hosted by Hamburg University Press

apropos [Perspektiven auf die Romania]

Sprache∫Literatur∫Kultur∫Geschichte∫Ideen∫Politik∫Gesellschaft

Atribución de autoría y humanidades digitales en el Siglo de Oro español

Alejandro Cienfuegos Pérez

apropos [Perspektiven auf die Romania]
hosted by Hamburg University Press
2022, 9
pp. 277-305
ISSN: 2627-3446

Online

https://journals.sub.uni-hamburg.de/apropos/article/view/1939

Zitierweise

Cienfuegos Pérez, Alejandro. 2022. „Atribución de autoría y humanidades digitales en el Siglo de Oro español.“ *apropos [Perspektiven auf die Romania]* 9/2022, 277-305.
doi: https://doi.org/10.15460/apropos.9.1939

2022, n°9
pp. 277-305
doi: https://doi.org/10.15460/apropos.9.1939

Alejandro Cienfuegos Pérez

Atribución de autoría y humanidades digitales en el Siglo de Oro español

Alejandro Cienfuegos Pérez
estudió lenguas romances (español y francés) en la Universidad de Göttingen (Alemania).
alejandro.cienfuegos.perez@gmail.com

Palabras clave

Literatura – lingüística – humanidades digitales – siglo de oro español – verificación de autoría – atribución de autoría

Este artículo se basa en una tesis de maestría que se realizó en el programa "TransRomania-Studien: Romanische Sprachen, Literaturen und Kulturen" en la Georg-August-Universität Göttingen bajo la supervisión de PD Dr. Nanette Rissler-Pipka. El trabajo ha sido abreviado y revisado para su publicación en la sección Premiers Travaux de *apropos*.

1. Introducción

Partimos de la hipótesis de que los nuevos métodos de la estilometría, en particular el empleo de la herramienta *stylo* de R (Eder 2018; Eder y Rybicki 2012; Eder y Rybicki 2011) puede si no dar una solución definitiva al menos aportar nuevas perspectivas a una discusión literaria largamente debatida: la de la autoría de *La tía fingida* atribuida por muchos a Cervantes. Así mismo, también consideramos relevante no obviar la perspectiva literaria cuando la determinación de autoría se ocupa del estilo literario de un autor [1].

Proseguiremos con una de las partes que consideramos más complejas en el entorno del Siglo de Oro que es llevar a cabo la compilación de un corpus suficientemente representativo y fiable de textos digitales. El camino desde las fuentes, pasando por la problemática y limitaciones hasta la cosecha de textos y configuración del corpus será expuesto aquí profusamente. También es ineludible la relación que esto tiene con el texto dubitado. De él se comentará su relevancia subrayando los avatares desde la aparición del manuscrito original hasta llegar a las múltiples versiones que existen hoy en día del mismo.

[1] A pesar de que consideramos la importancia del uso del lenguaje inclusivo y la distinción autor/autora y similares, evitaremos su empleo a fin de garantizar una mejor legibilidad del texto. En cualquier caso, entiéndase que utilizamos un masculino genérico para referirnos a ambos sexos.

Llegados a este punto, se empleará la herramienta *stylo* para las pruebas estilométricas utilizando diferentes funciones. En base a los datos obtenidos en los experimentos podremos pasar al análisis de *La tía fingida* y aquella/s novela/s que presenten una mayor similitud estilométrica desde una perspectiva literaria. Esta última parte del trabajo se omite en este artículo, aunque se puede consultar en Cienfuegos-Pérez (2022).

2. Corpus

2.1 Problemática de los textos del Siglo de Oro

El Siglo de Oro español es de gran interés para los trabajos de atribución de autoría debido a diversos motivos que tienen que ver sobre todo con las circunstancias de transmisión de los mismos. Sin embargo, esas mismas circunstancias de transmisión también dan lugar numerosos problemas para la investigación como aquí expondremos.

A pesar de que actualmente están a disposición de la investigación numerosos textos y ediciones digitales, como comenta Rißler-Pipka (2018) para el análisis estilométrico es necesario disponer de los textos digitales, pero también los metadatos deben estar completos. Estos contienen informaciones extralingüísticas de los textos: autor, fecha y lugar de publicación, etc. Su importancia se debe al hecho de que la calidad de las ediciones de los textos de este periodo es difícilmente comprobable. Juola (2008, 247) apunta en este sentido a otro inconveniente que puede existir en las ediciones digitales como las de Google Scholar, JSTOR, o el Project Gutenberg que podrían haber sido corrompidas debido al proceso de escaneado o reescritura.

Otros de los aspectos problemáticos para la atribución de autoría es la determinación del corpus como abierto o cerrado (*open/closed set*) que vimos en Juola (2008). En este sentido, la primera pregunta que habría que responder es ¿sabemos a ciencia cierta que el posible autor se encuentra entre los seleccionados? Para nuestro objetivo un análisis exhaustivo conllevaría reunir todo el corpus de obras de los posibles autores, que aun suponiendo que pudiera ser Cervantes y existiendo determinadas hipótesis podría ser cualquiera que, por ejemplo, hubiera realizado trabajos del mismo género (novela breve) en torno a la fecha de publicación de la obra cuya autoría queremos determinar. Como veremos, existen limitaciones para poder reunir este corpus ideal.

Siguiendo con los manuscritos es necesario reseñar que los textos que mejores resultados estadísticos dan en cuanto a fiabilidad son aquellos cuyo contenido es el expresamente creado por el autor. Pues bien, incluso tomando como punto de partida los manuscritos del Siglo de Oro, estos presentan una serie de problemas dado que los procesos de impresión y el concepto de propiedad intelectual no eran tal y como los conocemos hoy en día. Francisco Rico (2000) publicó un trabajo coordinado al respecto de esta problemática con el nombre *Imprenta y crítica textual en el Siglo de Oro*. En este libro, podemos no solamente hacernos una idea

de las ediciones que se realizaban por el propio impresor de los textos originales sino también del origen del problema que nos ocupa: la atribución de autoría.

Por otra parte, Arellano-Ayuso (1997, 41-42) comenta las vicisitudes de la transmisión literaria en el Siglo de Oro observando la trayectoria de la difusión de los textos desde el manuscrito pasando por la edición y la posterior venta del libro. Arellano-Ayuso (1997, 42) divide su artículo distinguiendo y explicando los avatares de los textos según su tipo. Dentro de los textos impresos en prosa, Arellano-Ayuso (1997, 51) da muestra de algunos problemas en el proceso de impresión que dificultaron el paso del manuscrito al libro. En cuanto a las *Novelas ejemplares* de Cervantes que se publican en 1613 comentan que algunas de ellas eran anteriores, pues en *El Quijote* de 1605 se cita al *Rinconete y Cortadillo.* Entre 1604-1606 fue compilada una colección miscelánea de obras para el cardenal Fernando Niño de Guevara por parte de Francisco Porras de la Cámara. En ella incluye sendas versiones con variantes de *Rinconete y Cortadillo* y de *El celoso extremeño,* además de *La tía fingida*. El manuscrito de Porras de la Cámara, desapareció cuando estaba en manos de Bartolomé José Gallardo (1835), que había realizado algunas copias. Hoy tenemos la copia manuscrita ("basada" en el códice Porras de 1604-1606) y la versión impresa de la novela *El celoso extremeño (1613)*. Lo curioso de estas dos versiones es que el final difiere. Esto ha llevado a discrepancias entre los estudiosos: Maregalli (1992) achaca el final de 1613 a la censura, Canavaggio (1992), sin embargo, considera que el mismo Cervantes habría llevado a cabo la reescritura para dar a la trama y los personajes mayor complejidad (cfr. Arellano-Ayuso 1997, 55). En cualquiera de los casos, nuestra intención no es en este punto entrar en este debate, sino exponer algunos ejemplos de las vicisitudes y los cambios realizados en los textos que son otro de los variados problemas inherentes a la literatura del Siglo de Oro.

Una vez expuestos los aspectos limitantes más generales pasaremos al siguiente apartado donde expondremos las fuentes de las que disponemos para poder compilar un corpus de trabajo.

2.2 Fuentes para la recopilación de textos del corpus

En otro orden de aspectos para un trabajo de estilometría es indispensable reunir un amplio corpus que sea fiable y representativo. Para este fin, existen en la actualidad algunos recursos que son de gran ayuda pero que sin embargo también presentan en muchos casos ciertas limitaciones. José Calvo Tello en Github[2] pone a disposición una lista de portales que aportan recursos digitales en español. Entre ellos podemos destacar la base de datos de la Biblioteca Virtual Miguel Cervantes (BVMC) que da acceso a un inmenso catálogo de libros entre los que se hayan una gran cantidad del Siglo de Oro. Sin embargo, como comenta Rißler-Pipka (2018) muchos de estos libros son ediciones antiguas por cuestiones de derechos de autor lo cual es un perjuicio a la hora de realizar un escaneado además de que su lenguaje puede no corresponder con el español moderno. Otras bases de datos reseñables que pueden servir de fuente para la recopilación de textos son: AHCT (Association

[2] <https://github.com/morethanbooks/Atlas-de-Datos/blob/master/atlas%20de%20datos.csv>.

for Hispanic Classical Theater); CORDE (Corpus diacrónico del español) que no pone a disposición ningún texto completo permitiendo solamente la búsqueda de títulos y autores en un periodo determinado.

Todos los aspectos problemáticos y limitantes expuestos en este apartado podrían hacer pensar que un trabajo como el que nos ocupa podría no presentar la suficiente fiabilidad y demasiados escollos para poder ser conclusivo. Pues bien, tampoco lo pretende, recuperemos en este punto las palabras de Holmes y Kardos (2003) que afirmaban que "la estilometría […] no pretende anular la escolástica tradicional de los expertos en literatura e historia, más bien trata de complementar su trabajo proveyendo significados alternativos a los trabajos de investigación sobre proveniencia dudosa"[3] (Holmes y Kardos 2003, 1, traducción propia) y adjuntaban una insistencia final donde enuncian que "la evidencia estilométrica tiene que ser contrastada con la provista por estudios más convencionales hechos por los estudios literarios"[4] (Holmes y Kardos 2003, 5, traducción propia). Por ende, el objetivo de aportar nuevas perspectivas sigue siendo legítimo y posible pese a las limitaciones que la época impone al corpus. Además, aunque debemos de tener en cuenta todos los aspectos condicionantes aquí señalados, estos no han sido óbice para obtener resultados satisfactorios en otros trabajos en los cuales nos basamos para aplicar sus métodos.

Por último, consideramos que la utilización de las MFW[5] como marcador de estilo puede en parte hacer frente a algunos de las restricciones como la alteración que pueden producir en la estadística las modificaciones y ediciones del texto. Pensamos que siendo estas a menudo las denominadas "palabras funcionales" son menos susceptibles de ser editadas pues no varían el significado o el mensaje de un determinado texto. Por otra parte, en algunos casos pueden faltar partes del texto para lo cual no existe a nuestro saber ninguna solución. No obstante, esta consideración es una mera hipótesis. Una respuesta a esta conllevaría un estudio y análisis profundos de las vicisitudes y los cambios más frecuentes en los textos del Siglo de Oro que estableciera exactamente cómo y en qué medida se modificaban los textos.

Vistas las fuentes, pasaremos pues en el próximo apartado al proceso de compilación de nuestro corpus y la explicación de su configuración.

2.3 Cosecha de los textos y configuración del corpus

Para compilar el corpus con el que vamos a trabajar llevamos a cabo en primer lugar una consulta en el CORDE[6]. Dentro de este, seleccionamos todos los autores entre

[3] Original: "Stylometry – the statistical analysis of literary style – does not seek to overturn traditional scholarship by literary experts and historians, rather it seeks to complement their work by providing an alternative means of investigating works of doubtful provenance" (Holmes y Kardos 2003, 1)

[4] Original: "stylometric evidence must always be weighed in the balance along with that provided by more conventional studies made by literary scholars" (Holmes y Kardos 2003, 5)

[5] Most frequent words: se refiere a las palabras que más frecuencia presentan en una determinada lengua. Suelen corresponder con las denominadas palabras funcionales (preposiciones p. ej.). A partir de este punto utilizaremos la abreviatura MFW.

