Nuevas pruebas estilométricas para negar la autoría cervantina del *Entremés de los romances*

Cristina Ruiz Urbón[1]
Universidad de Valladolid / Universidad de Salamanca
cristina.ruiz.urbon@uva.es / cristinaruizurbon@usal.es

Recepción: 09/03/2023, Aceptación: 24/11/2023, Publicación: 31/12/2023

Resumen
En este trabajo ofrecemos una serie de análisis cuantitativos que avalan la autoría no cervantina del *Entremés de los romances*, basados en la comparación de ese texto con treinta y dos entremeses del Siglo de Oro español, de diez autores distintos, entre los que se encuentran los ocho entremeses indubitados del autor del *Quijote*.

Palabras clave
Cervantes; *Entremés de los romances*; atribución de autoría; estilometría; teatro del Siglo de Oro.

Abstract
English title. New stylometric tests to deny the *Entremés de los romances* cervantine authorship.
In this paper we offer a series of quantitative analyzes that support the non-Cervantine authorship of the *Entremés de los romances*, based on the comparison of that text with thirty-two Spanish Golden Age entremeses, by ten different authors, among which are the eight unquestionable entremeses from *Don Quixote's* author.

Keywords
Cervantes; *Entremés de los romances*; Authorship attribution; Stylometry; Golden Age Theatre.

1. Investigadora posdoctoral Margarita Salas, gracias a una ayuda financiada por el Ministerio de Universidades a través de la Unión Europea (NextGenerationEU).

ISSN 2462-6813 (papel), ISSN 1988-1088 (en línea)
https://doi.org/10.5565/rev/studiaaurea.541

Estado de la cuestión

El *Entremés famoso de los romances*, publicado por primera vez en la *Parte tercera de las comedias de Lope de Vega y otros autores* (Valencia, 1611), recoge la historia de Bartolo, un ladrador recién casado que, obsesionado con las historias que lee en el Romancero, decide "dejar su esposa y tierra" para enrolarse en la guerra contra los ingleses, llevándose a su vecino Bandurrio como sirviente y escudero. Tras inmiscuirse en la discusión de pareja de dos pastores, de la que sale malparado, es encontrado por sus familiares y llevado de vuelta a casa, con la esperanza de que, tras una siesta, pueda recuperar la cordura. La comicidad de la pieza reside en el habla arromanzada de los personajes, que van construyendo sus intervenciones sobre la base de más de una treintena de romances populares, fácilmente reconocibles por los espectadores de la época: *Ensílleme el asno rucio*, *La más bella niña*, *Hermano Perico*, *Cabizbajo y pensativo*, *Mira Tarfe que a Daraja*, etc.

Aunque este entremés es uno de los más conocidos del teatro aurisecular español, no lo es tanto por su valía literaria como por su cercanía temático-argumental con la primera salida de don Quijote (*Quijote*, I, capítulos 1-6), lo que ha generado que, desde hace ciento cincuenta años, se debata en torno a su fecha de composición (anterior o posterior a la obra magna cervantina)[2] y su autoría; entre los candidatos propuestos, encontramos a Luis de Góngora, Juan de Salinas, Gabriel Lobo Lasso de la Vega, Liñán de Riaza, Lope de Vega o el propio Miguel de Cervantes.[3]

El primero en atribuir este entremés a Cervantes fue Adolfo de Castro, quien consideraba que el autor de *Quijote* "hizo lo que los grandes pintores: trazó un borrón o un ligero dibujo de un gran cuadro, primitivo pensamiento

2. En contra de la tendencia crítica mayoritaria, creemos que el entremés es posterior. Nuestros argumentos se basan en: (i) la presencia del escudero (frente a la ausencia de Sancho Panza en los primeros capítulos de la obra magna cervantina); (ii) la falta de coherencia en la caracterización de los personajes cuerdos, que utilizan un habla arromanzada incluso cuando el destemplado protagonista no está presente en la escena (algo que no le ocurre a don Quijote); (iii) el empleo de la forma versificada, bastante inusual a principios del XVII (de hecho, aunque el primer entremés en verso del que tenemos noticia —*Melisendra*— se publicó en 1605, el mismo año en que vio la luz la primera parte del *Quijote*, el uso de la forma prosificada siguió siendo bastante habitual hasta 1620, momento en que empezó a ser desbancada por la forma versificada, sobre todo de la mano de Quiñones de Benavente); (iv) el empleo de la fórmula sintética *vuesastedes*, frente al uso analítico anterior *vuesa/s merced/es*, que es el que encontramos en el *Quijote*.

3. A lo largo de todos estos años, las teorías en torno a su paternidad han sido muy diversas: Castro (1874), Menéndez Pelayo (1941), Northup (1922) y Rodríguez López-Vázquez (2019), como explicamos más adelante, creen que el autor es Cervantes; Rey Hazas (2006) y Blasco (2013) apuestan por Gabriel Lobo Lasso de la Vega; Millé y Giménez (1930) y López Navío (1959-1960) piensan en algún enemigo de Lope de Vega, como Góngora, Salinas o el mismo Cervantes; y Pérez López (2009), por el contrario, señala a Liñán de Riaza, gran amigo del Fénix, que pudo escribirla, tal vez, en connivencia con él.

que luego desenvolvió en un libro admirable" (1874: 140); una atribución tan defendida por unos (Menéndez Pelayo 1941, Northup 1922 y Rodríguez López-Vázquez 2019) como rechazada por otros (Menéndez Pidal 1964, Márquez Villanueva 1983, Campo 1948, Carreira 1998, Rey Hazas 2006, Pérez López 2009, Ruiz Urbón 2010 y Blasco 2013).

En un trabajo anterior, publicado en el año 2010, defendimos que Cervantes no era su autor, basándonos para ello en una serie de pruebas (temático-argumentales, estructurales, métricas, léxicas, gramaticales, etc.) que alejaban el idiolecto de esta pieza del cervantino (Ruiz Urbón 2010); y aunque Alfredo Baras Escolá considera que "ya nadie es partidario de la autoría de Cervantes" (2012: 200), lo cierto es que recientemente Alfredo Rodríguez López-Vázquez ha vuelto a poner el foco sobre él (2019). Tras analizar en profundidad dos pasajes del entremés —uno en quintillas (30 versos) y otro en redondillas (20 versos) y, en consecuencia, no contaminados por la inserción de romances ajenos— asegura que,

> la obra debe ser considerada como cervantina, en tanto no se proponga a un autor que en el período 1590-1600 use secuencias de quintillas del tipo *aabba* y use también la forma rústica ¡Dolo (a huego)!, manteniendo además un uso similar al de Cervantes para esos otros trece índices autoriales (238).

Vayamos por partes. Rodríguez López-Vázquez argumenta, en primer lugar, que la utilización de una secuencia de seis quintillas seguidas de tipo 5 (*aabba*) es una rareza, tomando como referencia el estudio de Morley y Bruerton sobre la cronología de las comedias de Lope de Vega, en donde este tipo de quintilla siempre aparece en combinación con una de tipo 1 (*ababa*) o de tipo 2 (*abbab*); advierte, sin embargo, que este uso tan particular sí se da en Cervantes, en su comedia *La casa de los celos*.[4] Aunque es cierta la cercanía combinatoria entre las quintillas que figuran en la apertura del entremés y este pasaje de la comedia cervantina, no nos parece acertado que se estipule la rareza de esta combinación basándose únicamente en su no aparición en las

4. Reproducimos, a continuación, el fragmento de *La casa de los celos* al que se refiere Rodríguez López-Vázquez (19):

[Espíritu]:	A tu primo entre esa yerba	a
	pondrás, que a mi se reserva	a
	y a mi fuente su salud,	b
	que hasta agora su virtud	b
	en ella el cielo conserva.	a
Malgesí:	Volveos por do vinistes	c
	figuras feas y tristes,	c
	que mi primo quedará	d
	adonde esperar podrá	d
	el remedio que no distes.	c

comedias de Lope de Vega. Consideramos, además, que un estudio riguroso de autoría debería abarcar el análisis global del metro, la estrofa y la rima del entremés (Ruiz Urbón 2010: 196-208), y no sólo el de unos cuantos versos seleccionados aleatoriamente.

Su segundo argumento en favor de la autoría cervantina de esta pieza es la utilización de la forma rústica *dolo* como pregunta y no como exclamación, que, según explica, era su uso habitual en el teatro popular breve del siglo XVI, a partir de la expresión "Doylo al huego > Doylo al fuego (del infierno)"; sin embargo, y tal y como reconoce más tarde, "ni Cervantes, ni Góngora, ni Gabriel Lobo Lasso de la Vega usan esta forma arcaizante" (235), por lo que el razonamiento atributivo no parece muy acertado. Además, tal y como anota Rey Hazas en su edición de este entremés, podría tratarse efectivamente de "*¡dolo!*, en el sentido de '¡malo, mala cosa!'", pero también de "*¿dólo?* [...], es decir, '¿por dónde [viene]?', que es lo más probable" (2006: 88).

En tercer y último lugar, Rodríguez López-Vázquez arguye, basándose en los datos ofrecidos por el *Corpus Diacrónico del Español* (*CORDE*), que trece de las quince unidades léxicas detectadas en esos dos pasajes del entremés están en Cervantes y sólo una vez en cada uno de los otros autores rastreados (Quevedo, Góngora, Vélez de Guevara, Lasso de la Vega y Lope de Vega) (234-237). No sabemos qué criterios ha seguido para seleccionar esas quince unidades, cuyos resultados, en todo caso, deben interpretarse dentro de la realidad que nos ofrece *CORDE* como repositorio limitado de textos. Tampoco entendemos los parámetros seleccionados, ya que unas veces decide restringir su búsqueda a un intervalo de tan sólo trece años (1600-1613) —intervalo en el que, por ejemplo, no se recoge ninguna obra dramática de Lasso de la Vega— y otras, sin embargo, no. Pero es que, además, y tal vez esto sea lo más relevante, no siempre da cuenta de los resultados obtenidos al margen de esos autores (que es lo que realmente permitiría establecer si se trata de un uso particular de Cervantes o no).

En definitiva: los argumentos atributivos utilizados por Rodríguez López-Vázquez nos parecen insuficientes para ahijar una obra a un determinado autor; argumentos que están basados, por otra parte, en el análisis arbitrario de algunos fragmentos sueltos del entremés (50 versos), ya que, en su opinión, "someter el texto íntegro [...] a un análisis de estilometría cuantitativa implica darle la misma importancia a los textos ajenos al autor que a los textos escritos en quintilla y redondilla, que sí son propios del autor de la obra" (238). Es cierto que la presencia de versos romanceriles complica enormemente el estudio atributivo de esta pieza; pero, si se identifican y eliminan todos esos versos, podemos aplicar diferentes técnicas estilométricas que permitan contrastar los usos lingüísticos de su autor con los utilizados por otros diez entremesistas de la época, incluido Cervantes, a fin de determinar si éste puede ser o no su autor y con qué grado de probabilidad.

Descripción del corpus de análisis

El corpus de análisis lo integran un total de treinta y tres entremeses del Siglo de Oro español (Figura 1):[5]

a) Texto dubitado: *Entremés de los romances*. Aunque el texto tiene 2243 palabras, nosotros trabajaremos con una versión reducida, de 1164 palabras, obtenida tras la eliminación de los versos que son reproducción literal o cuasi literal de conocidos romances de la época.[6]

b) Textos indubitados: Treinta y dos entremeses de diez autores distintos (Francisco de Ávila, Gaspar de Barrionuevo y Carrión, Luis de Belmonte Bermúdez, Alonso de Castillo Solórzano, Miguel de Cervantes, Antonio Hurtado de Mendoza, Francisco de Quevedo, Luis Quiñones de Benavente, Alonso Jerónimo de Salas Barbadillo y Luis Vélez de Guevara). Hemos contemplado entremeses en prosa (EP) y en verso (EV), si bien para ciertos análisis utilizaremos únicamente aquellos escritos en verso, ya que éste es el vehículo de expresión del entremés dubitado.

5. Los entremeses han sido tomados de la Biblioteca Virtual Miguel de Cervantes, de la *Flor de entremeses y sainetes de diferentes autores* (Madrid, 1657) (Menéndez Pelayo, ed., 1903), de la *Colección de entremeses* de Emilio Cotarelo y Mori (1911), del *Itinerario del entremés* de Eugenio Asensio (1971) y de Arellano y García Valdés (1997). Para facilitar el tratamiento automatizado de los textos y posibilitar su comparación, los hemos convertido a UTF-8 y hemos modernizado su ortografía, pero manteniendo las variantes gráficas que implican una variante fonética (como *agora* o *escura*), las contracciones (*desto*, *deso*, *dello*, *daquel*, *aqueso*, *esotro*...), las desinencias verbales enclíticas (del tipo *cogióla* o *teníanme*) y las formas verbales arcaicas (*gustallos*, *enternecella*, *escribille*...). Hemos optado por eliminar los títulos, la lista de figuras y el nombre del personaje que antecede a cada intervención; sí hemos mantenido, en cambio, las acotaciones, que en ciertos autores —como Cervantes— resultan idiosincrásicas.

