A Stylometric Analysis of *Amadís de Gaula* and *Sergas de Esplandián*

Yoshifumi Kawasaki
The University of Tokyo
ykawasaki@g.ecc.u-tokyo.ac.jp

Abstract

Amadís de Gaula (AG) and its sequel *Sergas de Esplandián* (SE) are masterpieces of medieval Spanish chivalric romances. Much debate has been devoted to the role played by their purported author Garci Rodríguez de Montalvo. According to the prologue of AG, which consists of four books, the author allegedly revised the first three books that were in circulation at that time and added the fourth book and SE. However, the extent to which Montalvo edited the materials at hand to compose the extant works has yet to be explored extensively. To address this question, we applied stylometric techniques for the first time. Specifically, we investigated the stylistic differences (if any) between the first three books of AG and his own extensions. Literary style is represented as usage of parts-of-speech n-grams. We performed principal component analysis and k-means to demonstrate that Montalvo's retouching on the first book was minimal, while revising the second and third books in such a way that they came to moderately resemble his authentic creation, that is, the fourth book and SE. Our findings empirically corroborate suppositions formulated from philological viewpoints.

1 Introduction

Amadís de Gaula (AG), which is a medieval Spanish chivalric romance published at the beginning of the sixteenth century, has long been considered a masterpiece of the genre. Its sequel *Sergas de Esplandián* (SE) came out a few years after. Both works have been attributed to Garci Rodríguez de Montalvo, a lower-class aristocrat from Medina del Campo in the present-day Valladolid prefecture. Note that no other work has been ascribed to him.

AG consists of four books. Together with its sequel SE, there are a total of five books in the series, even though the latter was published separately. According to the prologue of AG, the author revised the first three books that were in circulation at that time, and *translated* the fourth book and SE from a Greek manuscript he had encountered. In reality, however, they are both considered his own creation; feigning a *translation* was a literary commonplace back then. Still, the extent to which the author modified the materials at hand to compose the extant version has yet to be extensively explored.

To delve into the enigmatic composition of Montalvo's works, we applied stylometric analysis for the first time, to the best of our knowledge. Stylometry is a field of study that, among other goals, aims to identify authorship of disputed or anonymous documents (Juola, 2006; Grieve, 2007; Zhao and Zobel, 2007; Stamatatos, 2009; Jockers and Witten, 2010). Specifically, we investigated the stylistic differences (if any) between the first three books of AG and his own extensions, that is, the fourth book of AG and SE. Literary style is represented as usage of parts-of-speech (POS) n-grams. Since the employment of syntactic features is supposed to be fairly unconscious and hardly imitable, POS n-grams, which capture partial syntactic information, can reasonably serve as stylistic fingerprints. We performed principal component analysis (PCA) and k-means to demonstrate that Montalvo's retouching on the first book was minimal, while revising the second and third books in such a way that they came to moderately resemble his original contributions, that is, the fourth book and SE. Our findings empirically corroborate suppositions formulated from philological viewpoints by Cacho Blecua (Rodríguez de Montalvo, 2020a).

The rest of the paper is organized as follows. In Section 2, we review related research. Section 3 describes the methodology utilized. In Section 4, we present experimental results, followed by a discussion in Section 5. Section 6 concludes the study by discussing future research directions.

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 1–7
November 20, 2022.

2 Related Work

Research on the genesis of the Amadisian oeuvre has been conducted by Hispanic philologists including Cacho Blecua (Rodríguez de Montalvo, 2020a,b), Domingo del Campo (1982), and Sainz de la Maza (Rodríguez de Montalvo, 2003). However, few scholars have exhaustively inspected the linguistic usage therein. Labrousse (2021) studied a variation of nominal phrases containing possessives in the first and fourth books of AG and found discrepancies between them. However, the second and third books of AG as well as SE were not included in her scope of study. Moreover, the analysis was restricted to the first 500 occurrences of the construction in question. To gain a more complete picture, a comprehensive scrutiny is needed.

Over the past few years, Spanish Philology has witnessed an increasing number of stylometric studies (Fradejas Rueda, 2016; Rißler-Pipka, 2016; de la Rosa and Suárez, 2016; Rojas Castro, 2017; Cerezo Soler and Calvo Tello, 2019; García-Reidy, 2019; Hernández Lorenzo, 2019). The style markers used have been mostly limited to functional words and frequent words. POS *n*-grams have been rarely adopted even though its effectiveness has been confirmed by various studies addressing literary works in multiple languages including English (Koppel et al., 2002; Clement and Sharp, 2003; Juola, 2006; Hirst and Feiguina, 2007; Eder, 2015; Pokou et al., 2016; Savoy, 2017), French (Kocher and Savoy, 2019), Japanese (Uesaka and Murakami, 2015), and recently in Spanish (Kawasaki, 2021).