[6] REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]. Corpus diacrónico del español. <http://www.rae.es> [30.10.21]

1547 y 1620, periodo que coincide prácticamente con la vida de Miguel de Cervantes (1547-1616). Es una selección amplia que pretende abarcar todas las hipótesis posibles por lejanas que parezcan. Dentro de este periodo nos interesa saber que autores realizaron prosa narrativa breve en España, tanto culta como tradicional. El buscador arroja un total de 63 documentos con 677554 palabras. De estos resultados eliminaremos los numerosos anónimos que aparecen, pues no nos interesa incluir más textos dubitados. También se excluirán cartas, memoriales, premáticas y otros textos que por su extensión excesivamente breve o su tipología pueden ser problemáticas al compararlas con la TF.

En un análisis ideal dispondríamos de suficientes textos de prosa breve similares a las *Novelas Ejemplares* de diferentes autores de los que además tendríamos numerosos ejemplos, no obstante, no es el caso y no hay nada que podamos hacer para remediarlo. Una nueva búsqueda incluyendo la narrativa extensa da en el CORDE como resultado 39 documentos con un total de 4061193 palabras. Las *Novelas Ejemplares* tienen alrededor de 5000 hasta ca. 23000 palabras y como vimos, para el análisis estilométrico, necesitamos textos que sean lo más similares en cuanto a su extensión, sin embargo, dados los límites del corpus de la prosa breve no estamos en condiciones de rechazar muestras.

A este corpus debemos adjuntar otros textos que tienen algunas similitudes con las *Novelas Ejemplares* pero que pertenecen a una época posterior a Cervantes. Los incluiremos para añadir más fuentes de autores que están presentes en los corpus anteriores y que ayudarán a identificar mejor su estilo. Creemos que es poco probable que alguno de los presentes de este subcorpus hayan sido los creadores de la TF, pues esta precede a las primeras publicaciones que realizaron como escritores. Estos textos tienen también un argumento a favor para su inclusión que es su disponibilidad en formato txt. a través del *Github* de Fradejas Rueda[7]. Por otra parte, existe un problema y es que no disponemos de todos los metadatos acerca del origen de la edición de los textos más allá de la explicación de su propietario[8] por este motivo trataremos de contrastar los textos y completar estos datos en nuestro corpus.

Es necesario señalar que los autores que están representados por una sola obra son problemáticos para la clasificación utilizando *stylo* pues el algoritmo va a tratar de encontrar los textos que presentan una mayor similitud. Al no existir más ejemplos del mismo autor, es probable que no se puedan observar las características estadísticas de forma clara y el texto sea situado en algún lugar erróneo dando prioridad a otras cuestiones como el género, la extensión o la temática y genere de este modo irregularidades. Por este motivo y para aplacar el obstáculo que supone la diferente extensión de los textos, vamos a fragmentar los documentos del corpus igualando su longitud entorno a las 6000-7000 palabras, lo cual también nos posibilita multiplicar las muestras de un determinado autor e

[7] https://github.com/7PartidasDigital

[8] "Los textos relacionados con M.ª de Zayas y Alonso Castillo Solórzano proceden de los ficheros de Alejandro García-Reydi (USAL), los del entorno del Quijote de Avellaneda los he cosechado en la Cervantes Virtual y algunos me los ha pasado Javier Blasco Pascual (UVa)". Fradejas Rueda en: <https://github.com/7PartidasDigital/NovelaBarroca>.

incorporarlo incluso disponiendo de un solo texto de base. Estas muestras podrían ser útiles a la hora de configurar los parámetros de la herramienta donde servirán quizás para tratar de anular algunas de las señales que se desvelan tras las pruebas estilométricas[9]

Tras este trabajo de compilación, hemos realizado varios (sub)corpora: un primer corpus de la prosa narrativa breve contemporánea a Cervantes, muy reducido; un segundo corpus con la prosa narrativa extensa también contemporánea a Cervantes que presenta un número de palabras por texto muy superior al de las *Novelas Ejemplares*; y un tercer corpus de obras del Siglo de Oro de autores que suponemos menos susceptibles de guardar relación con la TF.

Al corpus añadiríamos por último la versión de la obra dubitada. Encontramos diferentes ediciones en la BVMC por lo que surge la cuestión en torno a cuál de ellas seleccionar. José Manuel Lucía Megías (2018) hace un comentario que nos parece relevante con respecto a las ediciones:

> La tía fingida, […] a pesar de los esfuerzos editoriales de los últimos años, aún carece de la edición crítica que dé cuenta de la complejidad de sus materiales textuales y el hecho de contar con testimonios de muy diversa naturaleza —copias antiguas y modernas— que dan cuenta de dos redacciones de la obra. (Lucía Megías 2018, 346)

De entre los testimonios que disponemos seleccionaremos la versión de Porras de la Cámara realizada por Francenson y Wolf (Berlín, 1818) en la edición realizada por Florencio Sevilla Arrollo que se encuentra disponible en formato digital en la BVMC. Esta edición, la berlinesa, es en palabras de Lucía Megías (2018, 344-345) "*muy correcta y fiel a su testimonio base, en que solo se aleja de los tres grandes grupos de cambios lingüísticos y ortográficos*" que como indica y presenta profusamente en el apéndice son: cambio de grafías, cambio de mayúsculas y minúsculas y cambios en la puntuación. Para nuestro análisis la modernización de las grafías es un aspecto positivo que hemos buscado también en el resto de textos para poder disponer de cierta homogeneidad. Los cambios en la puntuación no son relevantes pues no son un elemento fiable en los textos de esta época por lo que no serán tenidos en cuenta en el análisis cuantitativo. Los datos de la edición seleccionada son los siguientes:

- Autor: desconocido ¿Cervantes?
- Título: La tía fingida
- Fecha de publicación: 16??
- Edición: Novela de La tía fingida [versión Porras de la Cámara por Francenson/Wolf]
- Versión digital: <http://www.cervantesvirtual.com/nd/ark:/59851/bmc1z4h>.

No se nos escapa el hecho de que el corpus que hemos recopilado parte de los datos que nos da el CORDE. Como ya comentamos, nos basamos en aspectos sincrónicos seleccionando todas aquellas obras disponibles que correspondieran

[9] que, como resumen Cerezo Soler y Calvo Tello (2019, 237) citando algunos ejemplos de trabajos al respecto, son: el género literario (Kestemont 2011), la época de composición (Jockers 2014) y otras.

con la categoría de prosa narrativa. Sin embargo, el CORDE tiene en cuenta aspectos lingüísticos, no literarios y por lo tanto da como resultado una serie de textos de diferentes géneros literarios. Aunque en base a los resultados del buscador del CORDE hay una cierta homogeneidad en realidad desde un punto de vista literario nos encontramos con un corpus más bien heterogéneo.

3. *Stylo* de R: justificación de la herramienta.

En este punto ya tenemos listo un corpus considerable de textos en prosa del Siglo de Oro. Así pues, el siguiente paso antes de poner en marcha el análisis es presentar la herramienta que vamos a utilizar para llevarlo a cabo para así poder también comprender los resultados que arrojará y su índice de fiabilidad.

R es un entorno y lenguaje de programación que está enfocado en el análisis estadístico pero que como apunta Rueda (2019) y muestran los numerosos estudios que se han servido de este entorno tiene un gran potencial para su uso en los estudios filológicos. Como ejemplo de los trabajos realizados utilizando R, Rueda (2019) hace referencia a estudios del ámbito de la lingüística (Gries 2013, Gries 2016, Desagulier 2017, Levshina 2015, Winter 2019) y de la literatura (Jockers 2014). Si bien esta lista, se puede ampliar inmensamente.

Dentro de este entorno, para llevar a cabo el análisis estilométrico existen múltiples herramientas. Una de las más recientes y que ha dado buenos resultados en varias investigaciones es el paquete *stylo* basado en R (Eder, Rybicki y Kestemont 2016) mantenido y desarrollado por el grupo de estilística computacional[10]. Este paquete aporta diferentes funciones para el análisis estilométrico además de una interfaz de uso sencillo y diagramas de gran calidad.

Recordemos en este punto que Juola (2008) comentaba respecto a las herramientas a utilizar para el análisis estilométrico que lo que es más importante para el usuario casual es la capacidad de seleccionar los algoritmos y las características a utilizar dinámicamente, en función del idioma, el género, el tamaño, de los documentos disponibles. En este sentido *stylo,* nos parece una herramienta que cumple con todas estas necesidades. Con respecto a la precisión nos remitiremos al trabajo de Blasco Pascual y Ruiz Urbón (2009) que realizan una evaluación y cuantificación de algunas técnicas de atribución de autoría en textos españoles. En su trabajo, muestran que el *perfil de palabra simple* es el más efectivo para los textos del Siglo de Oro con un 70% de precisión en un corpus de 10 autores que se eleva al 90% entre dos autores. Además, dan cuenta de que el perfil más efectivo es el de *marca de puntuación simple1*[11] que presenta un 85% de fiabilidad en un corpus de 10 autores; sin embargo este método no es aplicable al Siglo de Oro dado que, como ya comentamos en el apartado sobre la problemática de los textos de esta época, en particular su puntuación no es suficientemente fiable[12].

[10] <https://computationalstylistics.github.io/resources/>

[11] Este perfil es calculado mediante la división de la frecuencia de una serie de signos de puntuación entre el número total de caracteres contenidos en el texto.

[12] Pascual y Urbón (2009) comentan en sus conclusiones que hay dos razones que suponen un problema para la atribución de autoría y que tienen que ver una con la fase de creación y otra con la fase de edición de los textos. Son las siguientes: *a) Los hábitos de escritura en los Siglos de Oro* b) Los correctores, componedores y

De este modo, de los diferentes métodos posibles para el análisis solo queda uno efectivo: "Desgraciadamente del *top ten* obtenido para el español sólo resulta operativo el *perfil de palabra simple*" (Blasco Pascual y Ruiz Urbón 2009, 44).

Así pues, llegados a este punto queda explicado el motivo del uso de *stylo*: medir una de las realidades textuales que es la frecuencia de palabras y dentro de estas *el perfil de palabra simple* que parece ser teóricamente el más adecuado y efectivo para este conjunto de textos[13]. Con el corpus compilado y la herramienta presentada podemos pasar a su empleo práctico.

4. Pruebas estilométricas

El primer paso a realizar es una serie de experimentos con el fin de determinar si nuestro corpus es organizado de manera lógica (por autores) por parte de la herramienta. Por el momento dejaremos la obra dubitada a un lado. Vamos a realizar un *Cluster Analysis* de cada uno de los subcorpora para crear un corpus final con el que trabajaremos. Eder (2015) considera que la extensión del texto que mejor define la señal autorial está en torno a las 5000 palabras y Hernández Lorenzo (2019, 192) comenta que el estilo en la narrativa suele estar más diseminado que por ejemplo en la poesía. Sin embargo, la novela breve tiende a concentrar el estilo. Además, como explicaremos a continuación hemos dividido las novelas extensas en fragmentos de unas 6000 palabras para lograr cierta homogeneidad de los fragmentos, así pues, utilizaremos el rango de las 500 MFW para nuestras pruebas estilométricas.