6. Estos son los romances utilizados por el anónimo entremesista, por orden de aparición: (1) *Ensíllenme el asno rucio* (vv. 31-45, 63-70, 71-74, 76-78; (2) *La más bella niña* (vv. 93-94, 97-98, 101-102, 105-106, 107-110, 126-127); (3) *Hermano Perico* (vv. 132-135, 136-139, 140-193); (4) *Cabizbajo y pensativo* (vv. 204-205, 206-221, 226-229); (5) *Mira Tarfe que a Daraja* (vv. 238-251, 258-261, 401); (6) *Mal hubiese el caballero* (vv. 287-288); (7) *De Mantua salió el marqués* (vv. 292-293, 306-313, 314-317, 318-321, 324-325, 324-339, 344-345, 349-355, 358-361); (8) *¿Dónde estás, señora mía?* (vv. 294-301); (9) *Dime, Bencerraje amigo* (vv. 384-389); (10) *Galiana está en Toledo* (v. 393); (11) *Si tienes el corazón* (vv. 396-397); (12) *Por una nueva ocasión* (v. 400); (13) *Rendido está Reduán* (v. 402); (14) *De las montañas de Jaca* (v. 403); (15) *Elicio, un pobre pastor* (v. 404); (16) *En una pobre cabaña* (v. 405); (17) *Con semblante desdeñoso* (v. 406); (18) *De pechos sobre una vara* (v. 407); (19) *Bravonel de Zaragoza* (v. 408); (20) *Discurriendo en la batalla* (v. 409); (21) *Por muchas partes herido* (v. 410); (22) *Rotas las sangrientas armas* (v. 411); (23) *Sale la estrella de Venus* (v. 412); (24) *Rompiendo la mar de España* (v. 413); (25) *Después que con alboroto* (v. 414); (26) *Entró la malmaridada* (v. 415); (27) *En un caballo ruano* (v. 416); (28) *¡Afuera, afuera! ¡Aparta, aparta!* (v. 417); (29) *Todos dicen que soy muerto* (v. 424); (30) *Dígasme tú la serrana* (v. 425); (31) *Azarque, indignado y fiero* (vv. 426-427); (32) *Azarque vive en Ocaña* (v. 429); (33) *Ardiéndose estaba Troya* (v. 470). Hemos eliminado los versos que son reproducción literal y también aquellos en los que la copia es más que evidente, aunque se alteren sólo algunas palabras, como ocurre, por ejemplo, con la afirmación de Bartolo "veintidós palos me han dado, / que el menor era mortal", que en boca de Valdovinos era "veinte y dos heridas tengo / que cada una es mortal".

Autor	Título	Etiqueta	Extensión
Anónimo	*Los romances*	Dub_EV_romances	2243 palabras
		Dub_EV_romances_R	1164 palabras
Ávila	*Los invencibles hechos de don Quijote de la Mancha*	Av_EV_quijote	2 584 palabras
	El mortero y chistes del sacristán	Av_EV_mortero	2268 palabras
Barrionuevo	*El triunfo de los coches*	Barr_EP_triunfo	4701 palabras
Belmonte	*Sierra Morena de las mujeres*	Belm_EV_sierra	1012 palabras
	Una rana hace ciento	Belm_EV_rana	871 palabras
Castillo Solórzano	*El barbador*	Cast_EV_barbador	1294 palabras
	El comisario de figuras	Cast_EV_comisario	1422 palabras
Cervantes	*El juez de los divorcios*	Cerv_EP_juez	3004 palabras
	El rufián viudo	Cerv_EV_rufian	2748 palabras
	La elección de los alcaldes de Daganzo	Cerv_EV_eleccion	2386 palabras
	La guarda cuidadosa	Cerv_EP_guarda	3684 palabras
	El vizcaíno fingido	Cerv_EP_vizcaino	4418 palabras
	El retablo de las maravillas	Cerv_EP_retablo	3300 palabras
	La cueva de Salamanca	Cerv_EP_cueva	3568 palabras
	El viejo celoso	Cerv_EP_viejo	3790 palabras
Hurtado de Mendoza	*Getafe*	Hurt_EV_getafe	1602 palabras
	El examinador Miser Palomo	Hurt_EV_miser1	2353 palabras
	Miser Palomo, el médico del espíritu	Hurt_EV_miser2	2108 palabras
Quevedo	*El marido Pantasma*	Quev_EV_marido	1748 palabras
	El niño y Peralvillo de Madrid	Quev_EV_niño	1390 palabras
	La vieja Mutañones	Quev_EP_vieja	2102 palabras

Autor	Título	Etiqueta	Extensión
Quiñones de Benavente	*El ventero*	Quin_EV_ventero	1437 palabras
	Los sacristanes	Quin_EV_sacristanes	1824 palabras
	El barbero	Quin_EV_barbero	986 palabras
	El borracho	Quin_EV_borracho	1859 palabras
Salas Barbadillo	*El buscaoficios*	Salas_EP_buscaoficios	3941 palabras
	Los mirones de la Corte	Salas_EP_mirones	2254 palabras
	El caprichoso en su gusto	Salas_EV_caprichoso	3510 palabras
	El comisario contra los malos gustos	Salas_EV_comisario	2729 palabras
Vélez de Guevara	*La burla más sazonada*	Vel_EV_burla	1144 palabras
	Los atarantados	Vel_EV_atarantados	1440 palabras
	Antonia y Perales	Vel_EV_antonia	1094 palabras

Figura1.
Corpus de análisis (autor, título, etiqueta y extensión)

Análisis estilométricos de atribución de autoría

Para dar respuesta al objetivo de nuestro trabajo, en los siguientes subapartados llevaremos a cabo siete pruebas de carácter estilométrico: (i) análisis de la distribución de la frecuencia de las palabras más empleadas en los textos; (ii) análisis de componentes principales aplicado a la distancia delta; (iii) clasificación supervisada por aprendizaje automático; (iv) verificación de autoría con General Imposters; (v) estudio de las palabras de función (orden de frecuencia, frecuencia relativa y PCA); (vi) análisis de la frecuencia de las distintas clases de palabras; y (vii) uso y frecuencia relativa de las conjunciones coordinadas y subordinadas adverbiales. Trabajaremos principalmente con lenguaje de programación R —unas veces a través de librerías y otras con *scripts* personalizados—, aunque para ciertos análisis utilizaremos los datos y gráficos ofrecidos por *Voyant Tools* (https://voyant-tools.org/).[7]

7. De un tiempo a esta parte, los análisis estilométricos —y, en especial, la distancia delta— han servido para ofrecer importantes teorías sobre la autoría de varias piezas del Siglo de Oro, como *El Lazarillo de Tormes* (Rosa y Suárez 2016), el *Quijote* apócrifo (Blasco 2016 y Rißler-Pipka 2016), *La conquista de Jerusalén* (Cerezo Soler y Calvo Tello 2019) o *La francesa Laura* (Cuéllar y Vega García Luengos 2023).

Análisis de la distribución de la frecuencia de las palabras más empleadas en los textos (Distancia delta, 1-grama de palabra)

La distancia delta es uno de los procedimientos estilométricos más usados hoy en día para cuantificar el grado de similitud y disimilitud entre un determinado conjunto de textos. Esta medida, propuesta por John Burrows, se sustenta en la idea de que la variación en las frecuencias de las palabras más empleadas en los textos permite establecer conclusiones sobre su autoría: "the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text" (2002: 271).[8]

Para calcular las redes léxicas resultantes de aplicar la distancia delta al texto expurgado del *Entremés de los romances* (Dub_EV_romances_R), utilizamos la función Stylo del paquete Stylo (Eder, Rybicki y Kestemont 2016) y seleccionamos el idioma español, el algoritmo delta de Eder, diseñado específicamente para las lenguas flexivas (Eder's Delta), y 1-grama de palabra para las 1000 palabras más frecuentes (MFW), sin restricciones (0% culled); parámetros que han demostrado tener una fiabilidad superior al 90% en el caso particular de los entremeses del Siglo de Oro español (Ruiz Urbón 2023: 77-83).

Tal y como muestran las tres siguientes figuras, el comportamiento del texto atribuido a Cervantes es muy distinto al de sus entremeses de autoría cierta.[9] En el primer caso, en el que hemos contemplado la totalidad de los textos indubitados, el *Entremés de los romances* se enmarca en una rama en la que están todos los entremeses de Quevedo, Quiñones de Benavente, Vélez de Guevara y Belmonte, pero sin asociarse directamente con ninguno de ellos; en la rama opuesta, se encuentran los de Cervantes, Hurtado de Mendoza, Ávila, Barrionuevo, Salas Barbadillo y Castillo de Solórzano (Figura 2). En el segundo caso, en el que únicamente hemos tenido en cuenta los textos indubitados en verso, el *Entremés de los romances* pasa a ubicarse en la rama contraria, vinculándose con los dos entremeses de Ávila, pero no de manera directa (Figura 3). Esta inestabilidad en el resultado del texto dubitado, unida al hecho de que su agrupación sea la que más a la izquierda se produce de todo el corpus (es decir, en un número más alto de MFW), nos lleva a pensar

8. Para una explicación detallada del algoritmo de Burrows (Burrows's Delta) y sus actualizaciones posteriores (Cosine Delta, Eder's Delta, Hoover's Delta, Linear Delta, Bray-Curtis, Euclidean, Manhattan, Rotated Delta, Quadratic Delta, etc.), véase Evert, Proisl, Jannidis, Reger, Pielström, Schöch y Vitt (2017).

9. Aunque Stylo proporciona todos los datos resultantes de cada análisis (lista de palabras, tabla de frecuencias, valores delta para cada pareja de textos, etc.), únicamente ofrecemos aquí los resultados en forma de gráfico, ya sea como análisis de conglomerados (*cluster analysis*) o como árboles de consenso (*bootstrap consensus tree*). El resto de elementos pueden consultarse en https://github.com/Cruizurbon/Entremes_de_los_romances.

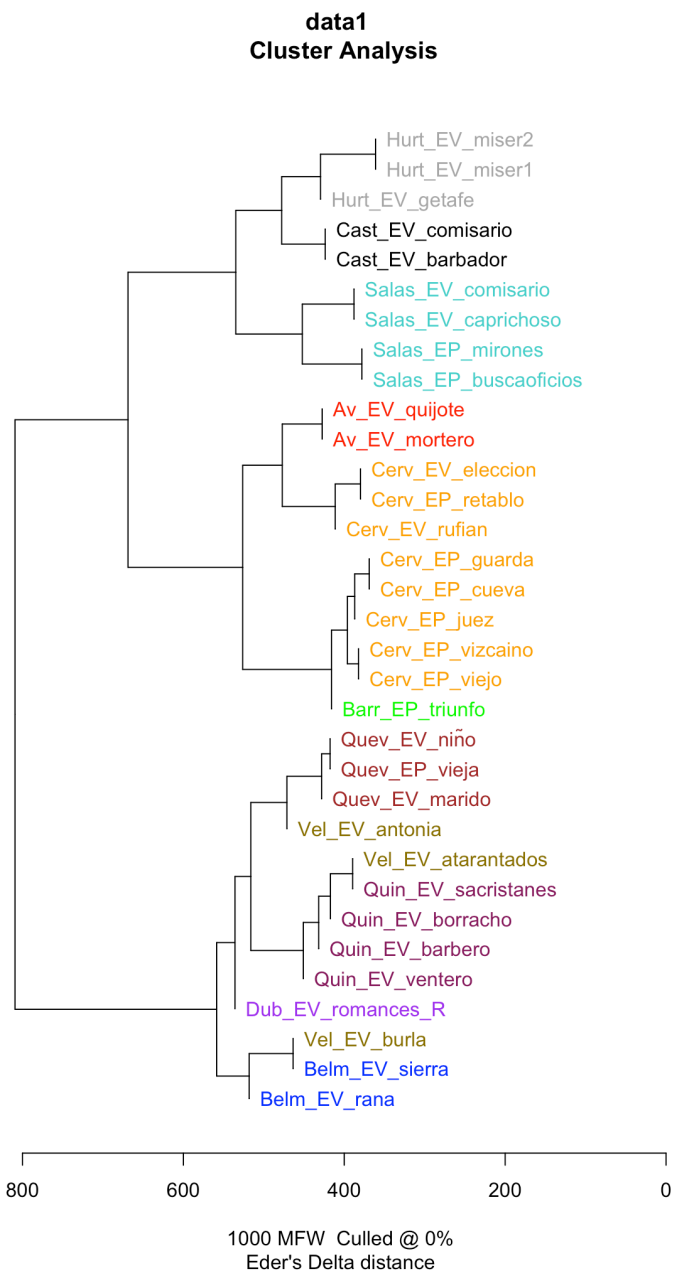

Figura 2.
Distancia delta de 1-grama de palabra entre el *Entremés de los romances* y el corpus de entremeses indubitados (*RStudio*. Stylo. Cluster Analysis. Eder's Delta. MFW 1000. Culled 0%)

