sample 2	100 MFW	Solís	Solís	Solís
	200 MFW	Solís	Salazar	Solís
	300 MFW	Solís	Solís	Solís
	400 MFW	Sor Juana	Solís	Solís
	500 MFW	Solís	Solís	Solís

References

Alatorre, A. (1990). 'La Segunda Celestina' de Agustín de Salazar y Torres: Ejercicio de crítica. *Vuelta*, **169**: 46–52.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks, *Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362.

Bia, A. and Pedreño, A. (2001). The Miguel de Cervantes digital library: the Hispanic voice on the web. *Digital Scholarship in the Humanities*, **16**(1): 161–77.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**: P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Burrows, J. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Oxford University Press.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Cafiero, F. and Camps, J.-B. (2019). Why Molière most likely did write his plays. *Science Advances*, **5**(11). https://doi.org/10.1126/sciadv.aax5489

Calvo Tello, J. (2019). Delta inside Valle-Inclán: stylometric classification of periods and groups of his novels. In Rißler-Pipka, N. (ed.), *Romanische Studien. Theorien von Autorschaft Und Stil in Bewegung: Stilistik Und Stilometrie in Der Romania*. Múnich: AVM.edition, pp. 151–64.

Calvo Tello, J., Schlör, D., Henny, U., and Schöch, C. (2017). *Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels, Digital Humanities 2017. Conference Abstracts*, McGill University and Université de Montréal, August 2017. https://dh2017.adho.org/abstracts/037/037.pdf (accessed 7 May 2021).

de la Cruz, S. J. I. (1700). *Fama y Obras Posthumas*. Imprenta de Manuel Ruiz de Murga. http://catalogo.bne.es/uhtbin/cgisirsi/0/x/0/05?searchdata1=Mima0000129454 (accessed 7 May 2021).

de la Cruz, S. J. I. (1692). *Segundo Volumen de las Obras de Soror Juana Ines de la Cruz, Monja Profesa en el Monasterio del Señor San Gerónimo de la Ciudad de Mexico*. Vols. **1 and 2**. por Tomás López de Haro ... http://catalogo.bne.es/uhtbin/cgisirsi/0/x/0/05?searchdata1=bima0000003357 (accessed 7 May 2021).

Eder, M. (2011). Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint. *Studies in Polish Linguistics*, **6:** 99–114.

Eder, M. (2017a). *Short Samples in Authorship Attribution: A New Approach. Digital Humanities 2017: Conference Abstracts*. Montreal: McGill University, pp. 221–24. https://dh2017.adho.org/abstracts/341/341.pdf (accessed 7 May 2021).

Eder, M. (2017b). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, **32**(1): 50–64.

Eder, M. (2016). Rolling stylometry. *Digital Scholarship in the Humanities*, **31**(3): 457–69.

Eder, M. (2013). Computational stylistics and biblical translation: how reliable can a dendrogram be? In Piotrowski, T. and Grabowski, L. (eds), *The Translator and the Computer*. Wroclaw: WSF Press, pp. 155–70.

Eder, M., Kestemont, M., and Rybicki, J. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, **16**(1): 107–21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html (accessed 8 May 2021).

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, **32**(Suppl_2): ii4–16. https://doi.org/10.1093/llc/fqx023

Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., Byszuk, J., and Rybicki, J. (2018). Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm. *Frontiers in Digital Humanities*, **5**. https://doi.org/10.3389/fdigh.2018.00004.

Gregg, K. C. (1977). Towards a definition of the 'Comedia de capa y espada'. *Romance Notes*, **18**(1): 103–6.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, **22:** 251–70. https://doi.org/10.1093/llc/fqm020

Hernández-Lorenzo, L. (2019). Fernando de Herrera y la autoría de Versos. Un primer acercamiento al drama textual desde la Estilometría. In Rißler-Pipka, N. (ed.), *Romanische Studien. Theorien von Autorschaft Und Stil in Bewegung: Stilistik Und Stilometrie in Der Romania*. Múnich: AVM.edition, pp. 75–90.

Jannidis, F., Pielström, S., Schöch, C., and Vitt, T. (2015). *Improving Burrows' Delta—An Empirical Evaluation of Text Distance Measures. Digital Humanities 2015 Conference Abstracts*. Sydney: ADHO. http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows_Delta__An_empi/JANNIDIS_Fotis_Improving_Burrows_Delta_ (accessed 19 November 2018).

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution.

Literary and Linguistic Computing, **25**: 215–23. https://doi.org/10.1093/llc/fqq001

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017). *Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents, 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, November 2017, pp. 19–24. https://doi.org/10.1109/ICDAR.2017.307

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., and Daelemans, W. (2016a). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, **63**: 86–96.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., and Daelemans, W. (2016b). *Authorship Verification with the Ruzicka Metric. Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University and Pedagogical University, pp. 246–49. http://dh2016.adho.org/abstracts/402 (accessed 8 May 2021).