4.1 Análisis de *clúster* para ajustar el corpus

Corpus 1: Prosa narrativa breve

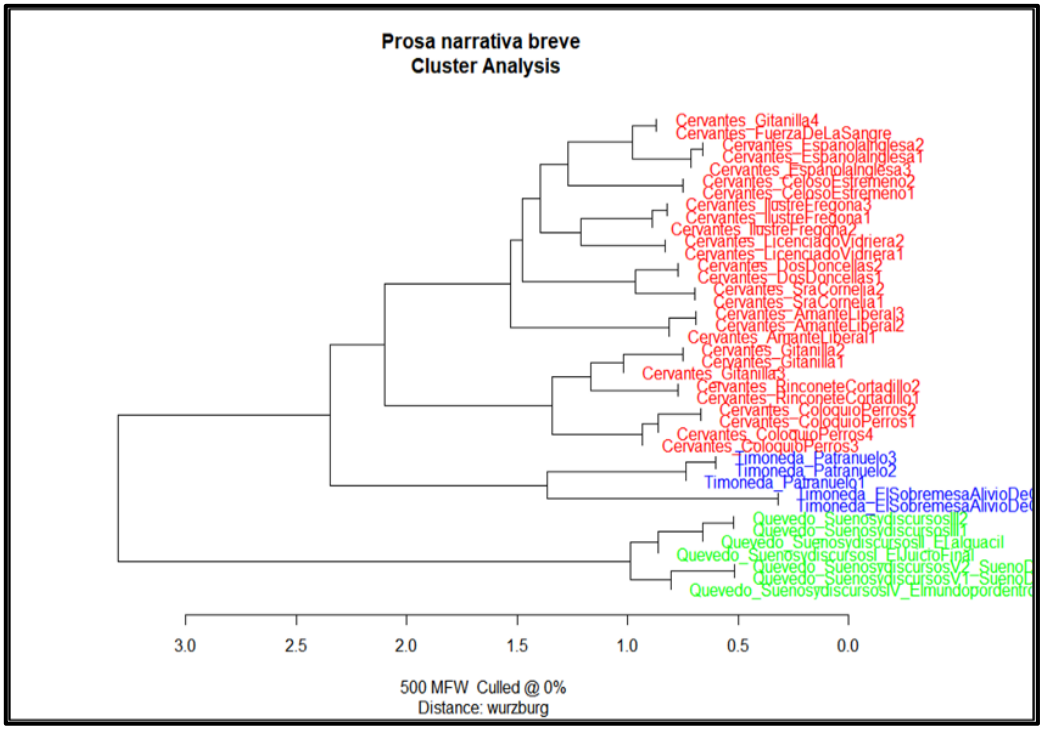
Tras varias pruebas con la función *stylo ()*, encontramos que la organización de los textos es eficiente utilizando de las 100 a las 500 MFW y *Cosine Delta*. También se observó en los primeros experimentos que existían algunas irregularidades que podrían ser debidas a la distinta extensión de los textos. Esta heterogeneidad de los textos está presente con mayor o menor medida en todo el corpus. Es por ese motivo por el cual hemos decidido fragmentar las obras en partes de aproximadamente 6000 palabras con el fin de homogeneizar lo más posible nuestros experimentos. La organización de los textos con 500 MFW se puede ver en la siguiente imagen. La herramienta reconoce sin problemas la obra de los autores y los organiza como se esperaba.

Si observamos el eje vertical, vemos que la herramienta reconoce dos Cervantes diferentes y que algunos de los textos no están organizados con sus correspondientes partes. Esto muestra que hay una cierta heterogeneidad en este

cajistas de imprenta podrían ser en alta medida los responsables de muchas de las marcas que nuestros procedimientos de medición actuales contemplan. (Blasco Pascual y Ruiz Urbón 2009, 44)

[13] Sería interesante en este punto presentar la funcionalidad y el uso de la herramienta estilo. Sin embargo, esto extendería mucho este artículo por lo que referimos al lector/a las publicaciones de sus creadores Maciej Eder, Jan Rybicki y Mike Kestemont (2016) y los tutoriales que se encuentran disponibles en la red, por ejemplo en: <https://computationalstylistics.github.io/stylo_nutshell/>.

corpus cervantino y pone quizás de manifiesto la relación entre ciertas partes de diferentes obras (por ejemplo, la cuarta parte de *La gitanilla* y *La fuerza de la sangre*). En comparación a otras pruebas hechas utilizando los textos completos, la división parece ayudar a que los textos se organicen mejor. Así mismo esto nos permitirá más adelante incluir algunos autores de los que solo disponemos de un texto que de otra manera sería imposible por no ser suficientemente representativo del "estilo" ya que el sistema trataría de asociarlo a un texto cercano y causaría imprecisiones. El corpus siguiente va a incluir por lo tanto los textos del corpus 1 y del corpus 2.

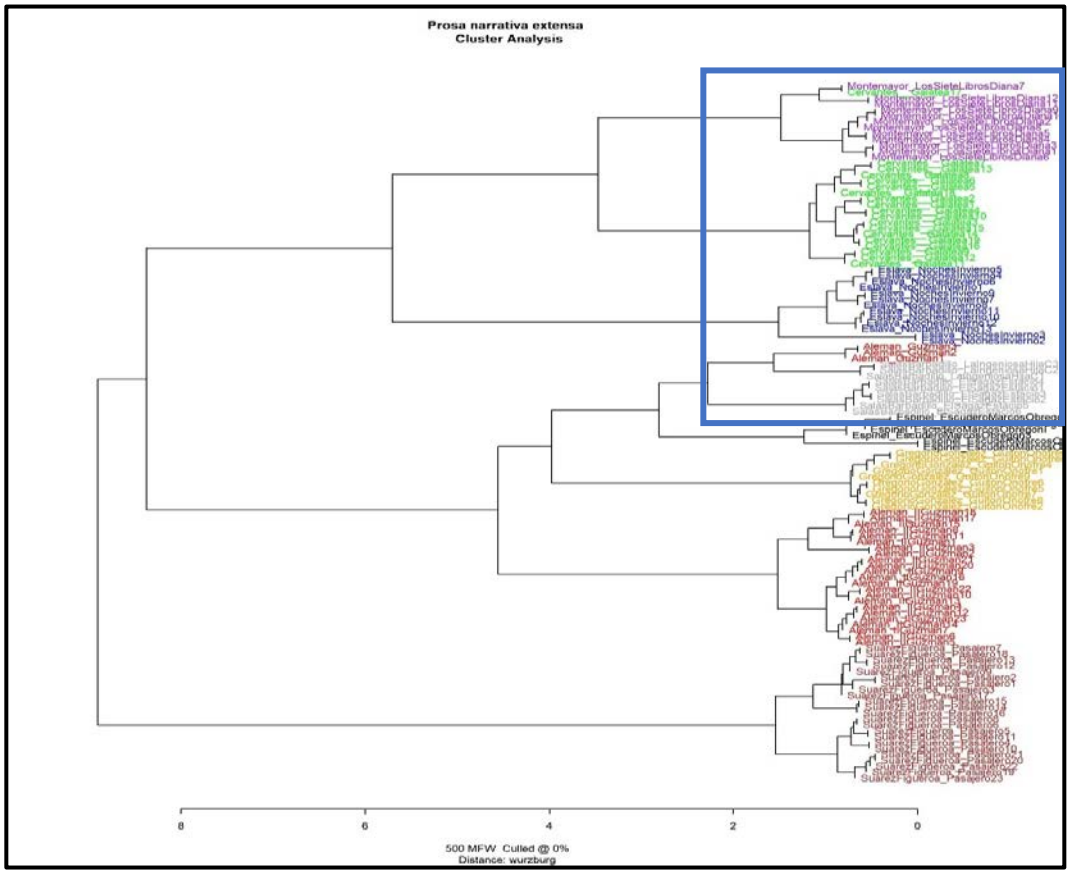


1 | prosa narrativa breve, análisis de *clúster*.

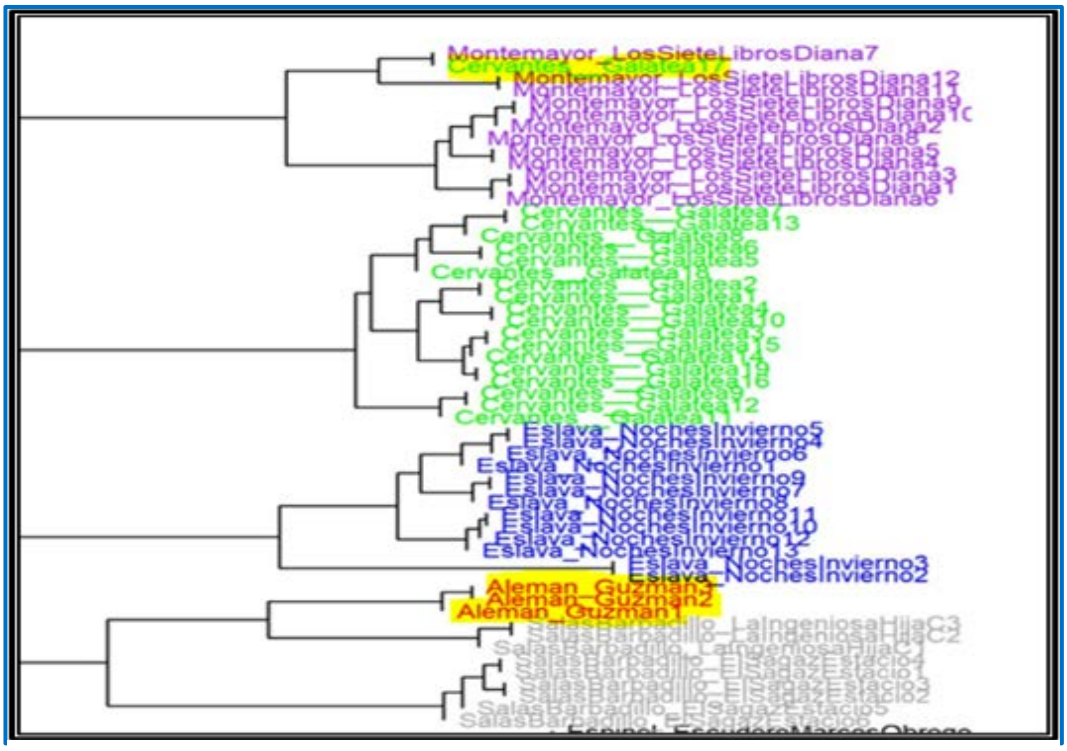
Corpus 2: Prosa narrativa extensa

Continuamos el experimento manteniendo la misma configuración de 500 MFW y *Cosine Delta*. Esta vez con nuestro segundo (sub)corpus que contiene las obras que corresponden a la prosa narrativa extensa fragmentada en partes de aprox. 6000 palabras. Vemos que en los resultados la organización también es la esperada (imagen 2). Solamente sorprende la parte diecisiete de *La Galatea* que aparece junto a la séptima parte *Los siete libros de Diana* de Montemayor. Igualmente, las tres primeras partes del *Guzmán de Alfarache* de Alemán aparecen separadas del resto de la obra en un subcluster en el que encontramos los textos de Salas Barbadillo. Aunque este dato podría ser interesante para el análisis literario, no es el tema que nos ocupa. Nuestro objetivo es conseguir un entorno libre de interferencias en el que podamos incorporar un texto dubitado. Por este motivo y considerando que las obras están suficientemente representadas, vamos a prescindir de esas partes de los textos.

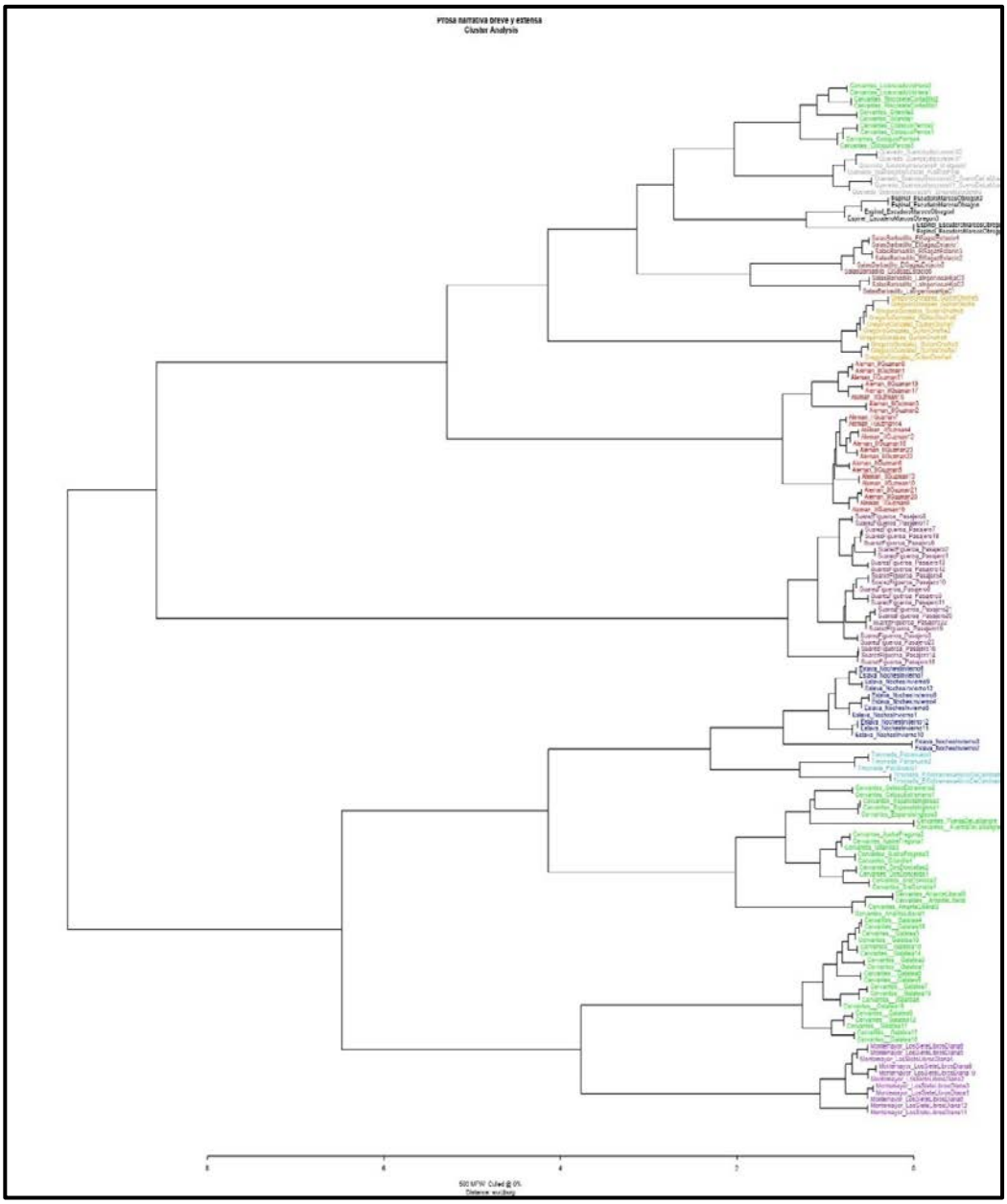
Al repetir la prueba sin estas partes, la organización de las obras por autores es satisfactoria. En la imagen se puede ver el dendrograma con los primeros resultados:



2 | prosa narrativa extensa, análisis de *clúster*, 500 MFW, *Cosine Delta*.