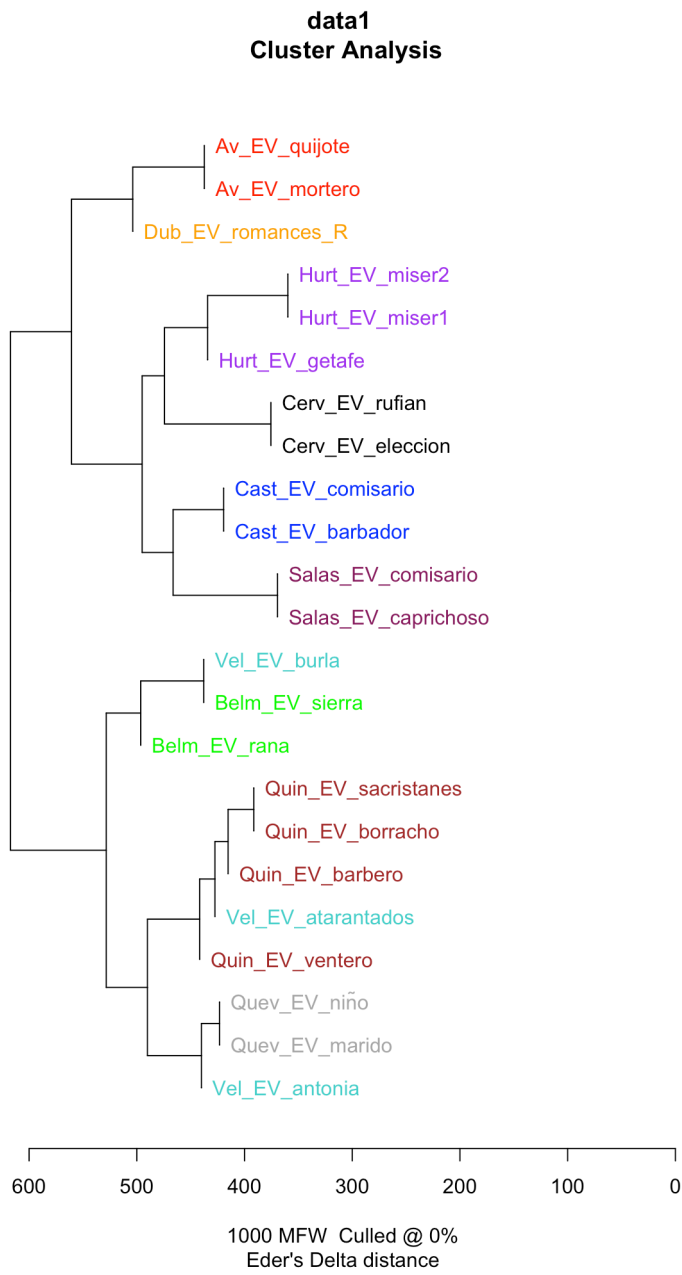

Figura 3.
Distancia delta de 1-grama de palabra entre el *Entremés de los romances* y el corpus de entremeses indubitados en verso (*RStudio*. Stylo. Cluster Analysis. Eder's Delta. MFW 1000. Culled 0%)

que el verdadero autor del texto anónimo no está entre ninguno de los que conforman este corpus.[10] En el tercer caso, en el que hemos reducido el corpus de contraste a los ocho entremeses indubitados de Cervantes, la falta de sintonía vuelve a ser evidente (Figura 4). Los ocho entremeses de Cervantes se agrupan en torno a dos ramas, mientras que *El entremés de los romances* aparece ubicado en una tercera rama, alejado de todos ellos. Esta diferenciación

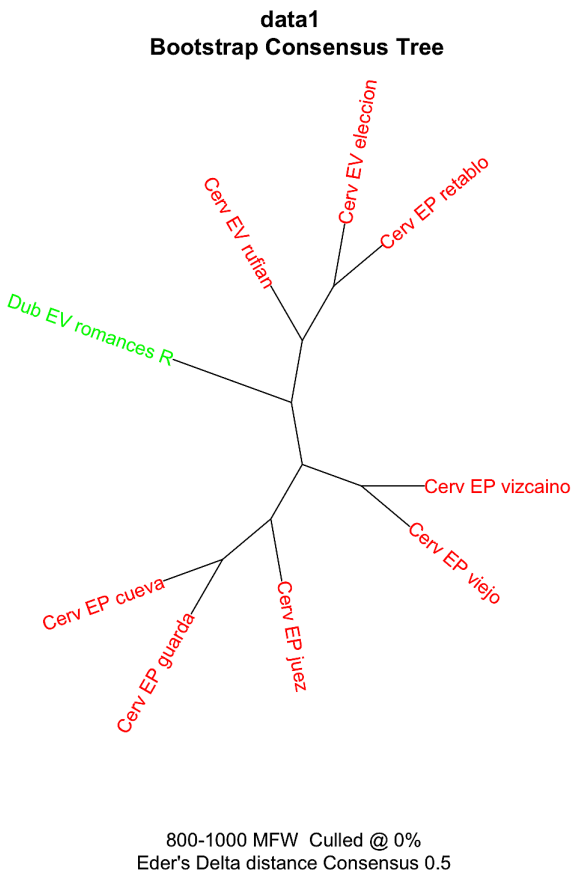

Figura 4.
Distancia delta de 1-grama de palabra entre el *Entremés de los romances* y los ochos entremeses indubitados de Cervantes (*RStudio*. Stylo. Bootstrap Consensus Tree. 0.5 consensus strength. Eder's Delta. MFW 800-1000)

10. Los únicos entremeses indubitados de autor conocido que no se asocian "correctamente" entre sí son los tres de Vélez de Guevara, ya sea por la baja calidad de las obras heredadas (contaminación autorial o textual); por la baja señal autorial de su escritura, con una clara tendencia a las prácticas pseudoplagiarias; o porque, en realidad, no son suyos (error de atribución).

viene reforzada, además, por la visualización en forma de árbol de consenso, que ofrece la agrupación más estable (entre las 800 y las 1000 MFW).[11]

Análisis de componentes principales aplicado a la distancia delta

El análisis de componentes principales (PCA) es una técnica estadística de síntesis de la información que consiste en describir un conjunto de datos en términos de nuevas variables (o componentes) no correlacionadas, con el objetivo de reducir la dimensionalidad de dicho conjunto (Jolliffe 2002: 2). En la siguiente figura mostramos el resultado de aplicar esta técnica a la distancia delta de 1-grama de palabra sobre las 1000 palabras más frecuentes (Figura 5):

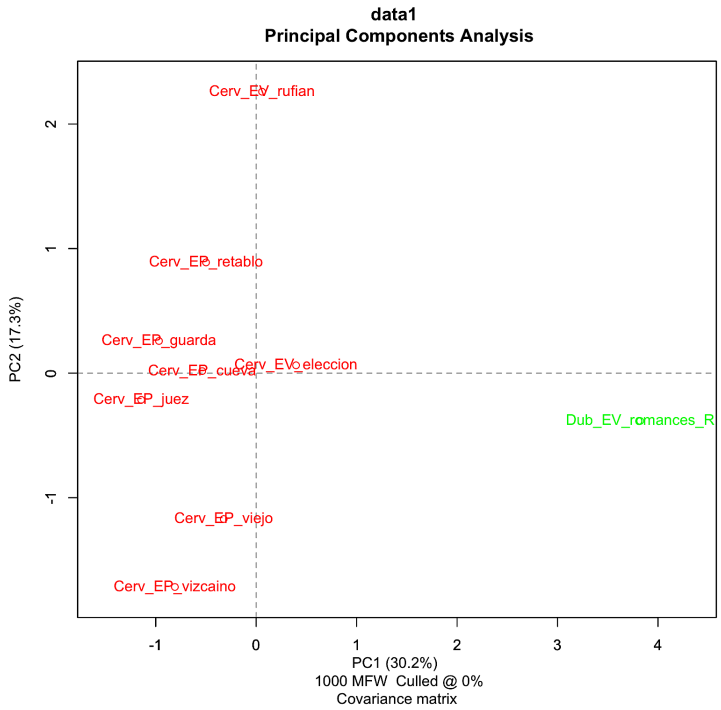

Figura 5.
PCA de la distancia delta aplicada a los ocho entremeses de Cervantes y al*Entremés de los romances* (*RStudio*. PCA. Eder's Delta. 1000 MFW)

11. El árbol de consenso, introducido por Eder para evitar la arbitrariedad del investigador en la selección de las palabras más frecuentes (*cherry picking*), ofrece la agrupación más sólida, es decir, aquélla que permanece a lo largo de un porcentaje mínimo de iteraciones (*consensus value*) (2013).

Los dos primeros componentes principales (PC1 y PC2) cubren respectivamente el 30,2 y el 17,3% de la varianza total. Llama la atención que el primer componente (eje horizontal) muestra una clara diferenciación entre entremeses en verso (en la mitad derecha) y en prosa (en la mitad izquierda), que se corresponde, muy probablemente, con diferentes momentos de escritura. Centrándonos en el asunto de la autoría, el *Entremés de los romances* se sitúa en el cuarto cuadrante, solo y muy alejado de los ocho indubitados cervantinos.

Esta disonancia refuerza la hipótesis de que Cervantes no es su autor. Disonancia que también aparece cuando repetimos el análisis con otros textos no cervantinos, como, por ejemplo, el *Entremés del marido pantasma*, de Quevedo (Figura 6), o el *Entremés del triunfo de los coches,* de Barrionuevo (Figura 7).

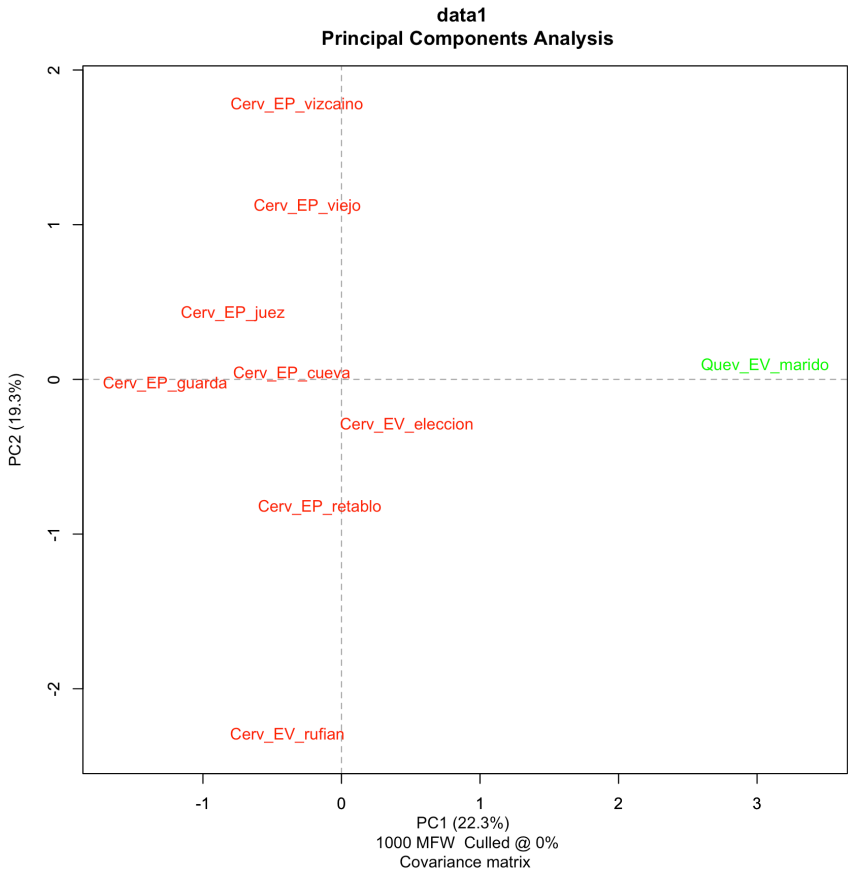

Figura 6.
PCA de la distancia delta aplicada a los ocho entremeses de Cervantes y al *Entremés del marido pantasma* (*RStudio*. PCA. Eder's Delta. 1000 MFW)

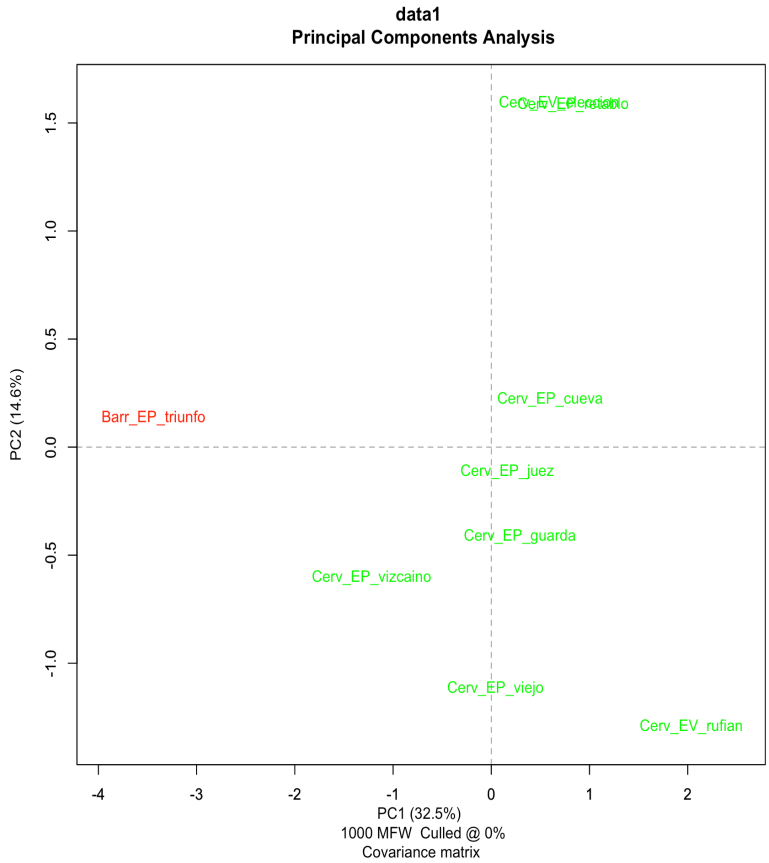

Figura 7.
PCA de la distancia delta aplicada a los ocho entremeses de Cervantes y al *Entremés del triunfo de los coches* (*RStudio*. PCA. Eder's Delta. 1000 MFW)

Como puede apreciarse en las figuras 6 y 7 —y al margen de las diferencias esperables entre los ocho entremeses cervantinos—, el *Entremés del marido pantasma* y el *Entremés del triunfo de los coches* aparecen en cuadrantes distintos, tal y como ocurría en el caso del texto dubitado. El análisis de componentes principales aplicado a la distancia delta, por tanto, refrenda la hipótesis de la autoría no cervantina del *Entremés de los romances*.