Koppel, M. and Schler, J. (2004). *Authorship Verification as a One-Class Classification Problem. Twenty-First International Conference on Machine Learning—ICML '04*. https://doi.org/10.1145/1015330.1015448

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26. https://doi.org/10.1002/asi.20961

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, **65**(1): 178–87. doi:10.1002/asi.22954; http://dx.doi.org/10.1002/asi.22954

Lutosławski, W. (1897). *The origin and growth of Plato's logic: With an account of Plato's style and of the chronology of his writings*. London: Longmans, Green & Co.

Luyckx, K., and Daelemans W. (2008). Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester: Coling 2008 Organizing Committee, pp. 513–20.

Méndez Plancarte, A. and Salceda, A. G. (eds) (1957). *Obras Completas de Sor Juana Inés de la Cruz. IV*. México: Fondo de Cultura Económica.

Mosteller, F. and Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Muehlberger, G., Seaward, L., Terras, M., *et al.* (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, **75**: 954–76. https://doi.org/10.1108/JD-07-2018-0114

Newman, M. E. J. (2006) Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(23): 8577–82.

Ochab, J. K., Byszuk, J., Pielström, S., and Eder, M. (2019). *Identifying Similarities in Text Analysis: Hierarchical Clustering (Linkage) versus Network Clustering (Community Detection). Digital Humanities* **2019**: *Book of Abstracts*. Utrecht. https://dev.clariah.nl/files/dh2019/boa/0981.html

Ochab, J. K. and Essler H. (2019). *Stylometry of Literary Papyri. Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. Association for Computing Machinery. https://doi.org/10.1145/3322905.3322930

O'Connor, T. A. (ed.) (1994). *Agustín de Salazar y Torres. Juan de Vera Tassis y Villarroel. Sor Juana Inés de la Cruz. El Encanto es la Hermosura y el Hechizo sin Hechizo. La Segunda Celestina*. Binghamton, New York: Medieval and Renaissance Texts and Studies.

O'Connor, T. A. (1992). Los Enredos de una Pieza. El Contexto Histórico-Teatral de *El Encanto es la Hermosura* o *La Segunda Celestina* de Salazar y Torres, Vera Tassis y Sor Juana. *Literatura Mexicana*, **3**(2): 283–303.

Oleza Simó, J. (2014). *Canon 60*. Valencia: Universitat de València.

Pascual Buxó, J. (1991). Las vueltas de Sor Juana. *Nuevo Texto Crítico*, **7**: 197–204. https://doi.org/10.1353/ntc.1991.0023

Paz, O. (1990). ¿Azar o justicia? *Proceso*, **710**: 53. https://www.proceso.com.mx/155130/azar-o-justicia (accessed 19 November 2018)

Sabat de Rivers, G. (1992). Los problemas de la Segunda Celestina. *Nueva Revista de Filología Hispánica*, **XL**(1): 493–512.

Sabat de Rivers, G. (1997). Una nueva edición de *El encanto es la hermosura y El hechizo sin hechizo/La segunda Celestina*. *Bulletin of Hispanic Studies*, **74**(3): 311–9.

Sahle, P. (2016). 2. What is a scholarly digital edition? In Driscoll, M. J. and Pierazzo, E. (eds), *Digital Scholarly Editing: Theories and Practices*. Cambridge: Open Book Publishers. http://books.openedition.org/obp/3397 (accessed 9 May 2021).

Sánchez Arteche, A. (1991). *La Segunda Celestina. Una Comedia que no Escribió Sor Juana*. México: Presencia.

Schmidhuber de la Mora, G. and Peña Doria, O. M. (eds) (1990). *Sor Juana Inés de la Cruz. Agustín de Salazar y Torres. La Segunda Celestina Una comedia perdida de Sor Juana*. México: Vuelta.

Schmidhuber de la Mora, G. (1991). 'La Segunda Celestina': Sor Juana y la estilometría. *Vuelta*, **15**(174): 54–60.

Schmidhuber de la Mora, G. (1994). La Segunda Celestina: Hallazgo de una comedia de Sor Juana Inés de la Cruz y Agustín de Salazar. In Ortega, J. and Amor y Vázquez, J. (eds), *Conquista y Contraconquista: La Escritura del Nuevo Mundo (Actas del XXVIII Congreso del Instituto Internacional de Literatura Iberoamericana)*. México: El Colegio de México—Brown University, pp. 315–24.

Schmidhuber de la Mora, G. (2016). *Hallazgo de una Comedia Perdida de Sor Juana: La Gran Comedia de la Segunda Celestina*. Alicante: Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/portales/sor_juana_ines_de_la_cruz/obra/hallazgo-de-una-obra-perdida-de-sor-juana-la-gran-comedia-de-la-segunda-celestina/ (accessed 9 May 2021).

Smith, P. W. H. and Aldridge, W. (2011). Improving authorship attribution: optimizing Burrows' Delta method. *Journal of Quantitative Linguistics*, **18**(1): 63–88.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, **21**(2): 421–39. https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/7 (accessed 9 May 2021).

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, **60**: 538–56.

Stamou, C. (2008). Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, **23**(2): 181–99. https://doi.org/10.1093/llc/fqm029

Traag, V., Waltman L., and van Eck N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, **9**(1): 5233. https://doi.org/10.1038/s41598-019-41695-z