The advantages of leveraging POS sequences are multi-fold: (i) their numerous occurrences provide reliable statistics; (ii) they are relatively independent from content; (iii) being out of conscious control of the author, they are supposed to be hardly imitable; and (iv) they partially capture syntactic patterns, which have been shown to be reliable style markers (Baayen et al., 1996).

3 Methods

The digitized texts of AG and SE were retrieved from *Corpus of Hispanic Chivalric Romances*[1]. For AG, we used the version published in Seville in 1539 by the printer Juan Cromberger[2]. For SE, we employed the version published in Rome in 1525 by the printers Jacobo de Junta and Antonio de Salamanca[3]. They were the only digitized texts available, although the first edition of AG goes back to 1508 and SE back to 1510.

AG consists of 133 chapters arranged across four books: AG1, AG2, AG3, and AG4. The token size amounts to 530,000 words. SE is composed of 184 chapters forming a single book. The token size adds up to 190,000 words. Since the chapter length varies considerably from one another, we decided to generate equal-length pieces of 10,000 words from respective books. The prologues and epistolary passages were omitted in advance. Note that book division was maintained for the subsequent analyses, while chapter division was disregarded. As for the final part of a book, where the piece length was below 10,000, it was treated as an independent one if it exceeded 6,000 words; otherwise, it was merged into the penultimate piece. Thus, AG1 resulted in 13 pieces, AG2 in 9, AG3 in 11, AG4 in 16, and SE in 18.

For stylistic features, we leveraged POS *n*-grams. The tags were assigned using a tagger designed for present-day Spanish `spaCy 3.3.1`[4]. The model employed was `es_dep_news_trf`, which is larger and more accurate. We utilized this tagger because there are no publicly available ones designed for Medieval Spanish, in which the Amadisian works are written. Based on the philological expertise, we modified extensively the texts prior to tagging to facilitate correct parsing; specifically, we applied as much orthographic modernization as possible. For instance, *auer* "to have" was transformed into its modern counterpart *haber* and certain words were separated, like *acostose* was separated into *acostó se* "he/she lay down".

AUX and PROPN were merged into VERB and NOUN respectively as their correct identification proved to be hardly feasible. For frequent functional words including auxiliary verbs, adverbs, conjunctions, and prepositions, we adopted surface forms in lieu of the assigned tags to make the most of their differing usage, for example, the preposition *de* "of" was not converted into ADP but maintained as such. As for verbs, we distinguished among infinite forms, that is, infinitives (INF), gerunds (GERUND), and past participles (PPART) and gave them distinct labels. In contrast,

[1] https://textred.spanport.lss.wisc.edu/chivalric/index.html

[2] https://textred.spanport.lss.wisc.edu/chivalric/textsoriginal/ama-text.txt

[3] https://textred.spanport.lss.wisc.edu/chivalric/textsoriginal/ag5-text.txt

[4] https://spacy.io/

the finite forms were uniformly given an identical label regardless of mode, tense, grammatical person, and number. In addition, we differentiated highly frequent verbs *haber* "to have" and *ser* "to be" by tagging the relevant forms with their infinitival forms. As for punctuation, we only retained periods and question marks representing sentence boundaries and omitted commas, colons, and semicolons that could stem from editorial interventions. These measures resulted in 54 tag types in total. Tagging performance was evaluated by computing an accuracy rate on randomly chosen five hundred-word passages: one from each of the four books of AG and another from SE. The mean accuracy was almost perfect at 0.99 ± 0.01. Note that, without manual modification of the texts and tags, the mean accuracy declined to 0.83 ± 0.02.

Every piece was represented as a vector whose elements represent z-transformed relative frequencies of the n-grams. We considered only the most frequent POS n-grams above a given rank threshold, while the remainder was aggregated under the label of OTHERS. To assess the robustness of our analyses, we varied the n-gram size n for $n \in \{1, 2, 3, 4\}$ and the rank threshold r for $r \in \{100, 300, 500\}$. For $n = 1$, r was fixed to 54, which was the number of unigram types.