Clasificación supervisada por aprendizaje automático (*Classify*)

Otra de las funcionalidades del paquete Stylo es Classify, que está basada en una serie de métodos de aprendizaje automático especialmente concebidos para su

aplicación a la estilística computacional: Delta, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes y Nearest Shrunken Centroids (NSC) (Jockers y Witten 2010 y Eder, Rybicki y Kestemont 2016). El procedimiento es sencillo: Classify extrae una tabla de frecuencias de los textos que incluimos en un corpus de entrenamiento (*primary set*), aprende los parámetros caracterizadores de cada texto, compara esas frecuencias con las de cada uno de los textos que conforman el corpus de prueba (*secondary set*) y propone asociaciones.

En la siguiente tabla mostramos los resultados obtenidos de aplicar Classify al estudio de la distribución de la frecuencia léxica, tras seleccionar el algoritmo Eder's Delta y un intervalo de 100 a 1000 palabras más frecuentes (con un incremento de 100 en 100). El corpus de entrenamiento está conformado por textos indubitados de Cervantes (Cerv_EV_eleccion y Cerv_EP_retablo), Ávila (Av_EV_mortero), Salas Barbadillo (Salas_EP_mirones), Quiñones de Benavente (Quin_EV_borracho), Hurtado de Mendoza (Hurt_EV_getafe) y Quevedo (Quev_EV_marido); y el de prueba, por el texto dubitado (Dub_EV_romances_R) y otros tantos indubitados de los mismos autores (Cerv_EV_rufian, Cerv_EP_guarda, Av_EV_quijote,Salas_EV_caprichoso, Quin_EV_barbero, Hurt_EV_miser1 y Quev_EV_vieja):

	Resultados		
100 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Av_EV_mortero
	Dub_EV_romances_R	→	**Cerv_EV_eleccion**
	Cerv_EP_guarda	→	Av_EV_mortero
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Cerv_EP_retablo
	Cerv_EV_rufian	→	Quin_EV_borracho
	Quev_EP_vieja	→	Cerv_EV_eleccion
200 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Av_EV_mortero
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Av_EV_mortero
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Cerv_EP_retablo
	Cerv_EV_rufian	→	Cerv_EV_eleccion
	Quev_EP_vieja	→	Quev_EV_marido
300 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Av_EV_mortero
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Cerv_EP_retablo
	Cerv_EV_rufian	→	Cerv_EV_eleccion
	Quev_EP_vieja	→	Quev_EV_marido

	Resultados		
400 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Av_EV_mortero
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Cerv_EP_retablo
	Cerv_EV_rufian	→	Quin_EV_borracho
	Quev_EP_vieja	→	Quev_EV_marido
500 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Av_EV_mortero
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Cerv_EP_retablo
	Cerv_EV_rufian	→	Quin_EV_borracho
	Quev_EP_vieja	→	Quev_EV_marido
600 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Av_EV_mortero
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Av_EV_mortero
	Cerv_EV_rufian	→	Quin_EV_borracho
	Quev_EP_vieja	→	Quev_EV_marido
700 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Av_EV_mortero
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Av_EV_mortero
	Cerv_EV_rufian	→	Cerv_EV_eleccion
	Quev_EP_vieja	→	Quev_EV_marido
800 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Cerv_EV_eleccion
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Av_EV_mortero
	Cerv_EV_rufian	→	Cerv_EV_eleccion
	Quev_EP_vieja	→	Quev_EV_marido

	Resultados		
900 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Cerv_EV_eleccion
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Av_EV_mortero
	Cerv_EV_rufian	→	Cerv_EV_eleccion
	Quev_EP_vieja	→	Quev_EV_marido
1000 MFW	Quin_EV_barbero	→	Quin_EV_borracho
	Salas_EV_caprichoso	→	Salas_EP_mirones
	Dub_EV_romances_R	→	**Av_EV_mortero**
	Cerv_EP_guarda	→	Cerv_EV_eleccion
	Hurt_EV_miser1	→	Hurt_EV_getafe
	Av_EV_quijote	→	Av_EV_mortero
	Cerv_EV_rufian	→	Cerv_EV_eleccion
	Quev_EP_vieja	→	Quev_EV_marido

Figura 8.
Clasificación supervisada por aprendizaje automático
(*RStudio*. Classify. Eder's Delta. MFW 100-1000)

El *Entremés de los romances* (Dub_Ev_romances_R) se asocia con Cervantes en las 100 palabras más frecuentes (Cerv_EV_eleccion) y con Ávila desde las 200 hasta las 1000 (Av_EV_mortero). La asociación con Cervantes no es significativa, ya que con ese número de palabras (100 MFW) el éxito atributivo, calculado en base a las asociaciones correctas entre indubitados (2 de 7), se sitúa en el 28,5%. Classify, por tanto, tampoco identifica al autor de este texto anónimo con Cervantes.

Verificación de autoría con *General Imposters*

Stylo también nos permite realizar un análisis de verificación de autoría mediante el método General Imposters (GI) (Koppel y Winter 2014), cuyo objetivo:

> is not to assess whether two documents are simply similar in writing style, given a static feature vocabulary, but rather, it aims to assess whether two documents are significantly more similar to one another than other documents, across a variety of stochastically impaired feature spaces (Eder, 2012; Stamatatos, 2006), and compared to random selections of so-called distractor authors (Juola, 2015), also called 'imposters' (Stover, Koppel, Karsdorp y Daelemans 2016: 88).

Este método —que trabaja con una métrica de "segundo orden"— compara, en varias iteraciones, el texto dubitado con un conjunto de textos de varios autores distintos, a fin de determinar si alguno de ellos pudo escribir ese texto. A cada caso se le asigna una puntuación entre 0 y 1; en términos teóricos, toda puntuación superior a 0,5 sugeriría la verificación de autoría, pero lo cierto es que cualquier puntuación entre 0,39 y 0,63 debe ser considerada dudosa, pues indica que el clasificador tuvo problemas para tomar decisiones claras.

Tal y como vemos en la siguiente tabla, la comparación en varias iteraciones entre el *Entremés de los romances* y los distintos autores del corpus de entremeses indubitados (Ávila, Barrionuevo, Belmonte, Cervantes, Hurtado de Mendoza, Quevedo, Quiñones de Benavente, Salas Barbadillo y Vélez de Guevara) no revela vinculaciones significativas:

	Av	**Barr**	**Belm**	**Cast**	**Cerv**	**Hurt**	**Quev**	**Quin**	**Salas**	**Vel**
500 MFW	0.43	0.00	0.19	0.19	0.00	0.17	0.11	0.47	0.00	0.22
800 MFW	0.29	0.00	0.39	0.44	0.00	0.19	0.22	0.19	0.00	0.34
1000 MFW	0.21	0.01	0.74	0.00	0.00	0.03	0.10	0.04	0.00	0.32

Figura 9.
Aplicación de GI al *Entremés de los romances* (*RStudio*. Stylo.
GI. Ruzicka Delta distance. 500, 800 y 1000 MFW)

El método GI no identifica a Cervantes como autor del *Entremés de los romances* en ninguna de las frecuencias analizadas (500, 800 y 1000 MFW), estableciendo dicha posibilidad en 0 sobre 1. Tampoco parece hacerlo con ningún otro autor del corpus de análisis, ya que sólo en un caso aislado se supera el umbral del 0,63 (Belmonte, con 0,74 para las 1000 MFW).

Estudio de las palabras de función (orden de frecuencia, frecuencia relativa y PCA)

Las palabras de función (preposiciones, conjunciones, determinantes, pronombres, verbos auxiliares, etc.) son otro de los elementos clave para discriminar autorías, ya que escapan del control consciente del escritor y, además, su uso no viene condicionado por el tema del texto.[12] En este apartado vamos a analizar (i)

12. Los verbos copulativos están a medio camino entre las formas verbales con significado pleno, que indican una acción (*cantar*), un proceso (*aprender*) o un estado (*saber*), y las que solo tienen

su orden de frecuencia, (ii) su frecuencia relativa y (iii) la aplicación de ésta al análisis de componentes principales (PCA).

Extraemos, en primer lugar, las diez palabras de función más frecuentes en el *Entremés de los romances* —que, en orden de mayor a menor frecuencia, son *que*, *y*, *de*, *a*, *el*, *no*, *la*, *se*, *en* y *es*— y las comparamos con las diez que aparecen en el conjunto de entremeses indubitados, agrupados por autores, a fin de comprobar si los comportamientos se asemejan o no (Figura 10).[13] Para facilitar su interpretación, marcamos en verde las palabras que coinciden con el texto dubitado en la misma posición; en azul las que coinciden, pero en distinta posición; y en rojo las que no coinciden:

	Listado de las diez primeras palabras de función									
	1ª	**2ª**	**3ª**	**4ª**	**5ª**	**6ª**	**7ª**	**8ª**	**9ª**	**10ª**
Dub_R	que	y	de	a	el	no	la	se	en	es
Av	que	de	y	el	a	la	en	no	por	es
Barr	que	y	de	a	no	v.m.	el	la	en	se
Belm	de	que	y	no	el	en	a	es	la	lo
Cast	de	el	que	y	en	a	no	un	con	la
Cerv	que	de	y	el	la	a	no	en	me	por
Hurt	de	que	el	y	en	a	no	la	es	que
Quev	de	y	que	el	la	a	no	en	lo	es
Quiñ	que	de	y	la	el	a	en	no	con	es
Salas	que	de	y	el	a	en	la	no	con	es
Vel	que	de	y	el	a	la	es	no	es	si

Figura 10.
Listado de las 10 primeras palabras de función
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

La tabla pone de manifiesto el uso de unas mismas palabras de función entre todos los entremesistas (alta presencia de las palabras marcadas en color azul),

significado gramatical, como los verbos auxiliares. Hemos optado por incluir el verbo *ser* en este listado de palabras de función.

13. Para garantizar la correcta cuantificación, hemos tratado los textos para deshacer aquellas formas contractas (*al*, *del*, *deste*, *dello*, etc.) o enclíticas (como *cayóseme* o *doyle*).

pero, también, que no todos las utilizan en el mismo orden de frecuencia (baja frecuencia de las palabras marcadas en color verde). El *Entremés de los romances* comparte ocho de sus diez palabras de función más frecuentes con Cervantes, pero sólo una de ellas en la misma posición (*que*), lo que aleja la idea de que sea su autor.[14] El autor que más se aproxima en esta variable al texto dubitado es Barrionuevo (con ocho palabras coincidentes, cuatro de ellas en la misma posición).

Calculamos, en segundo lugar, la frecuencia relativa de las diez palabras de función más utilizadas en *Los romances* y la comparamos con la que tienen en el resto de autores de nuestro corpus (Figura 11).

	Frecuencia relativa de las palabras de función									
	que	**y**	**de**	**a**	**el**	**no**	**la**	**se**	**en**	**es**
Dub_R	0.0390	0.0390	0.0297	0.0212	0.0195	0.0195	0.0178	0.0169	0.0153	0.0127
Av	0.0565	0.0348	0.0469	0.0262	0.0322	0.0137	0.0176	0.0070	0.0154	0.0129
Barr	0.0549	0.0469	0.0416	0.0258	0.0218	0.0249	0.0180	0.0139	0.0163	0.0131
Belm	0.0436	0.0294	0.0488	0.0163	0.0220	0.0283	0.0157	0.0115	0.0199	0.0157
Cast	0.0360	0.0338	0.0518	0.0198	0.0360	0.0151	0.0125	0.0103	0.0209	0.0118
Cerv	0.0506	0.0452	0.0502	0.0239	0.0262	0.0198	0.0241	0.0103	0.0193	0.0093
Hurt	0.0411	0.0330	0.0513	0.0216	0.0333	0.0214	0.0160	0.0064	0.0217	0.0137
Quev	0.0422	0.0449	0.0505	0.0253	0.0320	0.0175	0.0260	0.0096	0.0166	0.0115
Quiñ	0.0418	0.0341	0.0341	0.0236	0.0281	0.0151	0.0286	0.0125	0.0156	0.0096
Salas	0.0456	0.0327	0.0445	0.0243	0.0316	0.0158	0.0209	0.0080	0.0213	0.0134
Vel	0.0452	0.0344	0.0393	0.0214	0.0236	0.0184	0.0209	0.0095	0.0160	0.0190

Figura 11.
Frecuencia relativa de las 10 palabras de función más frecuentes
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

Acompañamos la tabla de un gráfico de barras para cada una de esas diez palabras (Figuras 12-21):

14. Hemos recogido el orden que refleja el análisis realizado con *RStudio*, si bien es cierto que, tal y como se desprende de la tabla de frecuencias relativas, la primera posición la comparten *que* e *y* y la quinta *el* y *no.* Si invertimos el orden de estos elementos, las conclusiones son las mismas.