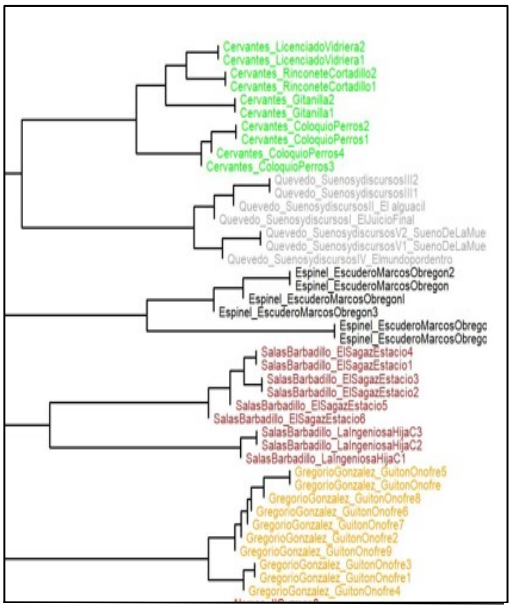


3 | detalle del dendrograma (Imagen 2), partes del texto problemáticas.

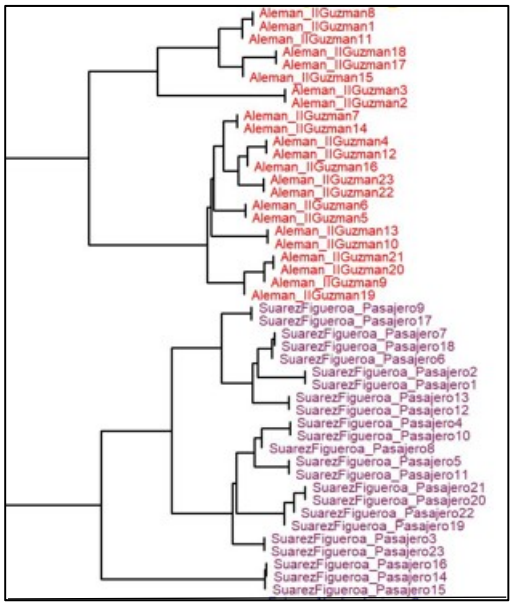


4 | prosa narrativa breve y extensa, análisis de *clúster*, 500 MFW, *Cosine Delta*.

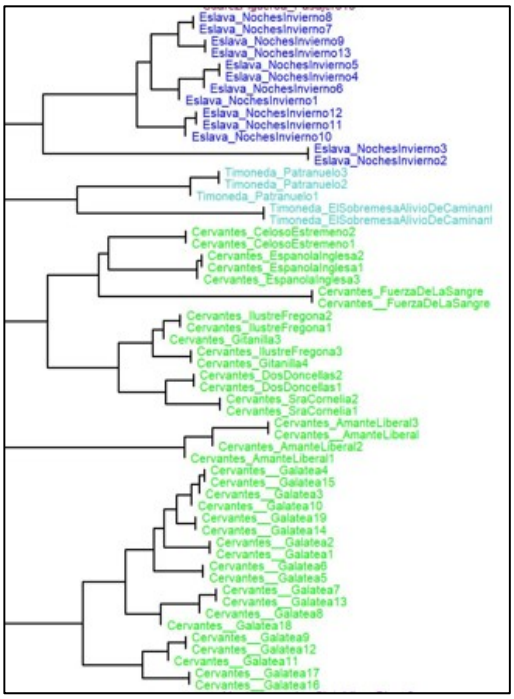
Llegados a este punto, vemos que una vez excluidos los fragmentos de los textos que causaban irregularidades, la herramienta es capaz de organizar las obras por autores en los (sub)corpus por separado por lo que ya podemos proceder a juntar ambos (sub)corpus para ver cómo se comportan en conjunto. La imagen de la página siguiente muestra que la organización de los autores y sus obras en *clusters* es correcta utilizando las 500 MFW y *Cosine Delta*. El amplio número de textos hace que la imagen del dendrograma se vea con dificultad por lo que incluimos imágenes con los detalles de las diferentes partes:



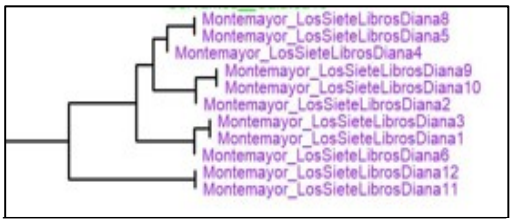
5 | sección 1 del dendograma (Imagen 4)



6 | sección 2 del dendograma (Imagen 4)



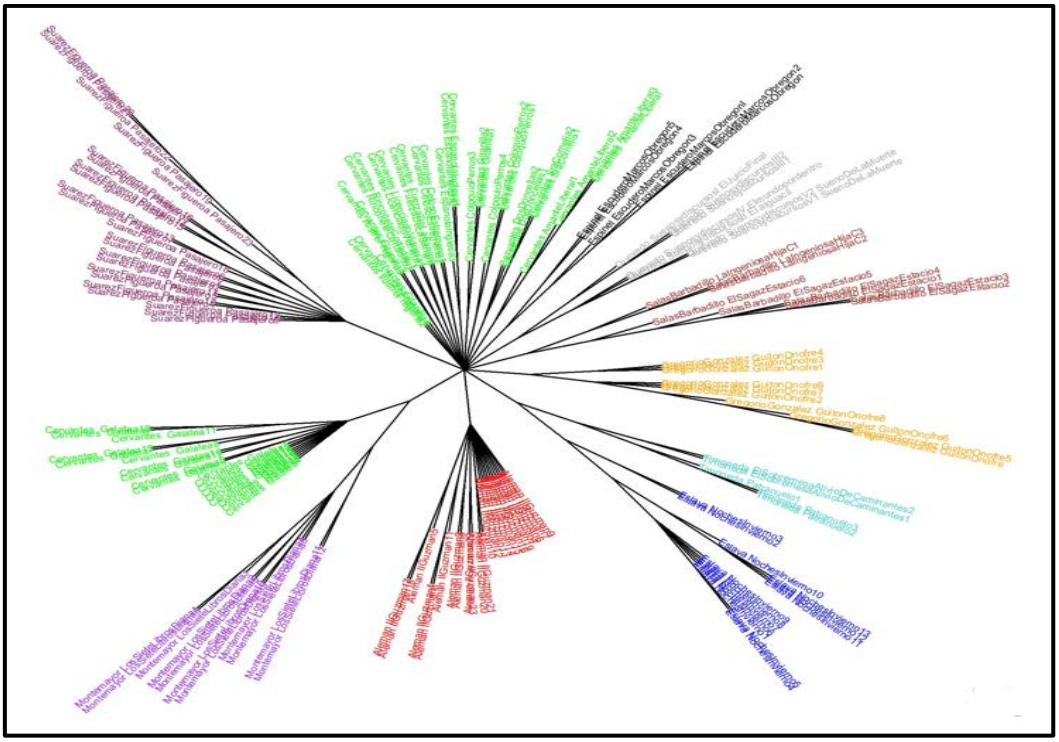
7 | sección 3 del dendograma (Imagen 4)



8 | sección 4 del dendograma (Imagen 4)

Algunos resultados dignos de mención son la clasificación de dos de las partes de *El escudero Marcos Obregón* de Espinel (imagen 5), de *Noches de Invierno de Eslava* (imagen 7) o de *La fuerza de la sangre* de Cervantes (imagen 7). En los dos primeros casos entendemos que son partes de los textos que se diferencian del resto por múltiples posibles motivos. Podrían ser una muestra de la señal del género literario o temática, por ejemplo. Esto no obstante no va a ocupar nuestro trabajo, aunque resulta interesante y podría analizarse en un trabajo específico. Lo que es para nosotros fundamental es que dichas partes fueron agrupadas correctamente en base a sus autores.

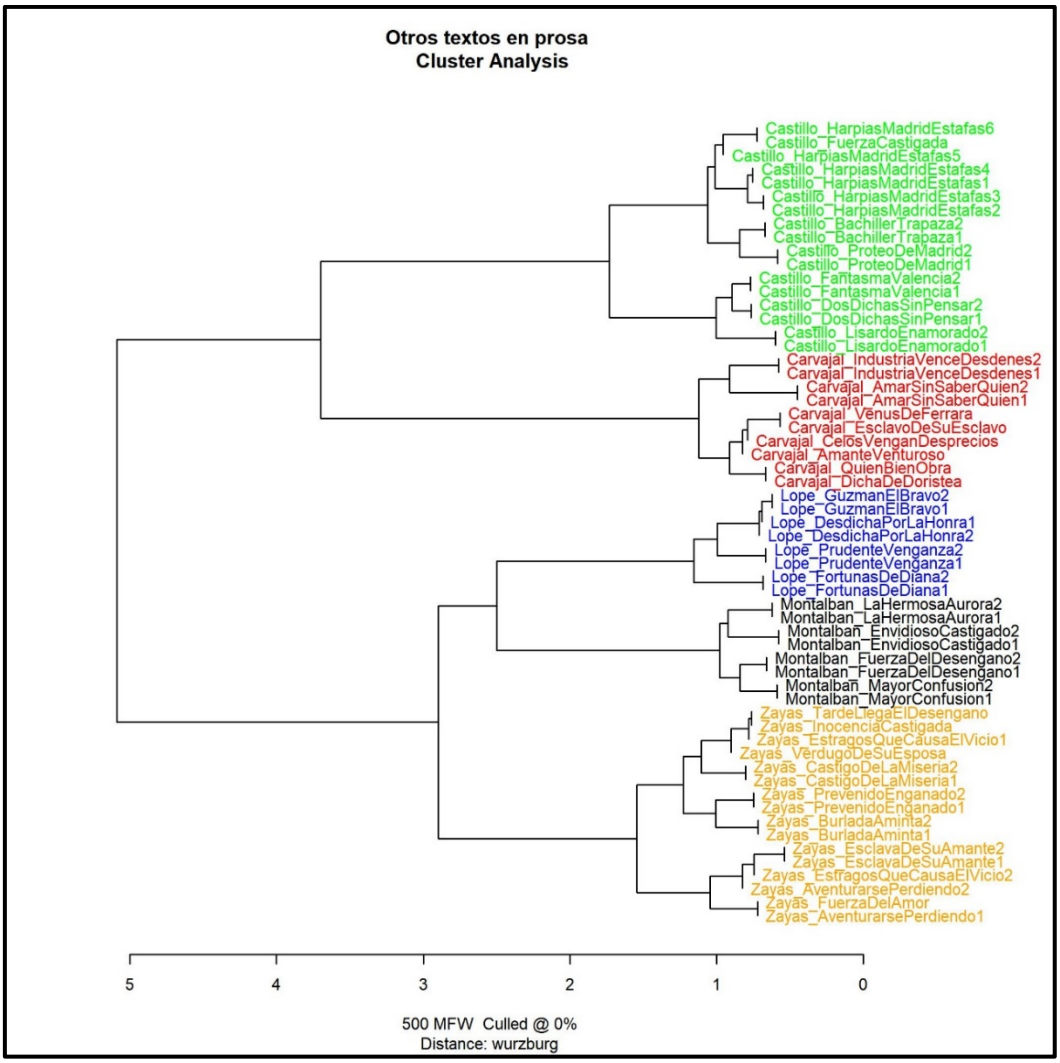
Ahora vamos a utilizar el árbol de consenso (BCT) que reúne los resultados de los dendrogramas que van desde las 100 MFW hasta las 500 MFW con el incremento especificado de 100 MFW. Esta nueva visualización nos permite observar de forma más evidente la relación entre los textos del corpus. Vemos que la clasificación de autores y obras es consistente en el sentido de que no se mezclan que como ya comentamos es lo que nos interesa en este punto. Sin embargo, también se puede observar que las ramificaciones correspondientes a Cervantes (verde claro) que salen del eje central corresponden a cada uno de los textos lo cual sugiere una gran heterogeneidad en su obra. Se podría decir que el programa reconoce a doce Cervantes. Por otra parte, la aparición de *La Galatea* a la izquierda separada del resto de su obra responde a la diferencia del género y su cercanía a la novela *Los siete libros de la Diana* de Montemayor lo respalda, pues esta novela pertenece también al género pastoril.