4 Analysis

For illustrative purposes, we present the results obtained with $(n, r) = (3, 300)$. Figure 1 displays the pair-wise distance scores between the pieces, computed as $\sqrt{\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2}{r}}$, where $\boldsymbol{x_i}$ represents the feature vector for the i-th piece. The bluer (redder) the cell, the more (less) similar the pair of pieces. Overall, we observe lower *intra*-book distance scores in contrast to larger *inter*-book ones. However, it is worth noting that the distance scores between AG2 and AG3 are relatively low and that these two books exhibit less dissimilarity with AG4 and SE.

Next, we conducted two types of exploratory multivariate analyses, PCA and k-means, to examine whether any stylistic difference was found across the books.

4.1 PCA

The first two PC scores obtained with $(n, r) = (3, 300)$ are plotted in Figure 2. Contribution ratios for PC1 and PC2 were 16.9% and 8.3%, respectively. PC1 can be reasonably interpreted as a representation of Montalvo's degree of contribution; on the left side are AG4 and SE, which are assumed to be his original creations, on the right side is AG1, which presumably best conserves the primitive appearance, and in between are AG2 and AG3, which were allegedly modified to some degree (Rodríguez de Montalvo, 2020a). PC2, which roughly dissociates AG4 and SE, can be regarded as reflecting Montalvo's internal stylistic variation. That the rest of books are found in between might be ascribed to their different origin, thereby remaining immune to Montalvo's literary style.

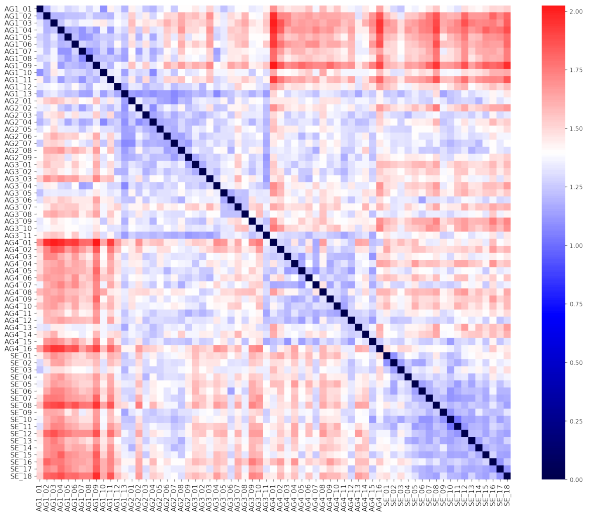

Figure 1: Pair-wise distance scores between the pieces computed with $(n, r) = (3, 300)$. The bluer (redder) the cell, the more (less) similar the pair of pieces.

4.2 *k*-means

We conducted k-means clustering using `sklearn.cluster.KMeans` with default setting (Pedregosa et al., 2011)[5]. The number of clusters k was varied for $k \in \{2, 3, 4, 5\}$. As the algorithm was sensitive to the initial centroids selected, we ran it 100 times and computed the mean concordance rate, which was defined as the average number of times a pair of pieces was classified into the same cluster. We supposed that no clear-cut pattern would emerge without stylistic differences across the books.

Figure 3 illustrates the pair-wise mean concordance rates obtained with $(k, n, r) = (2, 3, 300)$. The darker the cell, the more often the pair of pieces belonged to the same cluster and are judged as similar. We can discern two clusters, one formed by

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

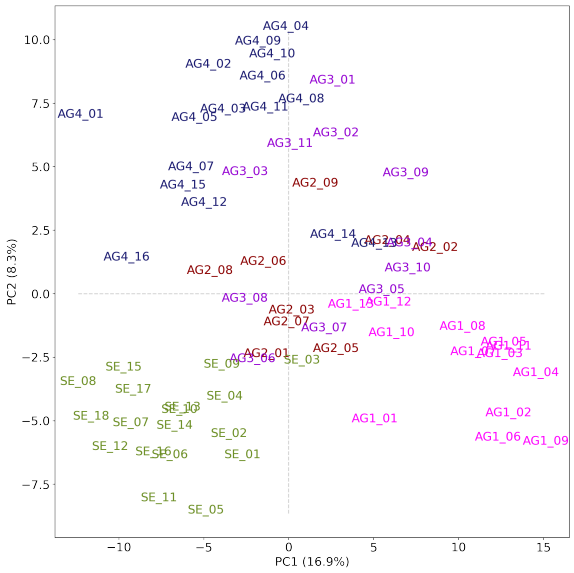

Figure 2: Scatter plot of PC1 and PC2 calculated with $(n, r) = (3, 300)$. PC1 can be understood as a representation of Montalvo's degree of contribution, whereas PC2 can be understood as reflecting his internal stylistic variation.