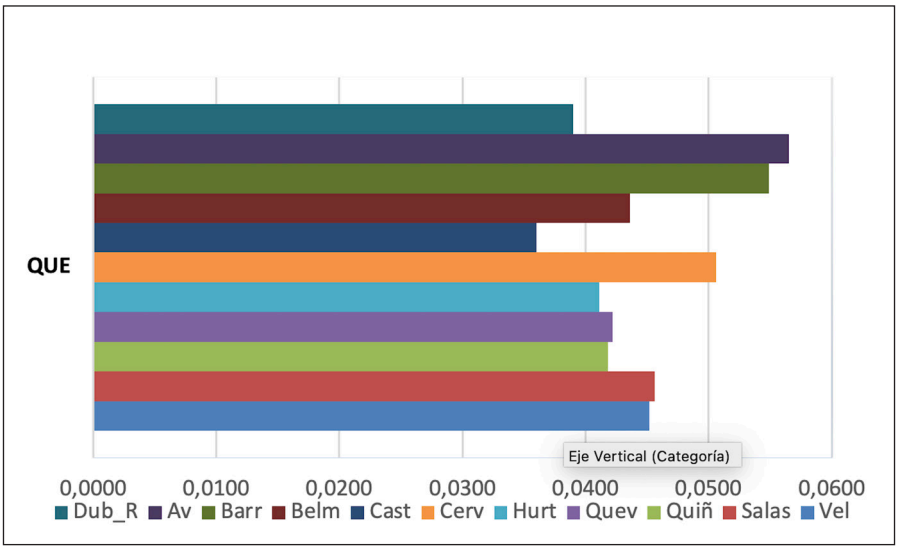

Figura 12.
Análisis comparado de la frecuencia relativa de *que*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

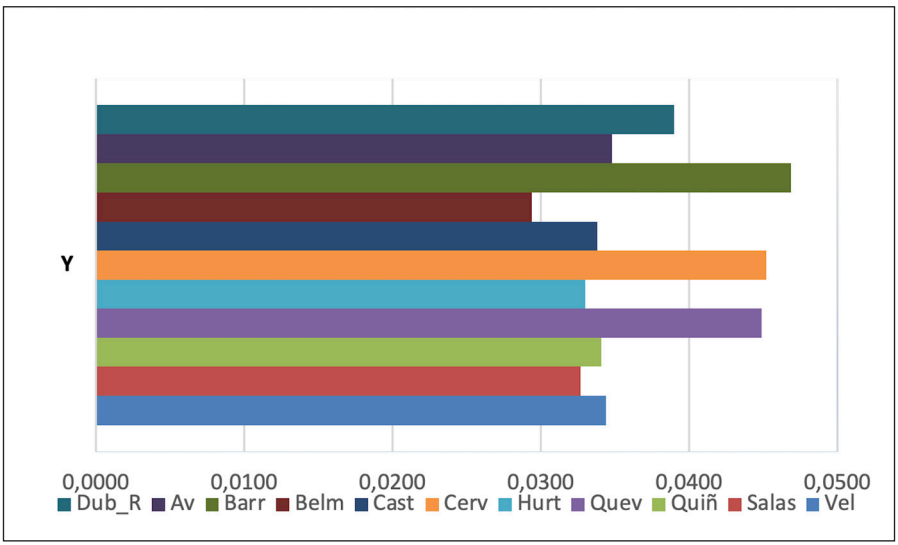

Figura 13.
Análisis comparado de la frecuencia relativa de *y*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

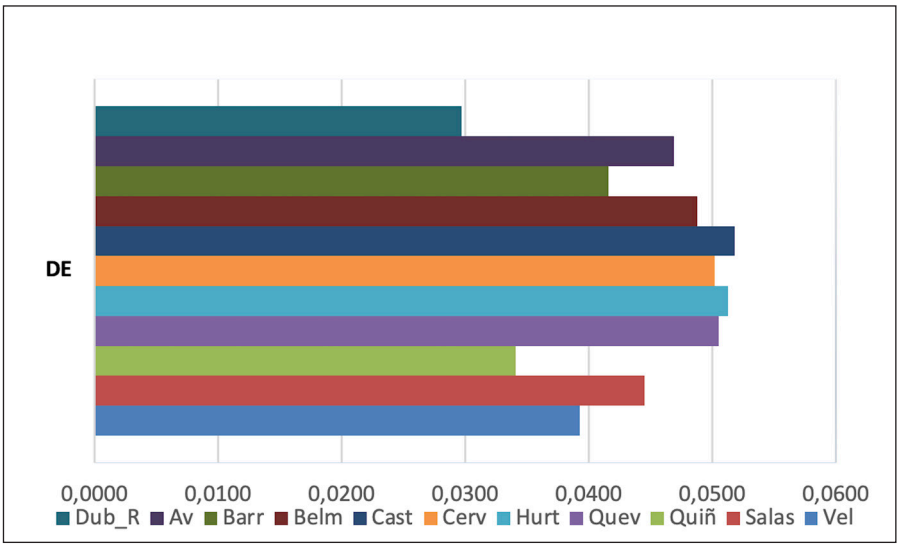

Figura 14.
Análisis comparado de la frecuencia relativa de *de*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

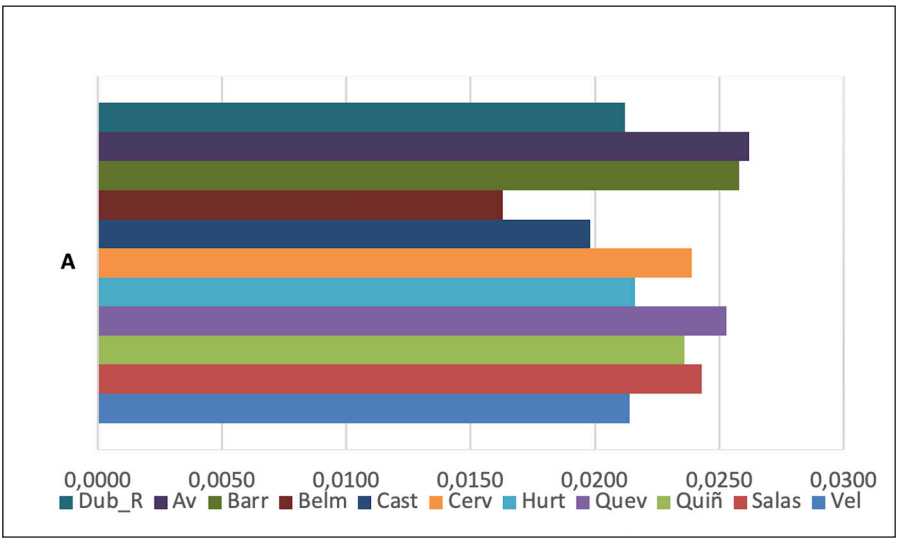

Figura 15.
Análisis comparado de la frecuencia relativa de *a*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

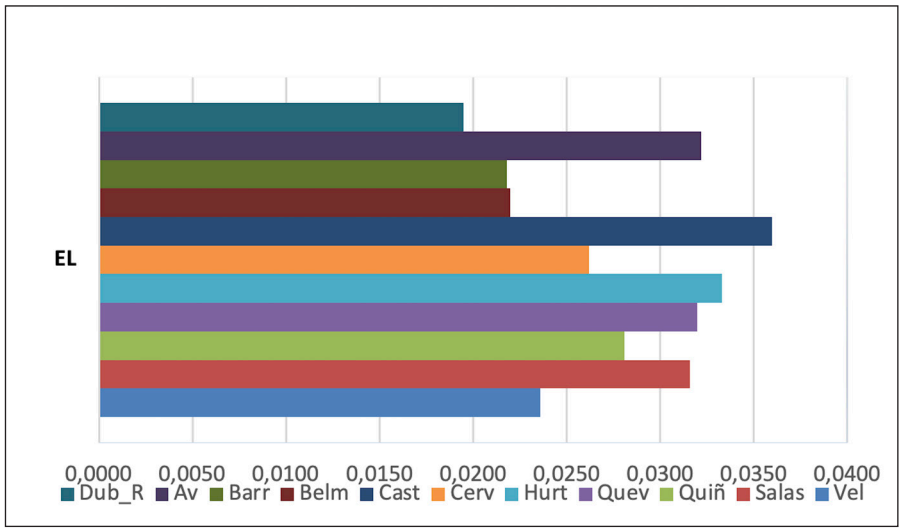

Figura 16.
Análisis comparado de la frecuencia relativa de *el*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

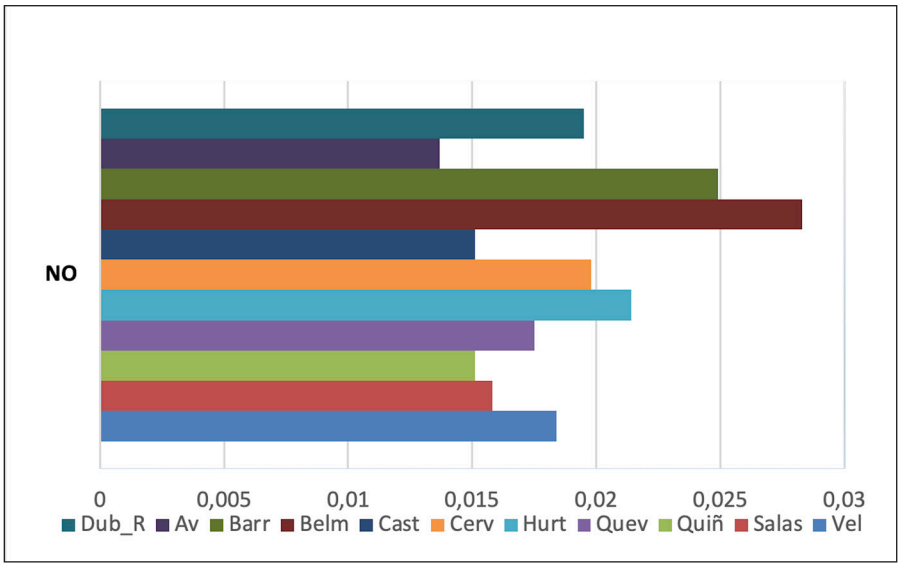

Figura 17.
Análisis comparado de la frecuencia relativa de *no*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

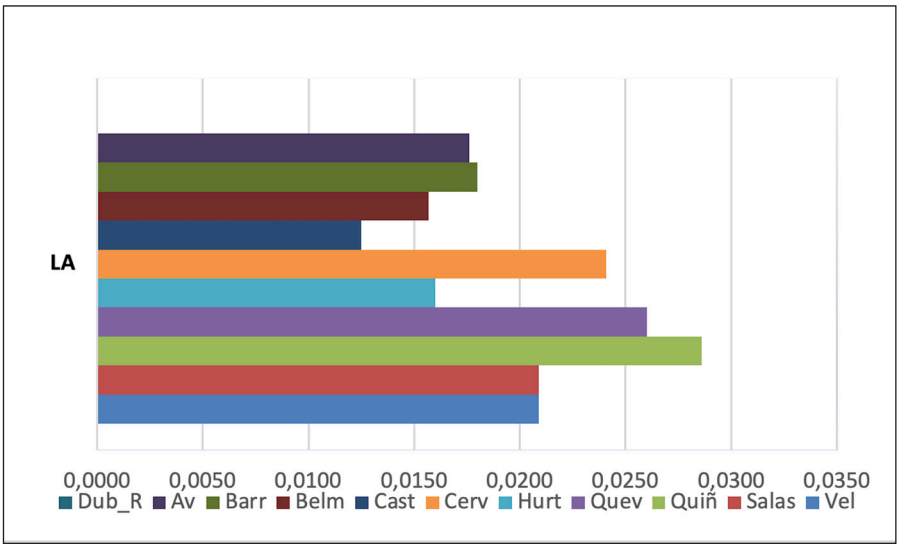

Figura 18.
Análisis comparado de la frecuencia relativa de *la*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

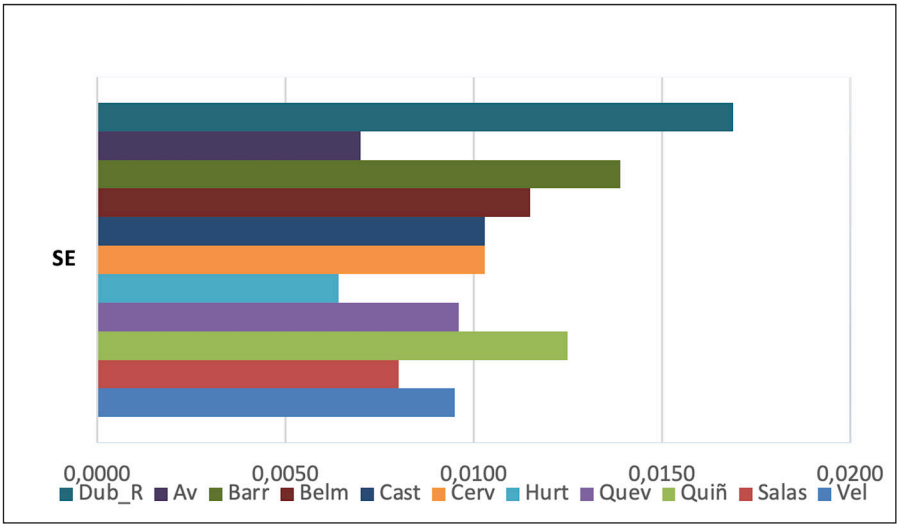

Figura 19.
Análisis comparado de la frecuencia relativa de *se*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