9 | prosa narrativa breve y extensa, *Bootstrap Consensus Tree,* 500 MFW, *Cosine Delta*.

Corpus 3: Otros textos en prosa

El último (sub)corpus que queda por organizar es el correspondiente a aquellos textos que incluyen a los autores que en un principio consideramos menos susceptibles de ser los creadores de la TF por una cuestión de cronología. Aquí realizamos nuevamente un análisis de *clúster*. Los resultados son los esperados utilizando nuevamente las 500 MFW con *Cosine Delta*:



10 | otros textos en prosa, análisis de *clúster*, 500 MFW, *Cosine Delta*.

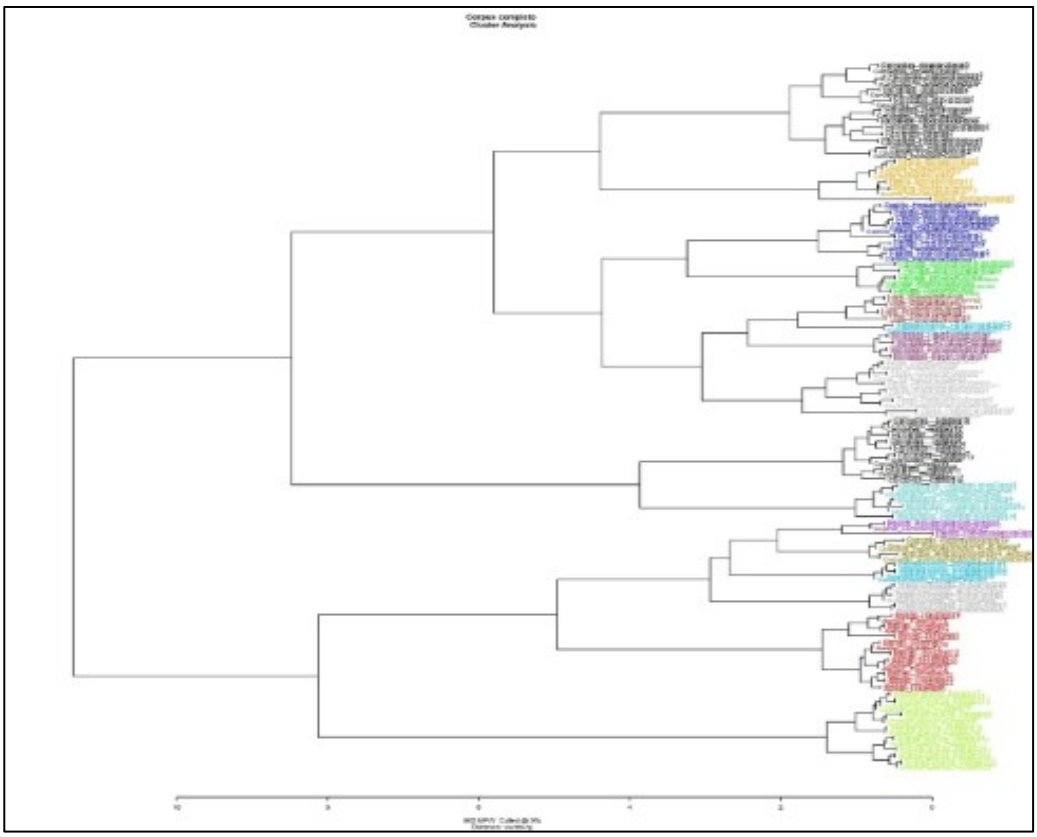
En este caso, no hay grandes sorpresas pues todos los textos son organizados en *clústeres* en los que se incluye un solo autor.

Corpus Completo

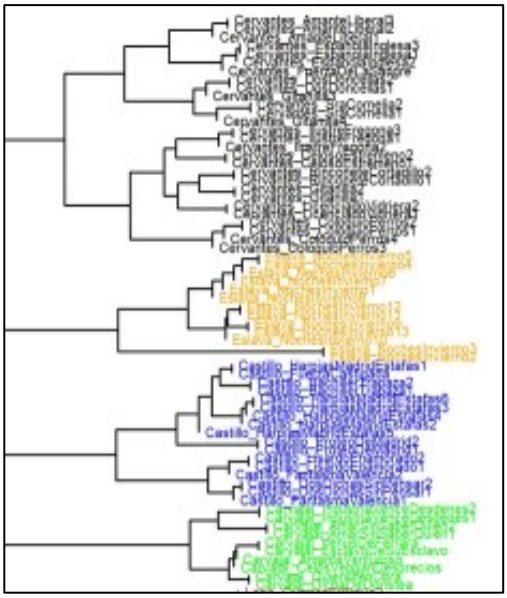
Tras esta progresiva comprobación de la organización de nuestro (sub)corpus por parte de *stylo* llegamos al que hemos llamado "corpus completo". En él incluimos todos los textos y autores de las pruebas anteriores para tratar de comprobar si la organización sigue siendo la esperada. Los resultados se vuelven más difícilmente visibles pero la organización de los textos por autores es correcta, aunque sigue existiendo cierta heterogeneidad en el corpus cervantino. Esto da lugar a que se cuelen dentro de un mismo *cluster* aunque separado en los subclusters *El patrañuelo* y *El sobremesa o alivio de caminantes* de Timoneda junto a las *Novelas Ejemplares*. También cercano encontramos el texto de Eslava, *Noches de Invierno*, pero este está algo más alejado en el eje horizontal donde se muestra que la relación aparece con respecto al *cluster* en el que están las *Novelas Ejemplares* y el texto de Eslava. Por otra parte, *La Galatea* se desmarca también y aparece en un

clúster en el que también está la obra de Montemayor. Sabemos a que se puede deber esto: se trata de dos novelas (prosa extensa) que pertenecen al género pastoril.

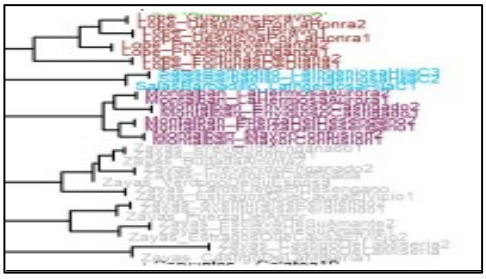
En este punto podemos plantearnos retirar la obra de Timoneda. *El sobremesa y alivio de caminantes* y *El patrañuelo*. Al hacer esto los resultados son idénticos, pero desaparece Timoneda del dendrograma:



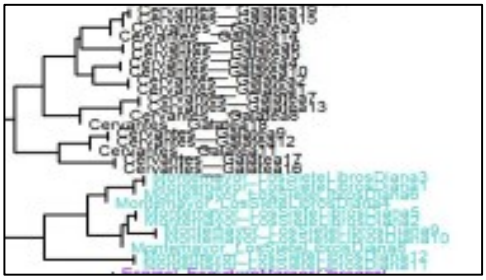
11 | corpus completo, análisis de clúster, 500MFW, *Cosine Delta*



12 | detalle 1 del dendograma (img. 11)



13 | detalle 2 del dendograma (img. 11)



14 | detalle 3 del dendograma (img. 11)

En este punto ya hemos puesto a prueba la herramienta en su función más básica y hemos comprobado que la organización de los textos en base a la figura autorial funciona correctamente con los (sub)corpora por separado y también en conjunto eliminando algunas partes y algunos textos. En el siguiente apartado vamos a proceder a comprobar la efectividad de estos resultados para luego introducir nuestro texto dubitado.

4.2 Precisión a través de la validación cruzada y clasificación

Para llevar a cabo la organización de los textos según su autor hemos utilizado la configuración de las 500 MFW y la distancia *Cosine Delta*. Sin embargo, aún nos queda responder a una pregunta de gran relevancia ¿con qué precisión se están organizando los textos? En este sentido, Eder añadió una revisión al observar que el límite fiable de las MFW puede variar según el corpus utilizado en el análisis. En base a esto, Eder (2017) llegó a la conclusión de que el factor más determinante de un texto es la fuerza de la señal autorial: a algunos les basta un número reducido de palabras mientras que otros continúan siendo dudosos pese a una mayor extensión. En nuestro caso utilizamos aún solo las 500 MFW adaptándonos a las circunstancias que impone la brevedad y la heterogeneidad de las longitudes de los textos del corpus. Las observaciones llevadas a cabo por Eder tienen como consecuencia que para poder conocer la fiabilidad de atribución de autoría para un autor es preciso llevar a cabo un análisis supervisado para comprobar el grado de fiabilidad que puede esperarse de la organización de los diferentes textos por parte de *stylo* antes de utilizarlo sobre el texto dubitado (cfr. Hernández Lorenzo 2019, 193-194). El código de este experimento puede consultarse en el anexo de Cienfuegos-Pérez (2022).

Los resultados reconocen quince clases en base a quince textos que se han tomado como referencia. Estos representan a los 14 autores que hemos seleccionado (Cervantes está repetido). El porcentaje general de acierto es del 100% con una desviación estándar del 0% utilizando desde las 100 hasta las 500 MFW con un incremento de 100 MFW. Si accedemos a los datos sobre el acierto atendiendo al número de palabras frecuentes utilizadas vemos que ya con las 100 MFW este porcentaje es del 100% y se mantiene así hasta las 500 MFW.

Sabiendo estos datos, utilizamos la función *classify ()* para la validación cruzada cuyo código y explicación se encuentra nuevamente en el anexo "código" (2) de Cienfuegos-Pérez (2022). Los resultados de esta prueba son otra vez correctos. Los autores predichos para los textos dados son los siguientes:

Aleman	Carvajal	Castillo	Cervantes	Cervantes
1	1	1	1	1
Eslava	Espinel	GregorioGonzalez	Lope	Montalban
1	1	1	1	1
Montemayor	Quevedo	SalasBarbadillo	SuarezFigueroa	Zayas
1	1	1	1	1

Tab. 1 | resultados de la validación cruzada, 500 MFW, *Cosine Delta*.

Así pues, vemos que utilizando tan solo las 500 MFW el resultado de acierto general es del 100%. El sistema ha conseguido reconocer la existencia de 15 clases correspondientes a los 14 autores. Para poder llegar a estos resultados hemos tenido que dividir los textos de aquellos autores de los cuales disponíamos de un solo ejemplo y homogeneizar las longitudes de todo el corpus. También, como ya se indicó, tuvimos que prescindir de algunos textos que causaban irregularidades e imprecisión por lo que consideramos que no nos permitirían introducir un texto dubitado en el próximo paso.