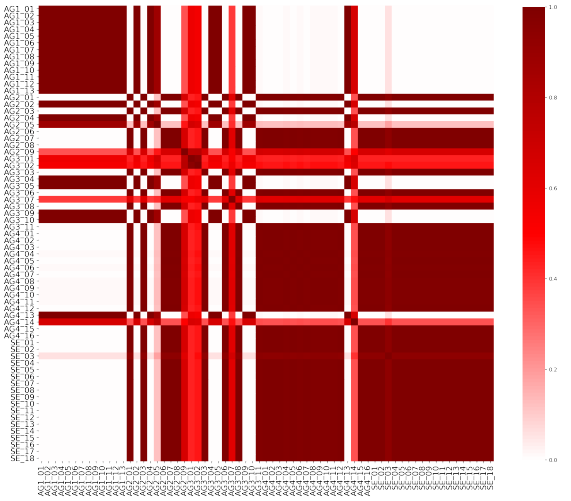

Figure 3: Pair-wise mean concordance rates computed from 100 iterations of k-means performed with $(k, n, r) = (2, 3, 300)$. The darker the cell, the more similar the pair of pieces.

AG1 only and the other by Montalvo's genuine writings, AG4 and SE. Meanwhile, AG2 and AG3 vacillated between the two clusters, which implies Montalvo's more extensive revisions there compared to AG1, which he might have retouched minimally. Our findings empirically corroborate suppositions formulated by Cacho Blecua (Rodríguez de Montalvo, 2020a).

5 Discussion

5.1 Sensitivity analysis of hyper-parameters

We examined the effects of the hyper-parameters and confirmed that the results were scarcely affected by n or r, which verifies the robustness of our findings. With respect to k-means, it is noteworthy that, even for $k = 5$, equal to the number of books, AG2 and AG3 jointly constituted a cluster instead of forming distinct individual groups, whereas AG1, AG4, and SE formed an individual one as shown in Figure 4. This result suggests that Montalvo accommodated AG2 and AG3 to his own literary style, to the point that they diverged from AG1, which seems almost intact.

5.2 Characteristic POS *n*-grams

We inspected n-grams whose frequency scores varied notably across the books and thus played a crucial role in the multivariate analyses. Figure 5 shows the trigrams among the top 300 for which the mean z-transformed relative frequency scores were above 1.0 or below -1.0 for any of the five books. Some of the sequences deserve special mention from the philological viewpoint:

CCONJ_VERB_PRON This trigram typically represents postposition of the pronoun to the finite verb (e.g., *y abrió lo* "and he/she opened it"). Its ratio was 0.63% in AG1, 0.44% in AG2, 0.52% in AG3, 0.32% in AG4, and 0.20% in SE. In his genuine creation, Montalvo apparently abstained from this syntactic pattern used recurrently in the first three books.

PRON_haber_PPART This trigram entails the use of the perfect tense (e.g., *lo he hecho* "I have done it"). Its ratio was 0.09% in AG1, 0.15% in AG2, 0.15% in AG3, 0.24% in AG4, and 0.16% in SE. We can see that Montalvo more frequently employed the perfect tense in his own works.

PUNCT_ADV_VERB This trigram represents the sentence beginning with an adverb followed by a finite verb (e.g., *Entonces dijeron* "Then they said"). Its ratio was 0.10% in AG1, 0.07% in AG2, 0.04% in AG3, 0.05% in AG4, and 0.05% in SE. This pattern was adopted more often in the first two books.

VERB_CCONJ_VERB This trigram typically represents two verbs joined with a coordinate conjunction (e.g., *cenaron y durmieron* "they had dinner and slept"). Its ratio was 0.36% in AG1, 0.28% in AG2, 0.32% in AG3, 0.29% in AG4, and 0.19%