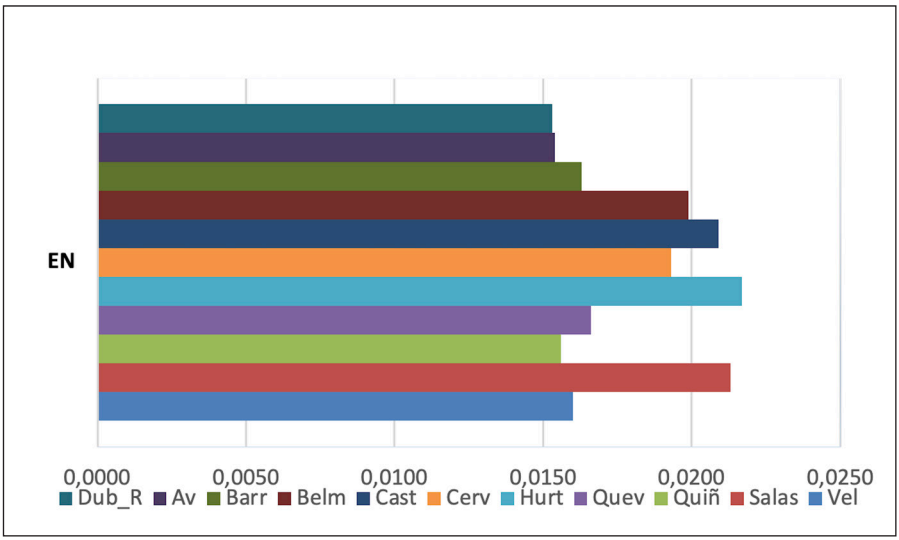

Figura 20.
Análisis comparado de la frecuencia relativa de *en*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

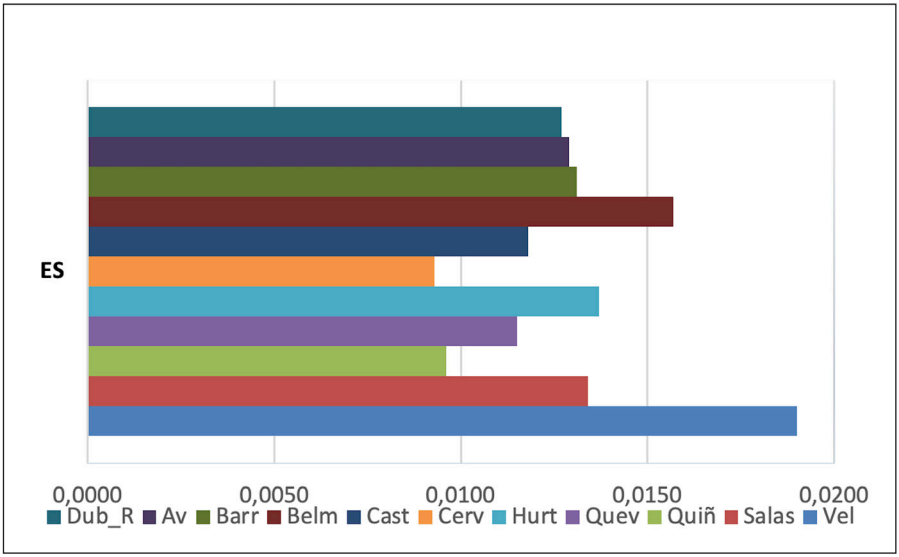

Figura 21.
Análisis comparado de la frecuencia relativa de *es*
(Elaboración propia, a partir de los datos obtenidos en *RStudio*)

Los diagramas de barras reflejan un comportamiento absolutamente diferencial entre el autor anónimo y Cervantes en el uso de estas diez palabras de función. Las únicas frecuencias relativas que parecen ser acordes en ambos son *no* y, en menor medida, *a*. Así queda también de manifiesto en el siguiente gráfico, realizado con *Voyant Tools*, en donde mostramos la tendencia de las diez palabras de manera conjunta:

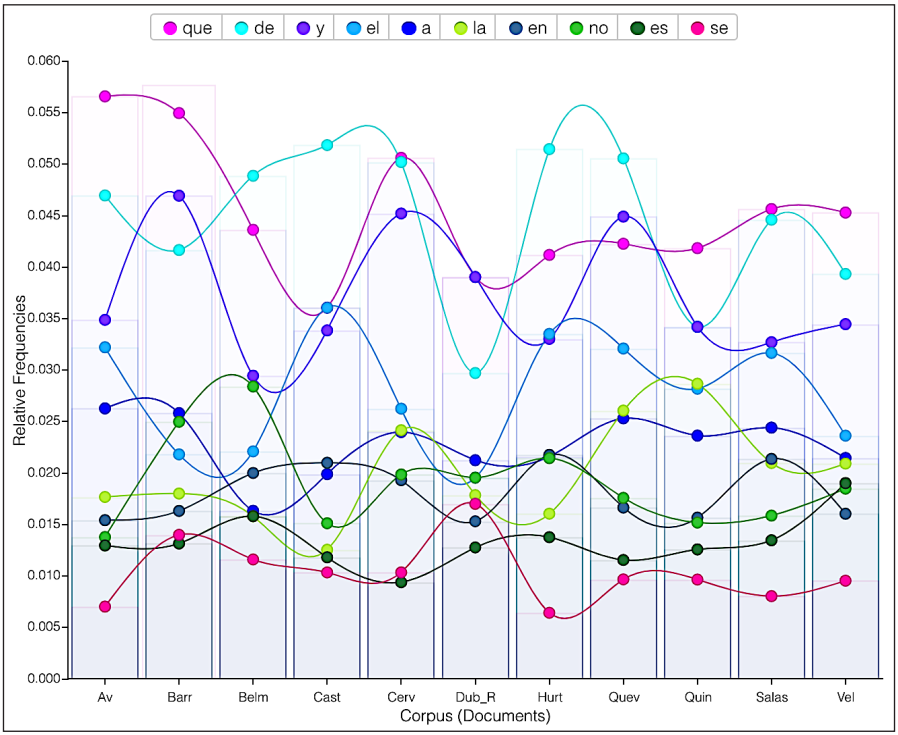

Figura 22.
Análisis comparado de la frecuencia relativa de las 10 primeras palabras de función en el *Entremés de los romances* (*Voyant Tools*. Tendencias)

Vamos a realizar ahora el PCA de la tabla de frecuencias relativas (Figura 11), con el objetivo de condensar la información en los componentes principales. Mostramos a continuación el resultado del gráfico de individuos (Figura 23), el de variables (Figura 24) y el biplot de individuos y variables (Figura 25). Las claves numéricas para cada autor son las siguientes: (1) Dubitado, (2) Ávila, (3) Barrionuevo, (4) Belmonte, (5) Castillo Solórzano, (6) Cervantes, (7) Hurtado de Mendoza, (8) Quevedo, (9) Quiñones de Benavente, (10) Salas Barbadillo y (11) Vélez de Guevara.

El gráfico de individuos (autores) refleja la cercanía entre el autor del texto

anónimo (1) y Hurtado de Mendoza (7), situados ambos en el tercer cuadrante, a escasa distancia. Cervantes (6) está muy alejado del centro, lo que significa que sus datos representan una variabilidad más extrema:

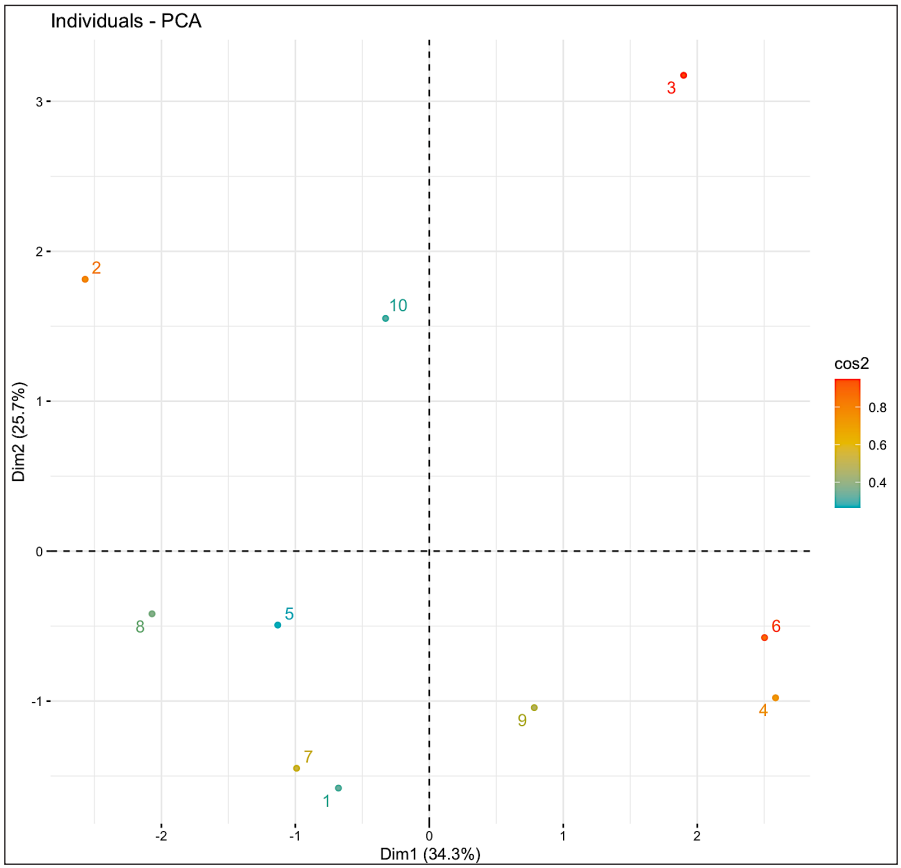

Figura 23.
PCA de las palabras de función. Gráfico de individuos.
(*RStudio*. PCA. Graph of individuals)

El gráfico de variables ofrece la visualización de las diez palabras de función en vectores. Las variables correlacionadas positivas apuntan al mismo lado de la gráfica, mientras que las negativas lo hacen en lados opuestos:

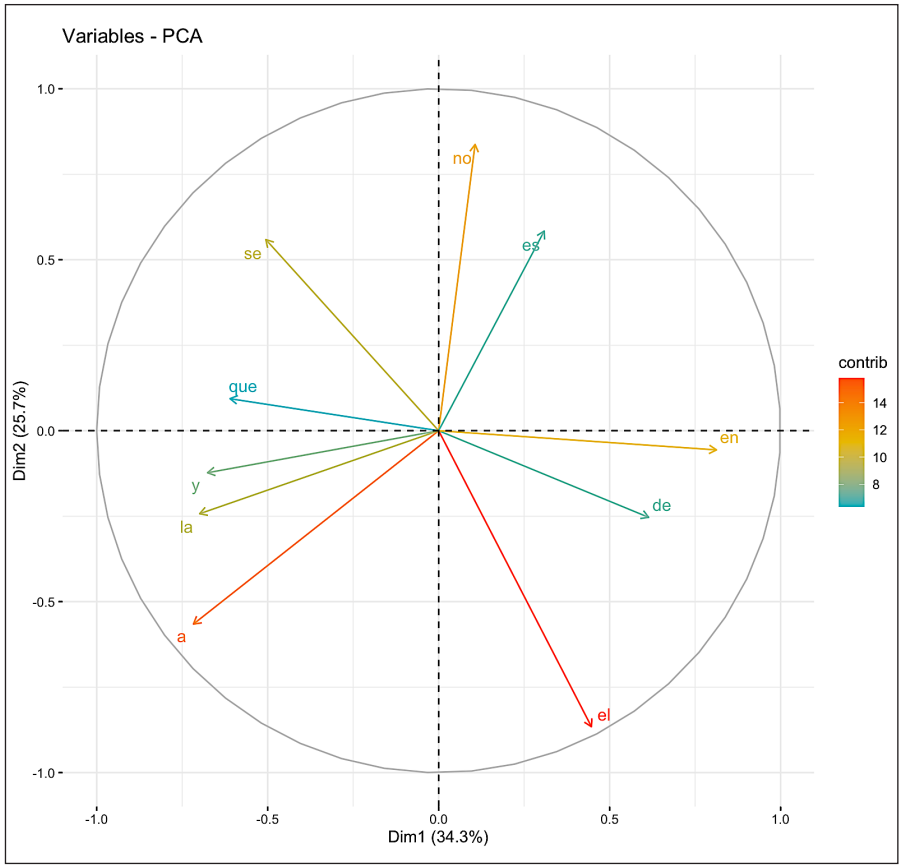

Figura 24.
PCA de las palabras de función. Gráfico de variables.
(*RStudio*. PCA. Graph of variables)

Por último, el biplot que relaciona a los individuos (autores) con las variables (palabras de función) muestra una evidente cercanía entre el autor del *Entremés de los romances* (1) y Hurtado de Mendoza (7) en el uso de *y*, *la* y *a*. Cervantes está en el cuarto cuadrante, próximo a Belmonte en el uso de las preposiciones *en* y *de*.