Por último, solo nos queda introducir el texto dubitado en este proceso. Así pues, repetimos la última prueba, pero esta vez introduciendo el texto dubitado en la carpeta *secondary_set* que es la que contiene las muestras de diferentes autores que van a ser comparadas con el *primary_set* conteniendo numerosos textos de ejemplo que le sirven al programa para "aprender" el estilo. Los resultados de las clases esperadas y predichas se pueden ver en la siguiente tabla:

Esperado ("expected")	Predicho ("predicted")
[1] "Aleman"	[1] "Aleman"
[2] "Carvajal"	[2] "Carvajal"
[3] "Castillo"	[3] "Castillo"
[4] "Cervantes"	[4] "Cervantes"
[5] "Cervantes"	[5] "Cervantes"
[6] "Desconocido"	[6] "Cervantes"
[7] "Eslava"	[7] "Eslava"
[8] "Espinel"	[8] "Espinel"
[9] "GregorioGonzalez"	[9] "GregorioGonzalez"
[10] "Lope"	[10] "Lope"
[11] "Montalban"	[11] "Montalban"
[12] "Montemayor"	[12] "Montemayor"
[13] "Quevedo"	[13] "Quevedo"
[14] "SalasBarbadillo"	[14] "SalasBarbadillo"
[15] "SuarezFigueroa"	[15] "SuarezFigueroa"
[16] "Zayas"	[16] "Zayas"

Tab. 2 | resultados de la función *classify*, corpus + La TF, 500 MFW, *Cosine Delta*

Como se puede ver en la tabla, el texto *Desconocido* que no es otro que *La tía fingida*, ha sido clasificado como perteneciente a Cervantes. Por lo tanto, el resultado que arroja la prueba de atribución de autoría con la función *classify* es que entre los autores seleccionados el texto dubitado presenta mayor similaridad estilística con Cervantes y por lo tanto parece pertenecer a este autor.

4.3 Verificación de autoría

Llegados a este punto en el que hemos comprobado la efectividad y hemos visto que la TF es atribuida a Cervantes podemos continuar con las pruebas en este caso de verificación de autoría: suponemos que el autor es Cervantes, pero queremos ponerlo a prueba. Esto sirve también como validación cruzada de los resultados anteriores. Para realizar este experimento vamos a utilizar otra característica de

stylo denominada *General Imposters* (GI) [14] también conocida como segundo sistema de verificación (o2). Kestemont et al. (2016) apuntan que esta característica fue introducida por Koppel y Winter (2014) y aplicada en el estudio de los escritos disputados de Julio Cesar.

Para ello seguimos el *script* que propone Eder (2018) adaptándolo a nuestro corpus (ver anexo "corpus" en Cienfuegos-Pérez 2022). La primera de las pruebas consiste en comparar las frecuencias relativas de las MFW de la TF con todos los demás textos a través de la función *imposters()*[15]. Utilizamos la medida de distancia por defecto que es *Delta de Burrows* obteniendo los siguientes datos:

Alemán	Carvajal	Gregorio Gonzalez	Lope
0.00	0.00	0.00	0.00
Castillo	Cervantes	Montalban	Montemayor
0.02	0.98	0.00	0.00
Eslava	Espinel	Quevedo	Salas Barbadillo
0.00	0.02	0.00	0.01
Suarez Figueroa	Zayas		
0.02	0.05		

Tab. 3 | resultados de la función *imposters* con Delta de Burrows.

Para entender los coeficientes solo hace falta saber que la atribución se da en los valores cercanos a 1 y se rechaza en el 0. Los resultados son los que nos da el algoritmo al tratar de comparar un texto anónimo determinado con un conjunto de textos de candidatos entre los que se encuentra también el probable autor. Dado que no indicamos otra cosa, el método trata de comprobar uno tras otro todos los autores disponibles que considera como candidatos potenciales. Sin embargo, también podemos focalizarnos en un solo autor haciendo que el método compare el texto dubitado con todos los textos de un candidato. Así pues, le indicamos al programa que queremos hacer esto. Lo llevamos a cabo de manera iterativa y con todos los autores siguiendo el consejo de Eder (2018) de forma que los resultados nos sirvan también a modo de validación cruzada. Tras llevar a cabo este laborioso proceso, vemos que los valores de las ocho pruebas varían levemente con respecto a la anterior destacando Zayas que obtiene un 0,15 en vez de 0,05, el resto de resultados de esta validación corresponden con los dados en el primer experimento con una desviación de 0,01/ 0,02 (ver tabla 5).

[14] Lo que hace esta característica es en palabras de Kestemont et al. (2016, 88): "La [..] GI no consiste en evaluar si dos documentos son simplemente similares en cuanto al estilo de escritura, dado un vocabulario de características estático, sino que pretende evaluar si dos documentos son significativamente más similares entre sí que otros documentos, a través de una variedad de espacios de características estocásticas y en comparación con selecciones aleatorias de los llamados autores distractores (Juola, 2015), también llamados 'impostores'."

[15] El script con el código tras esta función puede consultarse en el Github del Computational Stylistics Group: <https://github.com/computationalstylistics/stylo/blob/master/R/imposters.optimize.R>.

Alemán	0	Eslava	0.01		
Carvajal	0	Espinel	0.01	Montemayor	0
Castillo	0.02	Gregorio González	0.01	Quevedo	0
Cervantes	0.97	Lope	0	Salas Barbadillo	0.03
Suarez F.	0.01	Zayas	0,15	Montalbán	0

Tab. 4 | resultados de la función imposters en las pruebas individuales para cada autor.

En este caso solo hay un valor que se acerque a 1.00 (atribución) y es Cervantes. El único candidato (candidata) que parece mostrar cierta similitud es Zayas. Sigamos pues con el experimento para ver si estos resultados son estables. Realizaremos de nuevo la primera prueba utilizando la función *imposters*, pero esta vez vamos a emplear *Wurzburg Delta* (*aka Cosine Delta*) que es la medida de distancia que hemos aplicado desde los primeros experimentos pues según Eder (2018) muestra una mejoría notable con respecto a las otras (*Burrow's Delta, Eder's Delta*). También incluiremos la medida *Eder Delta* para poder contrastar los resultados. Tras realizar este proceso obtenemos los siguientes datos:

Autores	Wurzburg (Cosine)	Eder	Burrows
Alemán	0.03	0.02	0.00
Carvajal	0.00	0.00	0.00
Castillo	0.00	0.02	0.02
Cervantes	0.96	0.95	0.98
Eslava	0.00	0.04	0.01
Espinel	0.00	0.02	0.02
Gregorio González	0.00	0.01	0.00
Lope	0.00	0.01	0.00
Montalbán	0.00	0.00	0.00
Montemayor	0.00	0.00	0.00
Quevedo	0.05	0.00	0.00
Salas Barbadillo	0.00	0.00	0.01
Suarez Figueroa	0.00	0.01	0.02
Zayas	0.04	0.13	0.05

Tab. 5 | resultados de imposters con diferentes distancias, *Cosine Delta*, *Eder's Delta y Burrow's Delta.*

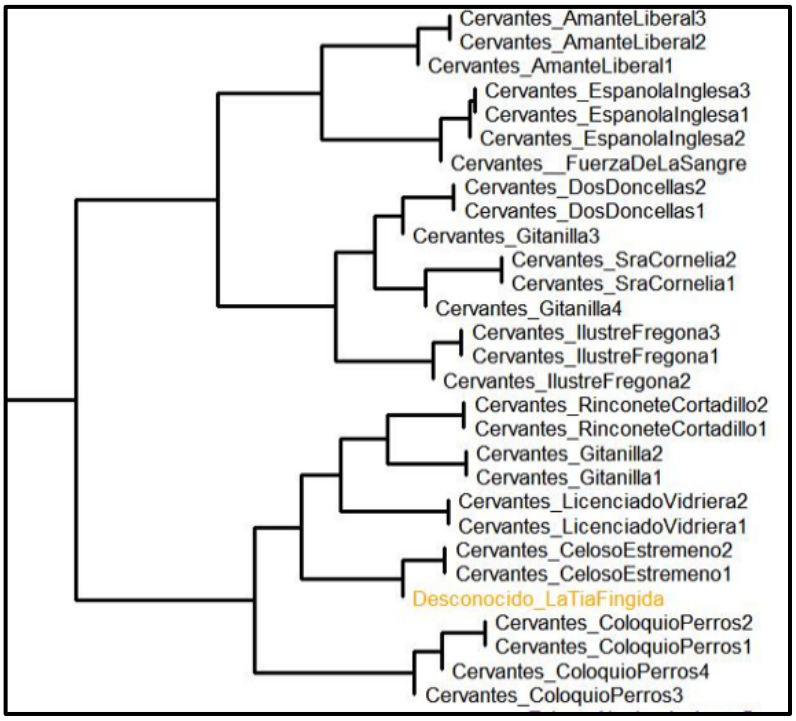
Para explicar los resultados, volvemos a la documentación de Eder (2018) quien afirma que cualquier resultado por encima de 0,5 puede ser indicativo de que la verificación de autoría para un candidato dado fue exitosa. Entonces, con los datos que obtuvimos podemos concluir que Cervantes es siempre reconocido como el autor más cercano a la obra dubitada con todas las distancias estadísticas utilizadas. No obstante, existen algunos autores que presentan un cierto grado de similitud. En un principio podríamos obviar este hecho dada la notable diferencia que presentan con respecto al alcalaíno, no obstante, hemos querido probar esta función para comprobar los resultados que nos da para Cervantes. Gracias a ella, ahora sabemos que utilizando diferentes distancias (*Eder's Delta, Burrow's Delta,*

Würzburg Delta) hay una sola atribución con una alta probabilidad pues la función *imposters()* apunta a que *La tía fingida* es de Cervantes. También vimos que la función *classify()* reconocía que el texto dubitado es del complutense. En definitiva, tanto la atribución como la verificación de autoría dan como resultado que el autor de esta novela es el Manco, aunque hay que insistir que en este corpus no están todos los posibles autores y textos que se podrían tener en consideración debido en parte a las limitaciones ya comentadas.

Por lo tanto, la única posible pega que vimos es que el programa parece tener algún problema para descartar de forma definitiva a varios de los autores destacando entre ellos Zayas que se queda siempre en la zona "de grises"[16]. Sin embargo, la atribución cervantina es tan constante y clara que no consideramos esto un impedimento. Además, en este punto conviene recordar el apunte que hace Eder (2018) con respecto a los resultados "cuando realizas el test nuevamente, el resultado final puede diferir levemente debido a la naturaleza estocástica del test". Por este motivo consideramos que debemos continuar indagando un poco. Además, nos interesa descubrir con qué obras se establecen los nexos entre *La tía fingida* y *Las Novelas Ejemplares* para poder realizar el análisis literario que nos permita comparar el estilo desde esa perspectiva.

En primer lugar, vamos a realizar un dendrograma utilizando el "corpus completo" y empleamos la función *stylo()* con las 500 MFW y la distancia *Cosine Delta*. Habíamos dejado para el final la comprobación de los resultados incluyendo a *La tía fingida*. La imagen que muestra el dendrograma es idéntica a la que se mostraba en el "corpus completo". La diferencia estriba en Cervantes, donde nos encontramos con *La tía fingida* que, aunque está sola en una ramificación, aparece asociada en un mismo *subcluster* con *El celoso extremeño*:

[16] Estas mismas pruebas fueron realizadas utilizando los textos completos (sin fragmentar) y en un rango de MFW más alto (3200). Esto nos obligaba a dejar fuera algunos autores de los cuales solo tenemos una obra, pero en ese caso la posible atribución a Zayas quedaba en todos los casos por debajo del rango inferior que daba la función *imposters.optimize()* y por lo tanto descartada.