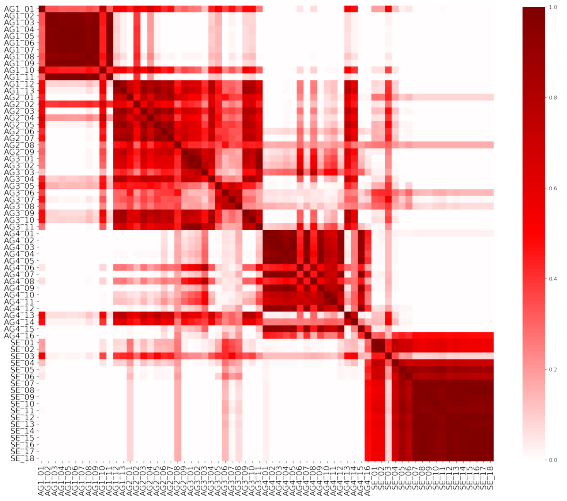

Figure 4: Pair-wise mean concordance rates computed from 100 iterations of k-means performed with $(k, n, r) = (5, 3, 300)$. The darker the cell, the more similar the pair of pieces.

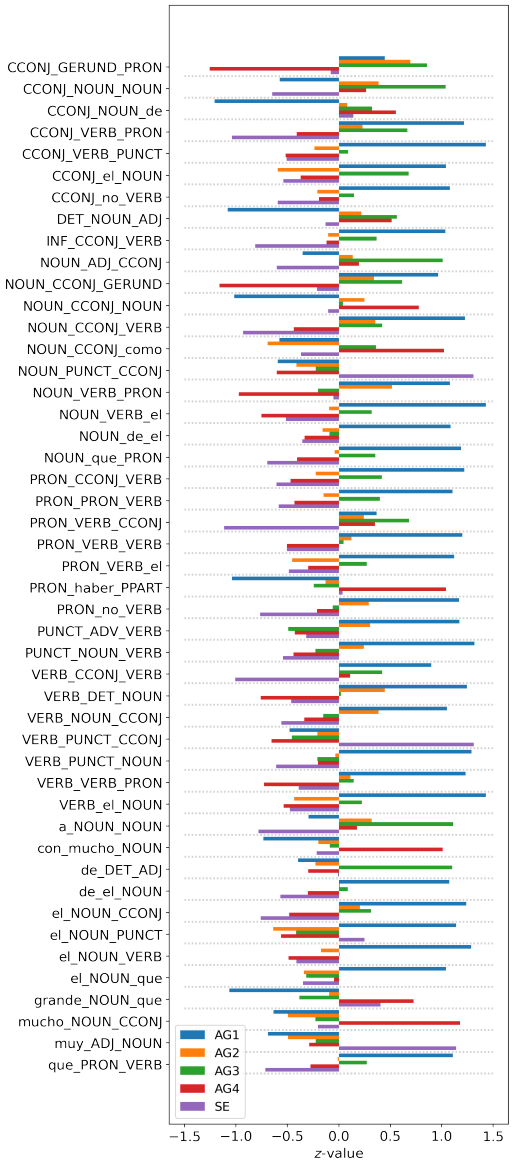

Figure 5: Trigrams among the top 300 for which the mean z-transformed relative frequency scores were above 1.0 or below -1.0 for any of the five books.

in SE. This syntagma might have been more frequently employed in the older versions of AG to which Montalvo had access.

VERB_PUNCT_NOUN This trigram represents closing a sentence with verb and opening the following one with (proper) noun. Its ratio was 0.25% in AG1, 0.15% in AG2, 0.14% in AG3, 0.14% in AG4, and 0.11% in SE. Montalvo seems to have avoided disposing verbs at sentence-final position.

grande_NOUN_que This trigram represents the noun preceded by adjective *grande* "great" and followed by relative pronoun *que* (e.g., *gran fatiga que* "great fatigue that"). Its ratio was 0.06% in AG1, 0.11% in AG2, 0.10% in AG3, 0.15% in AG4, and 0.14% in SE. Montalvo tended to utilize the syntagma more frequently in his own creation.

muy_ADJ_NOUN This trigram represents the nominal phrase of the type *muy leal caballero* "very loyal knight." Its ratio was 0.04% in AG1, 0.05% in AG2, 0.06% in AG3, 0.06% in AG4, and 0.13% in SE. This construction is found prominently in SE.

6 Conclusions

This study addressed a long-standing enigma concerning the genesis of the two monumental works authored by Montalvo. Applying stylometric techniques, we demonstrated that Montalvo's retouching on AG1 was minimal, while revising AG2 and AG3 to such an extent that they came to moderately resemble his authentic creations, AG4 and SE. Our findings empirically corroborate suppositions formulated from philological viewpoints by Cacho Blecua (Rodríguez de Montalvo, 2020a).