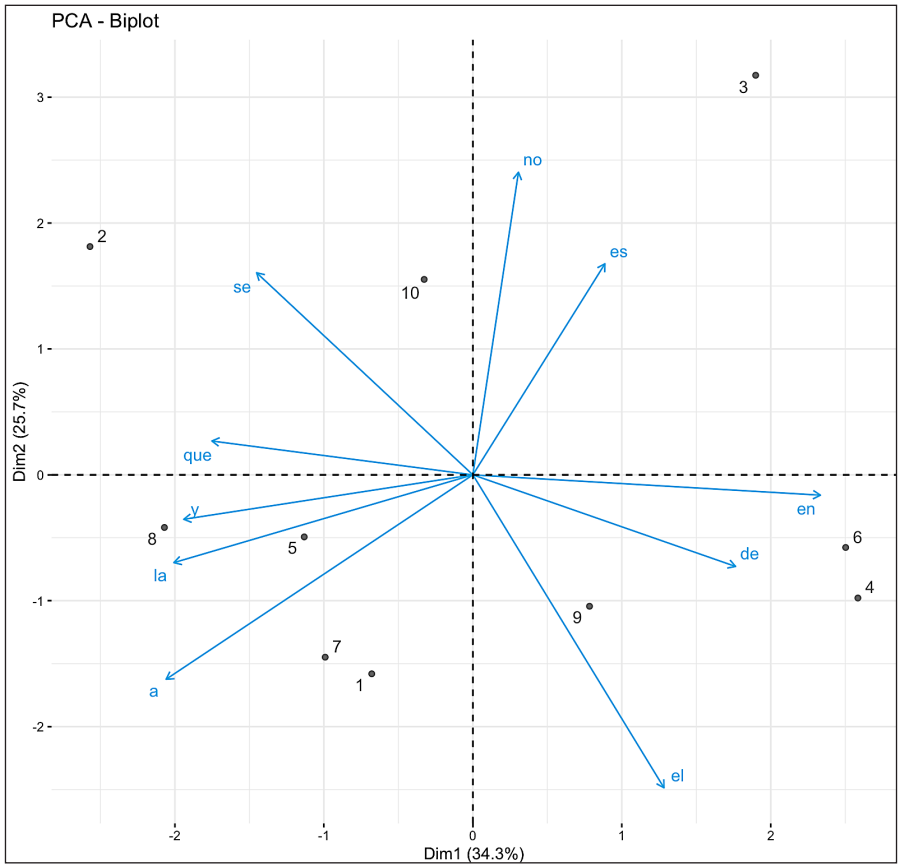

Figura 25.
PCA de las palabras de función. Biplot de individuos y variables
(*RStudio*. PCA. Biplot of individuals and variables)

El estudio del orden y la frecuencia de las palabras de función también contribuye a reforzar la hipótesis de que Cervantes no es el autor del *Entremés de los romances*.

Análisis de la frecuencia de las distintas clases de palabras

Otra de las variables características del idiolecto de un determinado individuo es la frecuencia de las distintas clases de palabras (sustantivos, adjetivos, verbos, etc.), que puede calcularse fácilmente con la ayuda de etiquetadores gramaticales automáticos, también llamados desambiguadores léxicos (en inglés, *Part-Of-Speech tagging* o *POS-tagging*). Los etiquetadores asignan a cada palabra una categoría gramatical, en función de su forma y contexto, y, a partir de esa información, elaboran listados de frecuencias de clases de palabras o de secuencias morfosintácticas (bigramas, trigramas...).

Existen varios etiquetadores automáticos, pero nos hemos decantado por FreeLing (versión 4.2_2), a pesar de que su utilización con R es compleja, ya que dispone de diccionario para el español. Aunque FreeLing funciona bastante bien para el español moderno, siempre es necesario supervisar el etiquetado, pues es posible encontrar errores de catalogación en aquellas formas que pueden pertenecer a más de una categoría (pensemos, por ejemplo, en la palabra "bajo", que puede ser adjetivo, sustantivo o preposición). En el caso del castellano del Siglo de Oro, el proceso de revisión es aún más necesario, pues encontramos errores frecuentes en aquellas palabras en desuso o en el caso de formas contractas y verbos con pronombres enclíticos.[15]

En la figura 26 mostramos el resultado de calcular la frecuencia de las distintas clases de palabras en el *Entremés de los romances* y en los dos entremeses en verso de Cervantes (*La elección de los alcaldes de Daganzo* y *El rufián viudo*), tras la revisión del etiquetado ofrecido por FreeLing. Como el análisis nos arroja frecuencias absolutas, hemos trasladado la información a un diagrama de barras, ya que así podremos advertir mejor las posibles similitudes o diferencias en el orden y frecuencia de uso de cada categoría (Figura 27).[16]

Clases de palabras	Dub_EV_Romances_R	Cerv_EV
A. Adjetivos (ADJ)	39	128
CC. Conjunciones coordinadas (CCONJ)	63	159
CS. Conjunciones subordinadas (SCONJ)	48	94
D. Determinantes (DET)	108	386

15. En nuestros textos, por ejemplo, FreeLing catalogaba *agora* como sustantivo (NCFS000) en lugar de adverbio (RG), *detuviéredes* como sustantivo (NCFS000) en lugar de verbo (VMSF2P0) o "destas" como sustantivo (NCMN000), sin separarlo en preposición (SP) y demostrativo (DD0FP0), como sí hace en las formas contractas actuales (*del* y *al*).

16. No estudiamos los bigramas y trigramas de clases de palabras ya que, al tratarse de obras en verso, los patrones morfosintácticos pueden verse condicionados por la necesidad de cuadrar el ritmo, el metro o la rima.

Clases de palabras	Dub_EV_Romances_R	Cerv_EV
F. Signos de puntuación (PUNT)	380	669
I. Interjecciones (INTJ)	20	24
M. Numerales (NUM)	2	32
NC. Sustantivos (SUST)	189	492
NP. Nombres propios (NPROP)	83	111
P. Pronombres (PRON)	176	310
R. Adverbios (ADV)	67	165
S. Preposiciones (PREP)	106	327
VA. Auxiliares (AUX)	40	89
VM. Verbos (VERB)	241	466

Figura 26.
Frecuencias absolutas por clases de palabras (*RStudio*. FreeLing. Datos)

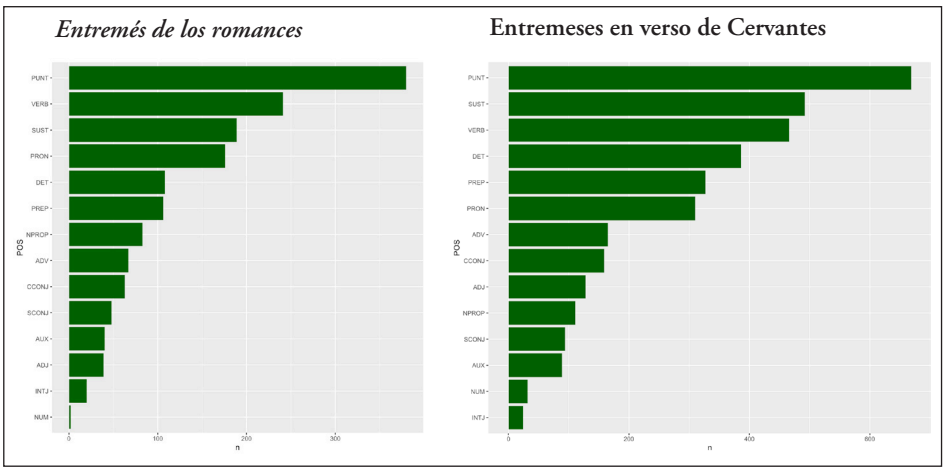

Figura 27.
Frecuencias absolutas por clases de palabras (*RStudio*. FreeLing. Ggplot)

Dejando a un lado la puntuación, que no es marca propia de autoría en el Siglo de Oro, apreciamos algunas diferencias entre ambos conjuntos textuales. En el *Entremés de los romances*, la categoría más utilizada es el verbo, seguida a cierta distancia del nombre común; en los entremeses en verso de Cervantes, lo es el nombre común, seguido muy de cerca por el verbo. También hay un comportamiento desigual en el uso de los adjetivos (con mucha mayor presencia en Cervantes) y los nombres propios y los pronombres (mucho más habituales en el dubitado). En el caso de las conjunciones, hay más subordinación en el texto dubitado que en Cervantes.

Uso y frecuencia relativa de las conjunciones coordinadas y subordinadas adverbiales

Analizamos, en primer lugar, las frecuencias relativas de las conjunciones coordinadas más habituales: la copulativas *y* e *ni*, la disyuntiva *o*, la explicativa *esto es* y las adversativas *pero*, *sino* y *mas* (Figuras 28 y 29).

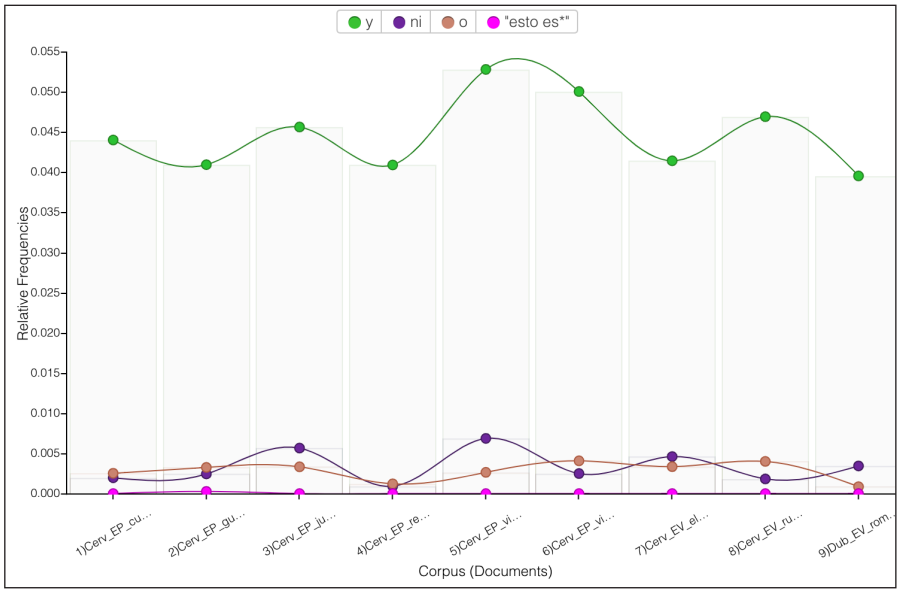

Figura 28.
Conjunciones coordinadas copulativas, disyuntivas y explicativas.
Frecuencias relativas (*Voyant Tools*. Tendencias).

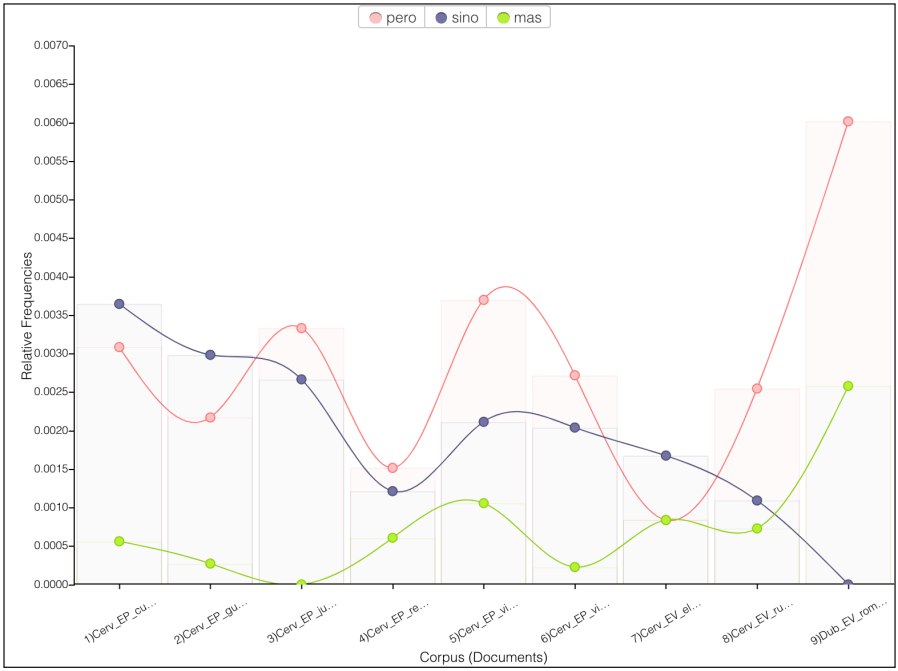

Figura 29.
Conjunciones coordinadas adversativas.
Frecuencias relativas (*Voyant Tools*. Tendencias).

Tal y como se desprende de la Figura 28, no hay grandes disonancias en el uso de las conjunciones coordinadas copulativas, disyuntivas y explicativas; las conjunciones *y* y *o* son algo menos usadas por el autor anónimo que por Cervantes, pero la diferencia es poco relevante. Mayores diferencias apreciamos en el uso de las conjunciones coordinadas adversativas (Figura 29), pues el texto dubitado muestra una clara predilección por *pero* y, en menor medida, por *mas*, mientras que en los ocho entremeses de Cervantes se da el comportamiento contrario: mayor frecuencia relativa de *sino* (que en el anónimo no aparece) y menor de *pero* y *mas*.

Pasamos ahora al análisis de las principales conjunciones subordinadas: *si* (condicional), *para que* (final), *pues* (causal), *porque* (causal), *ya que* (causal), *cuando*/*cuándo* (temporal), *donde*/*dónde* (de lugar), *aunque* (concesiva en la mayoría de los casos), *según* (modal), *más que* (comparativa) y *menos que* (comparativa).