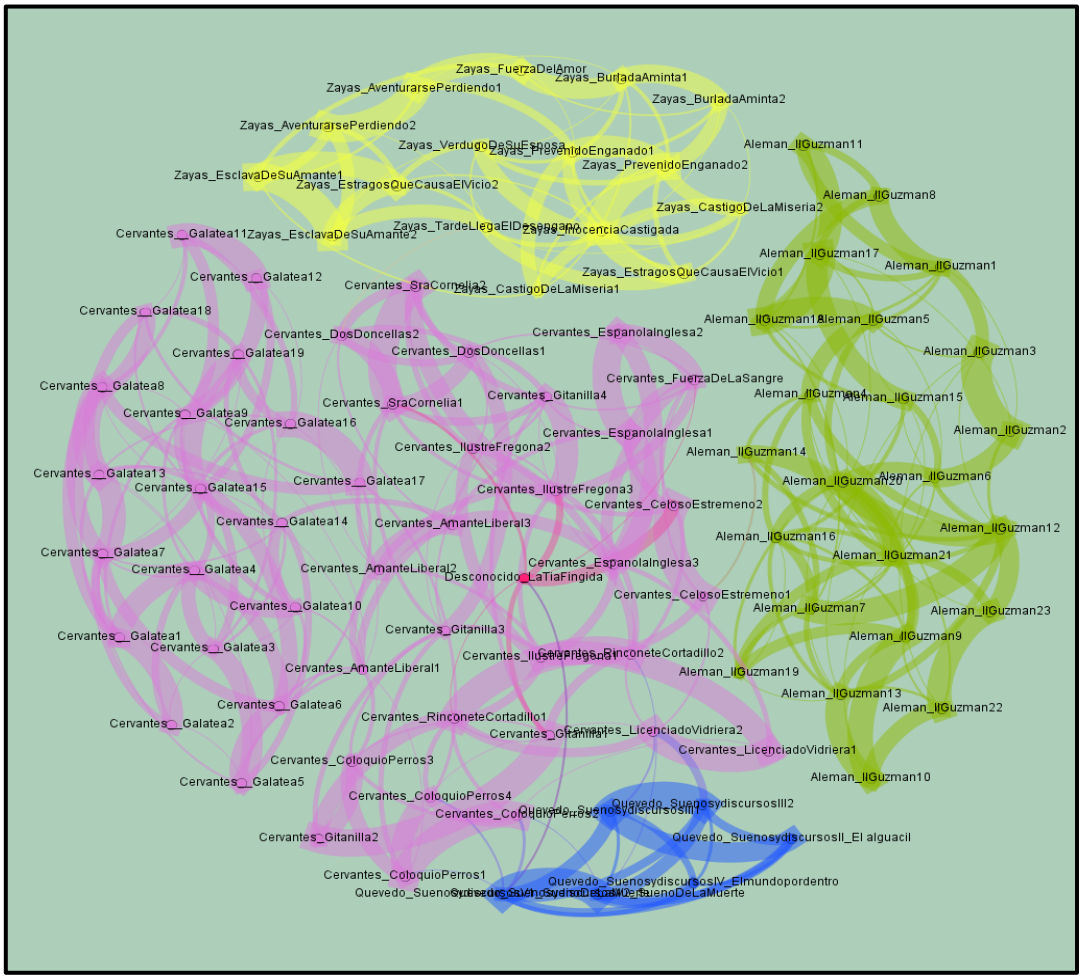


16 |detalle (obras de Cervantes) del análisis de clúster con el corpus "completo", 500 MFW, *Cosine Delta*.

Para seguir con la comprobación de los resultados podemos echar un vistazo a las relaciones entre todos los autores y el texto dubitado. Esto lo vamos a realizar en el siguiente apartado donde llevaremos a cabo un experimento que nos permitirá visualizar los parentescos entre textos.

4.4 Stylo.network () y Gephi

En la última prueba se puso de manifiesto que parece existir una relación entre la obra cervantina y *La tía fingida*. Además, sabemos que Zayas, Quevedo y Alemán parecen tener un cierto grado mínimo de similitud estilística con Cervantes. El problema es que no conocemos donde residen estas relaciones. Para obtener esta información ya hemos utilizado el dendrograma y este nos muestra como resultado un parentesco con *El celoso extremeño*. Ahora nos queda por comprobar un segundo método de visualización: el *Bootstrap Consensus Tree* (BCT) a partir de la frecuencia relativa de las 100 a las 500 MFW utilizando aquellos autores que la función *imposters()* no acababa de descartar del todo. La función *stylo()* nos va a servir para crear el gráfico. Vamos a utilizar la función *stylo.network()* con la que creamos un BCT cuyos resultados exportaremos a *Gephi* (cf. Bastian, Heymann y Jacomy 2009). Se trata de dos tablas de Excel en formato csv conteniendo los datos de los nodos y las aristas. Con ello empleamos el *layout* de Fruchterman y Reingold (1991) con el que, tras ajustar los parámetros de visualización, obtenemos las siguientes imágenes:



17 | gráfico *Gephi* del BCT de cuatro autores (Zayas, Cervantes, Quevedo, Alemán) con distancia *Cosine Delta* (500 MFW)

18 | detalle de la imagen 17, *La tía fingida* y sus nexos.

A través de este gráfico podemos ver con mayor claridad las relaciones que tiene la TF con otros textos en base a los parámetros que seleccionamos con la función *stylo()*. Vemos efectivamente que no tiene ningún nexo con Zayas, pero sí alguna conexión con el primer fragmento de la quinta parte de los *Sueños y discursos*, "el sueño de la muerte" de Quevedo. Sin embargo, las relaciones con las *Novelas Ejemplares* de Cervantes son mucho más notorias y variadas entre las que destaca una relación que sobresale por encima del resto: se trata de la segunda parte de *El celoso extremeño*.

En este punto dada la evidente conexión que aparenta tener *La tía fingida* con el de *El celoso extremeño* procedimos en este punto a llevar a cabo el análisis literario de ambas obras y su confrontación con el objetivo de poder descifrar algunos de los elementos que se esconden tras los datos que hemos visto hasta ahora. Los resultados de este análisis pueden ser consultados en Cienfuegos-Pérez (2022).

5. Conclusiones

En nuestros experimentos con *stylo* procuramos que la organización de textos indubitados fuera hecha por autor para anteponer la relevancia del empleo de determinadas palabras funcionales a cualquier otro aspecto como marcador del estilo autorial. A través de los experimentos vimos que la organización de los textos por autor de las obras no dubitadas era correcta[17]. Al introducir *La tía fingida* en el corpus después de realizar la validación y demostrar que el porcentaje de fiabilidad con las 500 MFW y la distancia *Cosine Delta* era del 100% vimos que esta obra era asociada a Cervantes y más particularmente a *El celoso extremeño*. Además, llevamos a cabo una serie de experimentos estilométricos utilizando diferentes funciones para determinar la autoría. El primero de ellos fue de atribución de autoría (función *classify*) seguido de la verificación de autoría (función *imposters*). En ambos casos la TF fue asociada a Cervantes sin que otro autor mostrara resultados tan evidentes. Asimismo, el empleo de diferentes distancias estadísticas como *Delta de Burrows* y *Delta de Eder* mostró que la opción de la autoría cervantina es constante. Para ver dónde se encontraban los nexos entre las obras de Cervantes y otros autores de forma más concreta se utilizaron los parámetros ya indicados, se empleó la función *stylo.network* con la que se creó un BCT (*Bootstrap Consensus Tree*) y se exportaron los datos a *Gephi* obteniendo así un gráfico que evidenció que la relación entre *La tía fingida* y *El celoso extremeño* era más notoria que otras. En este punto, decidimos analizar las dos obras que presentaban la interrelación más destacada.

En el análisis literario nos centramos sobre todo en aspectos técnicos y expresivos que no se pueden medir cuantitativamente como son los contenidos, y las figuras retóricas. Pudimos observar que en el plano de los contenidos ambas obras pertenecen a un género, tipo y temática similar y utilizan técnicas narrativas semejantes. En el análisis de la expresión se reconocieron también algunas

[17] Es necesario precisar que, aunque no se hayan incluido en el trabajo, se llevaron a cabo numerosas pruebas previas utilizando los textos sin fragmentar, con diferentes configuraciones de las MFW y utilizando diversas medidas de distancia. Es así como llegamos hasta los parámetros que se incluyen en los experimentos finales.

semejanzas a través del uso de figuras retóricas compartidas, pero se evidenciaron divergencias notorias que están recogidas en Cienfuegos-Pérez (2022) dentro del apartado "Síntesis del análisis literario" y que se podrían englobar en la afirmación: *El celoso extremeño* se sirve de recursos que presentan una mayor complejidad comparados con los de *La tía fingida*.

Llegados a este punto solo nos queda responder a la gran pregunta ¿quién fue el autor de *La tía fingida*? Pues bien, hay determinados aspectos a considerar antes de hacer cualquier afirmación terminante sobre su identidad. En primer lugar, la elección del corpus no incluye todos los textos que serían ideales para una investigación de este tipo ni tampoco todos los autores que se podrían considerar. Además, la fiabilidad de los textos es solo relativa pues no cumplen siempre la premisa de poseer muestras incorruptas que señalaba Juola (2008). La problemática de los textos del Siglo de Oro muestra la necesidad de seguir trabajando en la calidad, la digitalización y la puesta a disposición de los mismos para el ámbito académico.

Además de esto, la decisión de contrastar determinados textos y no otros en el experimento con *classify* no es 100% objetiva. En el conjunto de los textos a comprobar (*secondary_set*, o *test_set*) elegimos ciertas obras para ver si el algoritmo las reconocía y las asignaba a sus respectivos autores, pero ¿Qué pasa si probamos con textos que se consideren estilísticamente más lejanos a un autor? Desafortunadamente la herramienta no ofrece ninguna solución a esto más allá de realizar la prueba con diferentes conjuntos de textos en el *secondary_set*. [18]

Entendidos estos escollos, podemos entrar en consideraciones. En base a las pruebas estilométricas solo tenemos una respuesta posible: *La tía fingida* es con una alta probabilidad obra de Cervantes. Sin embargo, en vista de los límites ya comentados lo que podríamos afirmar es más bien que: de entre este conjunto de autores y con el corpus seleccionado el único autor posible es Cervantes. Por otra parte, desde la perspectiva literaria las dos obras comparadas mostraron convergencias y divergencias. En la obra no dubitada se pudo observar una obra técnicamente más elaborada ¿Se puede desde el análisis literario de ciertas figuras retóricas y algunos elementos morfológicos decidir la autoría de la obra? Consideramos que no, tampoco era nuestro objetivo, pero sí sirve en conjunto con los experimentos estilométricos para llegar a ciertas hipótesis.

La primera de estas hipótesis es que esta obra podría ser una novela que Cervantes compuso con anterioridad a sus otras novelas breves, un periodo en el cual el autor aún estaba puliendo su pluma en este género. No olvidemos que la fecha de escritura aproximada que manejamos es la de la compilación de los manuscritos que realizó Porras de la Cámara para el cardenal Niño de Guevara (1604-1606). Por lo tanto, no sabemos a ciencia cierta cuando fue redactada ni en qué orden cronológico con respecto a los otros textos aparecidos en el manuscrito Porras.

[18] Evidentemente, realizamos pruebas con diferentes "sets" sin embargo consideramos que la selección es siempre arbitraria. De todos modos, las diferentes configuraciones clasificaban los textos de autor conocido correctamente y atribuían la TF a Cervantes. En este contexto la utilización de textos *estilísticamente* lejanos al autor puede servir en la investigación para validar los resultados.

Esto explicaría las coincidencias a la vez que visto en perspectiva podría servir como ejemplo de análisis de la evolución del alcalaíno. Así pues, aunque compartimos aquí su hipótesis de la autoría cervantina diferimos de Marín (1944, 40) en que la novela sea excelente, aunque emplea un vocabulario variado, numerosas técnicas expresivas y representa una buena muestra del género de la novela breve primigenia. Consideramos que no tiene el grado de complejidad y excelencia en el uso de los recursos técnicos y expresivos de los que sí hace gala *El celoso extremeño*. Concordamos por lo tanto en parte con la opinión de Foulché-Delbosc (1899) al reconocer cierta distancia estilística con respecto a Cervantes, pero al contrario que el hispanista francés nos parece que esto no sirve como argumento para no considerarla del complutense, no vemos tanta distancia como para poder afirmar rotundamente que no es obra suya.

La segunda presunción es que la obra fuera de un temprano imitador como hipotetizaba Márquez Villanueva (1995). Sin embargo, aunque un imitador pudiera emplear el vocabulario y las expresiones más características de Cervantes resulta poco creíble que prestara atención a todas las palabras funcionales y su frecuencia relativa. Opinamos por lo tanto que la posibilidad de un imitador no se puede descartar utilizando solamente la comparación de determinadas expresiones, pasajes textuales e índices verbales. Estas, en nuestra opinión, no sirven tampoco como argumentos definitivos para desmentir u otorgar la autoría de la TF como hicieron por ejemplo Icaza (1916) y De Val (1953) para rechazarla como obra de Cervantes o Madrigal (2003) para adjudicársela.

Otro aspecto relevante que nos queda por comentar es el de justificar para qué serviría considerar la TF de Cervantes ¿qué aporta esto a la literatura? Aunque desde nuestro punto de vista la obra nos parezca más prosaica que el resto de novelas cortas, el conocimiento de sus parecidos y diferencias nos da cierta perspectiva sobre la evolución del autor. También vemos en sus contenidos crítica social, dicotomía de espacios casa-calle, una serie de ideas sobre la mujer, imagen del mundo estudiantil, entorno del hampa y otras múltiples posibilidades de análisis que podrían ser contrastadas con el resto de su corpus conformando en su totalidad el significado último de determinados propósitos e ideas, el empleo y desarrollo de ciertas técnicas narrativas, etc. No olvidemos que el genial Cervantes es considerado el creador de la novela breve española[19] por lo que saber más sobre el proceso y la evolución de este género es descubrir también las estrategias que se esconden tras su nacimiento.