One limitation of our study is the lack of distinction between narration and conversation. The distinction is desirable, because varying proportions of the two components across the books could potentially affect the study's outcome. In so doing, we can also examine if authorial fingerprints are more clearly detectable in one part than in the other.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP18K12361.

References

Harald Baayen, Hans van Halteren, and Fiona Tweedie. 1996. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Juan Cerezo Soler and José Calvo Tello. 2019. Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de *La conquista de Jerusalén*. *Anales Cervantinos*, 51:231–250.

Ross Clement and David Sharp. 2003. Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*, 18(4):423–447.

Francisca Domingo del Campo. 1982. *El lenguaje en el* Amadís de Gaula. Tesis doctoral, Universidad Complutense de Madrid, Madrid.

Maciej Eder. 2015. Does Size Matter? Authorship Attribution, Small Samples, Big Problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

José Manuel Fradejas Rueda. 2016. El análisis estilométrico aplicado a la literatura española: Las novelas policiacas e históricas. *Caracteres. Estudios culturales y críticos de la esfera digital*, 5(2):196–245.

Alejandro García-Reidy. 2019. Deconstructing the Authorship of *Siempre ayuda la verdad*: A Play by Lope de Vega? *Neophilologus*, 103:493–510.

Jack Grieve. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Laura Hernández Lorenzo. 2019. Fernando de Herrera y la autoría de Versos: Un primer acercamiento al drama textual desde la Estilometría. *Romanische Studien*, 6:75–90.

Graeme Hirst and Ol'ga Feiguina. 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417.

Matthew L. Jockers and Daniela M. Witten. 2010. A Comparative Study of Machine Learning Methods for Authorship Attribution. *Literary and Linguistic Computing*, 25(2):215–223.

Patrick Juola. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Yoshifumi Kawasaki. 2021. Stylometric Analysis of Avellaneda's *Don Quijote*. In *12th International Conference on Corpus Linguistics*, Universidad de Murcia (Online). Spanish Association for Corpus Linguistics.

Mirco Kocher and Jacques Savoy. 2019. Evaluation of Text Representation Schemes and Distance Measures for Authorship Linking. *Digital Scholarship in the Humanities*, 34(1):189–207.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

Mallorie Labrousse. 2021. Los sistemas de los posesivos en el *Amadís de Gaula*, reflejo de un cambio lingüístico. *Revista de Historia de la Lengua Española*, 16:35–66.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yao Jean Marc Pokou, Philippe Fournier-Viger, and Chadia Moghrabi. 2016. Authorship Attribution Using Variable Length Part-of-Speech Patterns. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, volume 2, pages 354–361.

Nanette Rißler-Pipka. 2016. Avellaneda y los problemas de la identificación del autor: Propuestas para una investigación con nuevas herramientas digitales. In *El otro Don Quijote. La continuación de Fernández de Avellaneda y sus efectos*, pages 27–51, Augsburgo. Institut für Spanien-, Portugal- und Lateinamerikastudien.

Garci Rodríguez de Montalvo. 2003. *Sergas de Esplandián*. Editorial Castalia, Madrid.

Garci Rodríguez de Montalvo. 2020a. *Amadís de Gaula I*, 12th edition. Cátedra, Madrid.

Garci Rodríguez de Montalvo. 2020b. *Amadís de Gaula II*, 12th edition. Cátedra, Madrid.

Antonio Rojas Castro. 2017. Luis de Góngora y la fábula mitológica del Siglo de Oro: clasificación de textos y análisis léxico con métodos informáticos. *Studia Aurea*, 11:111–142.

Javier de la Rosa and Juan Luis Suárez. 2016. The Life of *Lazarillo de Tormes* and of His Machine Learning Adversities. *Lemir*, 20:373–438.

Jacques Savoy. 2017. Analysis of the Style and the Rhetoric of the American Presidents over Two Centuries. *Glottometrics*, 38:55–76.

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

Ayaka Uesaka and Masakatsu Murakami. 2015. Verifying the authorship of Saikaku Ihara's work in early modern Japanese literature; A quantitative approach. *Digital Scholarship in the Humanities*, 30(4):599–607.

Ying Zhao and Justin Zobel. 2007. Searching With Style: Authorship Attribution in Classic Literature. In *Proceedings of the Thirtieth Australasian Computer Science Conference*, volume 62 of *CRPIT*, pages 59–68, Ballarat. Australian Computer Society.