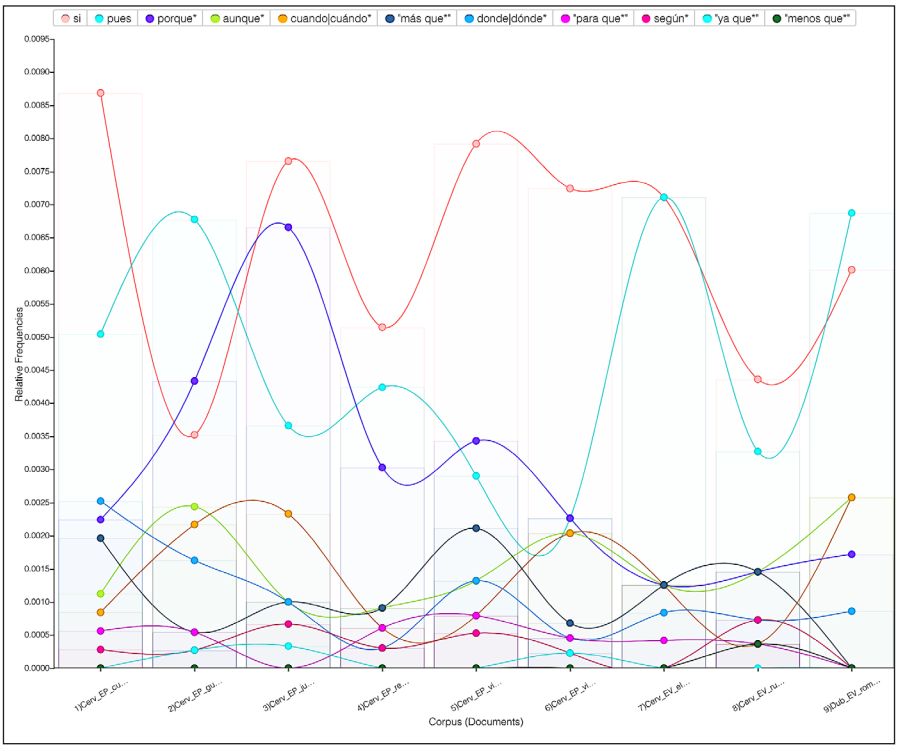

Figura 30.
Conjunciones coordinadas adversativas.
Frecuencias relativas. (*Voyant Tools*. Tendencias)

Las líneas de tendencia de la figura 30 nos permiten hacer algunas consideraciones:

— En las ocho obras cervantinas, hay una alta variabilidad en las frecuencias relativas de ciertas conjunciones, como *si*, *pues* o *porque*, lo que impide establecer conclusiones fiables sobre ellas. En todo caso, el comportamiento del texto dubitado se ubica en los tres casos dentro de los intervalos cervantinos.
— La frecuencia de las conjunciones *donde* /*dónde* y *menos que* es similar en todos los textos.
— Las conjunciones *para que* y *según* aparecen en siete de los ocho entremeses de Cervantes, pero no en el texto dubitado.
— La conjunción *más que* está presente en los ocho textos cervantinos, con una frecuencia de uso de [0,0005 , 0,0021], pero no en el *Entremés de los romances*.
— La conjunción *aunque* aparece en todos los textos, pero su frecuencia en el dubitado (0,0026) es ligeramente superior a la que presenta en los

entremeses indubitados [0,009 , 0,0024]. Lo mismo ocurre con *cuando* / *cuándo*, con una presencia en el anónimo (0,0026) superior a la que tiene en los textos de Cervantes ([0,0004 , 0,0023]).

Conclusiones

Los análisis contrastivos realizados en este trabajo nos permiten establecer una serie de conclusiones sobre la posible autoría cervantina de la parte original de esta pieza, es decir, aquella ajena a los versos parodiados del Romancero, que vienen a complementar las alcanzadas en 2010:

1. Tal y como muestran los distintos análisis de la distribución de la frecuencia de los primeros 1000 gramas de palabra, realizados con el algoritmo Eder's Delta (*cluster analysis* y *bootstrap consensus tree*), el texto dubitado siempre se ubica en una rama diferente a la que integra los ocho entremeses indubitados de Cervantes.
2. El análisis de componentes principales de la distancia delta (Eder's Delta) muestra una clara disparidad entre el *Entremés de los romances* y las ocho piezas cervantinas. El texto dubitado se sitúa en el cuarto cuadrante, en el que no figura ninguno de los ocho entremeses de autoría cierta.
3. El análisis de clasificaciones supervisadas por aprendizaje automático (Classify) asocia correctamente todos los textos indubitados a partir de las 700 palabras más frecuentes; el texto dubitado sólo se asocia con una obra de Cervantes en las 100 MFW, pero, a partir de esa cantidad, lo hace siempre con una obra de Francisco de Ávila (200-1000 MFW).
4. El análisis de verificación de autoría, realizado con el método General Imposters (GI), permite reforzar la hipótesis no sólo de que Cervantes no es el autor del texto dubitado, sino también de que probablemente no lo sea ninguno de los autores del corpus de respaldo (pues sólo en una ocasión se aproxima a Belmonte por encima del 0,63).
5. El estudio del orden de frecuencia, la frecuencia relativa y el análisis de componentes principales (PCA) de las palabras de función tampoco permite identificar al autor del texto anónimo con Cervantes.
6. El *Entremés de los romances* y los dos entremeses en verso de Cervantes presentan diferencias en el uso de ciertas clases de palabras. El autor del texto dubitado utiliza más verbos, nombres propios, pronombres y conjunciones subordinadas que Cervantes y menos adjetivos y nombres comunes.
7. El autor del texto anónimo se aproxima bastante a Cervantes en el uso de las conjunciones coordinadas copulativas (*y* y *ni*), disyuntivas (*o*) y explicativas (*esto es*), pero no en el de las adversativas: mientras él prefiere el uso de *pero* y, en menor medida, de *mas*, el autor del *Quijote* se

decanta por el uso de *sino* (que en el anónimo no aparece) y utiliza en menor proporción las formas *pero* y *mas*. En el caso de las conjunciones subordinadas adverbiales, hay ciertos usos similares (como ocurre con *donde* / *dónde*) pero también notables diferencias (*más que*, *para que*, *según*, *aunque* y *cuando* / *cuándo*).

A la luz de todo lo analizado, podemos determinar que Miguel de Cervantes no es el autor del *Entremés de los romances*, con un grado de probabilidad bastante alto (nivel -4, en una escala de opinión verbal del -5 al +5), ya que el idiolecto que se desprende del estudio de sus entremeses indubitados presenta muy pocos rasgos coincidentes con el idiolecto del texto dubitado.

Bibliografía

Arellano, Ignacio y Celsa Carmen García-Valdés, "El *Entremés el marido pantasma*, de Quevedo", *La Perinola*, núm. 1 (1997), pp. 41-68.

Asensio, Eugenio, *Itinerario del entremés, desde Lope de Rueda a Quiñones de Benavente; con cinco entremeses inéditos de D. Francisco de Quevedo*, Madrid, Gredos, 2ª edición revisada, 1971.

Baras Escolá, Alfredo, ed., Miguel de Cervantes, *Entremeses*, Madrid, Real Academia Española / Barcelona, Galaxia Gutenberg y Círculo de Lectores, 2012.

Blasco, Javier, "Avellaneda desde la estilometría", en *Cervantes: los viajes y los días*, ed. Pedro Ruiz Pérez, Madrid, Sial Ediciones, 2016, pp. 97-116.

Blasco, Javier, "Más allá del romancero: *Entremés de los romances*", *Edad de Oro*, XXXII (2013), pp. 31-45, en línea, <https://revistas.uam.es/edadoro/issue/view/178/78>.

Burrows, John Frederick, "«Delta»: A measure of stylistic difference and a guide to likely authorship", *Literary and Linguistic Computing*, XVII, núm 3 (2002), pp. 267-287.

Campo, Agustín del, ed., Miguel de Cervantes, *Entremeses*, Madrid, Clásicos Castilla, 1948.

Carreira, Antonio, ed., Luis de Góngora, *Romances*, Barcelona, Quaderns Crema,1998, 4 vols.

Castro, Adolfo de, *Varias obras inéditas de Cervantes,* Madrid, A. de Castro e hijos editores, 1874.

Cerezo Soler, Juan y José Calvo Tello, "Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de *La conquista de Jerusalén*", *Anales Cervantinos*, LI (2019), pp. 231-250, en línea <https://analescervantinos.revistas.csic.es/index.php/analescervantinos/article/view/453/453>.

Cotarelo y Mori, Emilio, *Colección de entremeses, loas, bailes, jácaras y mojigangas desde fines del siglo XVI á mediados del XVIII*, Madrid, Casa Bailly Bailliére, 1911, tomo I, vol. I.

Cuéllar, Ávaro y Germán Vega García-Luengos, "*La francesa Laura*. El hallazgo de una nueva comedia del Lope de Vega ultimo", *Anuario Lope de Vega. Texto, literatura, cultura*, XXIX (2023), pp. 131-198, en línea <https://revistes.uab.cat/anuariolopedevega/article/view/v29-cuellar-vega-garcia-luengos/492-pdf-es>.

Eder, Maciej, "Bootstrapping Delta: A safety net in open-set authorship attribution", *Digital Humanities 2013. Conference abstracts. University of Nebraska/Lincoln, USA. 16-19 July 2013*, 2013, pp. 169-172, en línea, <http://dh2013.unl.edu/abstracts/ab-135.html>.

Eder, Maciej, Mike Kestemont y Jan Rybicki, "Stylometry with R: A Package for Computational Text Analysis", *The R Journal*, VIII, núm 1 (2016), pp. 107-121.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, y Thorsten Vitt, "Understanding and explaining Delta measures for authorship attribution", *Digital Scholarship in the Humanities*, XXXII, núm. 2 (2017), pp. 4-16, en línea <https://academic.oup.com/dsh/article/32/suppl_2/ii4/3865676>.

Jockers, Matthew y Daniela M. Witten, "A comparative study of machine learning methods for authorship attribution", *Literary and Linguistic Computing*, XXV, núm. 2 (2010), pp. 215-523.

Jolliffe, I. T. (2002), *Principal Component Analysis*, New York, Springer, 2ª edición, 2002.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp y Walter Daelemans, "Authenticating the writings of Julius Caesar", *Expert Systems with Applications*, LXIII (2016), pp. 86-96.

Koppel, Moshe y Yaron Winter, "Determining if two documents are written by the same author", *Journal of the Association for Information Science and Technology*, LXV, núm. 1 (2014), pp. 178-87.

López Navío, José, "*El entremés de los romances*, sátira contra Lope de Vega, fuente de inspiración de los primeros capítulos del *Quijote*", *Anales cervantinos*, VIII (1959-1960), pp. 151-239.

Márquez Villanueva, Francisco, "Lope, infamado de morisco", *Anuario de Letras*, XXI (1983), pp. 147-182, en línea, <https://revistas-filologicas.unam.mx/anuario-letras/index.php/al/article/view/50/50>.

Menéndez Pelayo, Marcelino, "Obras inéditas de Cervantes", en *Estudios y discursos de crítica histórica y literaria*, Madrid, CSIC, 1941, tomo I, pp. 269-302.

Menéndez Pelayo, Marcelino, ed., *Flor de entremeses y sainetes de diferentes autores* (1657), Madrid, Imprenta de Fortanet, 2ª edición corregida, 1903.

Menéndez Pidal, Ramón, "Un aspecto en la elaboración del *Quijote*", en *De Cervantes y Lope de Vega*, Madrid, Espasa-Calpe, 1964, pp. 9-60.

Millé y Giménez, Juan, *Sobre la génesis del Quijote: Cervantes, Lope, Góngora, el "Romancero general", el "Entremés de los Romances", etc.,* Barcelona, Araluce, 1930.

Northup, George T., "Review de 'Un aspecto en la elaboración del *Quijote*' de Menéndez Pidal", *Modern Philology*, XIX, núm. 4 (1922), pp. 435-436.

Pérez López, José Luis, "Los romances del realismo bucólico de Liñán de Riaza y de Lope de Vega, el *Entremés de los romances* y el *Quijote*", *Anuario Lope de Vega*, XV (2009), pp. 169-201.

Rey Hazas, Antonio, *El nacimiento del* Quijote*. Edición y estudio del* Entremés de los romances*,* Guanajuato, Museo Iconográfico del *Quijote*, 2006.

Rissler-Pipka, Nanette, "Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales", en *El otro don Quijote. La continuación de Fernández de Avellaneda y sus efectos*, ed. Hanno Ehrlicher, Augsburgo, Universität Augsburg, 2016, pp. 27-51.

Rodríguez López-Vázquez, Alfredo, "*El entremés de los romances*, entre Cervantes y Góngora", *Atalanta: Revista de Letras Barrocas*, VII, núm. 2 (2019), pp. 221-239, en línea, <https://www.revistaatalanta.com/index.php/ARLB/article/view/133/144>.

Rosa, Javier de la y Juan Luis Suárez, "The Life of *Lazarillo de Tormes* and of his Machine Learning Adversities. Non-Traditional Authorship Attribution Techniques in the Context of the *Lazarillo*", *Lemir*, XX (2016), pp. 373-438, en línea, <https://parnaseo.uv.es/lemir/Revista/Revista20/09_Rosa_Javier_de_la.pdf>.

Ruiz Urbón, Cristina, "El *Entremés de los romances*: una atribución cervantina largamente dubitada", en *Hos ergo versículos feci... Estudios de atribución y plagio*, eds. Javier Blasco, Patricia Marín Cepeda y Cristina Ruiz Urbón, Madrid, Iberoamericana, 2010, pp. 171-260.

Ruiz Urbón, Cristina, "Sobre la validez de los análisis cuantitativos en los estudios de autoría de textos breves: el caso particular de los entremeses del Siglo de Oro", *Ogigia. Revista electrónica de estudios hispánicos*, núm. 33 (2023), pp. 69-96, en línea, <https://revistas.uva.es/index.php/ogigia/article/view/7143>.