Llegamos pues al final habiendo visto que nuestra hipótesis inicial no estaba desencaminada. La estilometría ha demostrado ser una herramienta útil en el proceso de atribución de autoría, que pese a sus límites nos ha ayudado a adoptar nuevas perspectivas literarias. Sin ella no hubiéramos encontrado con tanta rapidez paralelismos con *El celoso extremeño*. La hipótesis de la autoría cervantina de la TF a pesar de que haya sido desdeñada por algunos críticos que no reconocían en esta obra la genialidad de Cervantes parece tener bastante sustento al menos mientras no encontremos una alternativa con textos que podamos contrastar y que presente

[19] Por ejemplo, Díez-Echarri y Roca Franquesa (1972, 222) o Montero Reguera (2006, 166)

unos resultados semejantes a los arrojados por el complutense en nuestros experimentos. Hasta entonces, si nos decantamos por la primera de las sospechas, la de la novela cervantina de composición temprana, tendríamos que admitir que quizás la excelencia de los trabajos conocidos del ingenioso Miguel de Cervantes es fruto de la evolución y mejoría del mismo o dicho de otro modo que sin esfuerzo y trabajo ni el gran talento de Miguel de Cervantes Saavedra hubiera alcanzado la calidad de las *Novelas ejemplares* ni dado vida al ingenioso caballero y a su leal escudero que viven hoy en el mundo entero.

Bibliografía

ARELLANO-AYUSO, Ignacio. 1997. "Las aventuras del texto: del manuscrito al libro en el Siglo de Oro." En *Unum ET Diversum: Estudios en honor de Ángel-Raimundo Fernández González*, ed. Spang, Kurt, 41-66, Pamplona: Ediciones Universidad de Navarra.

BASTIAN, Mathieu, Sebastien Heymann & Mathieu Jacomy. 2009. "Gephi: an open-source software for exploring and manipulating networks." *Proceedings of the international AAAI conference on web and social media* 3 (1), 361-362.

BLASCO PASCUAL, Francisco & Cristina Ruiz Urbón. 2009. "Evaluación y cuantificación de algunas técnicas de atribución de autoría en textos españoles." *Castilla: Estudios de literatura* 0: 27-47. <https://doi.org/10.24197/cel.0.2009> (09.10.21).

CANAVAGGIO, Jean. 1992. *Cervantes: en busca del perfil perdido*. Besalú: Llibres Detot.

CEREZO SOLER, Juan & 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.

CERVANTES SAAVEDRA, Miguel. 2018. *La tía fingida*. Ediciones Cátedra.

CERVANTES SAAVEDRA, Miguel. 1818. *La tia fingida: novela inédita*. Edición digital basada en C. F. Franceson y F. A. Wolf, Berlín: Librería de G. C. Nauck. <https://www.cervantesvirtual.com/nd/ark:/59851/bmc1z4h1>.

CERVANTES, Miguel. 2018. *Novelas ejemplares*. MVP.

CIENFUEGOS PÉREZ, Alejandro (2022). *Atribución de autoría y Humanidades Digitales: Métodos de estilometría y aplicación al Siglo de Oro*. Tesis de Master, Georg-August-Universität Göttingen. DARIAH-DE. <https://doi.org/10.20375/0000-000f-3246-a>.

DE ICAZA, Francisco A. 1916. *De cómo y por qué "La tía fingida" no es de Cervantes: y otros nuevos estudios cervánticos*. Madrid: Imprenta Clásica Española.

DE VAL, Manuel Criado. 1953. *Análisis verbal del estilo: índices verbales de Cervantes, de Avellaneda y del autor de "La tía fingida"*. Madrid: Consejo superior de investigaciones científicas.

DESAGULIER, Guillaume. 2017. *Corpus linguistics and statistics with R*. Berlin: Springer International Publishing.

DÍEZ ECHARRI, Emiliano & José María Roca Franquesa. 1972. *Historia de la literatura española e hispanoamericana*. Madid: Aguilar.

EDER, Maciej. 2018. "Authorship verification with the package *stylo*." *Computational Stylistics Group Blog*. <https://computationalstylistics.github.io/blog/imposters/> (18.11.2022).

EDER, Maciej. 2017. "Short Samples in Authorship Attribution: A New Approach." *Digital Humanities. Montreal 8.-11.8.2017.* <https://dh2017.adho.org/abstracts/341/341.pdf> (18.11.2022)

EDER, Maciej. 2015. "Does size matter? Authorship attribution, small samples, big problem." *Digital Scholarship in the Humanities* 30 (2), 167-182.
<https://doi.org/10.1093/llc/fqt066> (22.11.22)

EDER, Maciej & Jan Rybicki. 2011. "Stylometry with R." *Digital Humanities*, 308-310.

EDER, Maciej, Jan Rybicki & Mike Kestemont 2016. "Stylometry with R: a package for computational text analysis." *The R Journal* 8 (1), 107–121.
<https://journal.r-project.org/archive/2016/RJ-2016-007/index.html> (15.08.21).

FOULCHE-DELBOSC, Raymond. 1899."Etude sur 'La tía fingida': I. le manuscrit Porras.- II. le manuscrit de la Colombine.- III. le texte de La tía fingida.- IV. l'attribution a Cervantes." *Revue hispanique: recueil consacré à l'étude des langues, des littératures et de l'histoire des pays castillans, catalans et portugais* 6 (19), 256-306.

FRUCHTERMAN, Thomas MJ & Edward M. Reingold. 1991. "Graph drawing by force-directed placement." *Software: Practice and experience* 21 (11), 1129-1164.

GALLARDO, Bartolomé José. 1835. "La tía fingida ¿es novela de Cervantes?" *El Criticón, papel volante de Literatura y Bellas artes* 1, 1-43.

GRIES, Stefan Th. 2016. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

GRIES, Stefan Th. 2013. *Statistics for Linguistics with R. A practical introduction*. Berlin/Boston: De Gruyter Mouton.

HERNÁNDEZ-LORENZO, Laura. 2019. "Poesía áurea, estilometría y fiabilidad: métodos supervisados de atribución de autoría atendiendo al tamaño de las muestras." *Caracteres: estudios culturales y críticos de la esfera digital* 8 (1), 189-228.

HOLMES, David I. & Judit Kardos."Who was the author? An introduction to stylometry." *Chance* 16 (2), 5-8.

JOCKERS, Matthew L. 2014. *Text Analysis with R for Students of Literature*. Cham: Springer International Publishing.

JUOLA, Patrick. 2015. "The Rowling case: a proposed standard analytic protocol for authorship questions." *Digital Scholarship in the Humanities* 30 (suppl_1), i100-i113.

JUOLA, Patrick. 2008. "Authorship attribution." *Foundations and Trends® in Information Retrieval* 1 (3), 233-334.

KESTEMONT, Mike et al. 2016. "Authenticating the writings of Julius Caesar." *Expert Systems with Applications* 63, 86-96.

KESTEMONT, Mike. 2011. "Een stylometrisch onderzoek naar Jan van Boendales auteurschap voor de Brabantse yeesten." *Revue belge de Philologie et d'Histoire* 89 (3), 1019-1048.

KOPPEL, Moshe & Yaron Winter. 2014. "Determining if two documents are written by the same author." *Journal of the Association for Information Science and Technology* 65 (1), 178-187.
<http://dx.doi.org/10.1002/asi.22954> (18.11.2022).

LEVSHINA, Natalia. 2015. *How to do linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam/Philadelphia: Benjamins.

LUCÍA MEGÍAS, José Manuel. 2018. "El códice Porras (casi) recuperado (la copia del Cigarral del Carmen de" "La tía fingida"). " *Anales cervantinos* 50, 333-351.

MADRIGAL, José Luis. 2003. "De cómo y por qué La tía fingida es de Cervantes." *Artifara* 2, s. p.

MARÍN, Luis Astrana. 1944. "Sobre La tía fingida" en *Cervantinas: y otros ensayos*. Vol. 6. Afrodisio Aguado, sa, 1944. 37-48.

MÁRQUEZ VILLANUEVA, Francisco. 1995. *Trabajos y días cervantinos*. Vol. 2. Biblioteca Estudios Cervantinos.

MAREGALLI, Franco. 1992. *Introducción a Cervantes* Vol. 111. Barcelona: Ariel.
MONTERO REGUERA, José. 2006. "El nacimiento de la novela corta en España (la perspectiva de los editores)." *Lectura y signo* 1, 165-175.
RICO, Francisco. 2000. *Imprenta y crítica textual en el Siglo de Oro, ed. lit. Pablo Andrés Escapa y Sonia Garza,* Valladolid, Centro para la Edición de los Clásicos Españoles.
RIßLER-PIPKA, Nanette. 2018. "Die Digitalisierung des goldenen Zeitalters – Editionsproblematik und stilometrische Autorschaftsattribution am Beispiel des Quijote Abstracts." *Zeitschrift Für Digitale Geistes-wissenschaften* 2018, s.p.
<https://zfdg.de/2018_004_v1> (18.11.2022).
RUEDA, José Manuel. 2019 "Estilometría y la Edad Media castellana" En *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania (=Romanische Studien*, Beiheft 6), ed. Rissler-Pipka, Nanette, 49-74.
RYBICKI, Jan and Maciej Eder. 2012. "Stylometry: Computer-Assisted Analysis of Literary Texts." Workshop *Culture & Technology, European Summer School in Digital Humanities*. Leipzig University.
WINTER, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York: Routledge.

Páginas web y recursos en línea.

Nota: los enlaces a las versiones digitales de los libros empleados para el corpus se encuentran en el anexo "Corpus" en Cienfuegos-Pérez (2022).
BVMC (Biblioteca virtual Miguel de Cervantes), Guía de obras atribuidas a Miguel de Cervantes Saavedra (15.06.21). En línea:
<http://www.cervantesvirtual.com/portales/miguel_de_cervantes/estados_atribuciones/>.
Los gráficos Gephi fueron creados con el software de la página:
<https://gephi.org/>.
Los textos digitales del subcorpus *otros textos en prosa* han sido recuperados de Fradejas Rueda, 7 Partidas Digital, Github:
<https://github.com/7PartidasDigital>.
ForTEXT & CATMA. „Tutorial: Stylo zur Analyse des Autoren-Stils nutzen" forTEXT & CATMA. 1 de agosto de 2019. Video, 12m22s.
<https://youtu.be/LuJe67898z0>.
Tutorial para el empleo de Stylo: Rueda, José Manuel Fradejas. "*Cuentapalabras. Estilometría y análisis de texto con R para filólogos*." (2020) Recuperado el 08.08.21 de:
<http://www.aic.uva.es/cuentapalabras/>.

Resumen

El presente trabajo expone el empleo de algunas de las herramientas que ofrecen las humanidades digitales para tratar de aportar nuevas perspectivas a un caso de autoría largamente discutido: el de la novela breve *La tía fingida* atribuida a Cervantes. Para llevar a cabo dicha tarea se mostrará el proceso de compilación del corpus y algunos de los problemas de los textos del Siglo de Oro. A través del uso de la herramienta *stylo* se presentarán los diferentes métodos y experimentos llevados a cabo, tanto de atribución como de verificación de autoría. Los resultados de la estilometría mostraron que *El celoso extremeño* y *La tía fingida* parecen tener un parentesco estilístico. Para tratar de confirmarlo se contrastarán ambas obras desde la perspectiva estilométrica y la literaria.

Abstract

This paper presents a series of experiments using tools offered by the digital humanities to bring new perspectives to a long-discussed case of authorship attribution: regarding the short novel *La tía fingida* ('The Pretended Aunt') attributed to Cervantes. The process of compilation of the corpus and some of the general problems of authorship attribution in the Spanish Golden Age are also discussed. By trying out the method of stylometry and using *stylo* as a tool, the different functions and experiments carried out, both for attribution and authorship verification, are presented. The results show that *El celoso extremeño* and *La tía fingida* seem to have a stylistic relatedness. To try to confirm this, both works will be contrasted from a stylometric and literary criticism perspective.