OGIGIA
Revista electrónica de estudios hispánicos

Ed UVa EDICIONES Universidad Valladolid

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

On the validity of quantitative analyzes in authorship studies of short texts: the particular case of Golden Age *entremeses*

CRISTINA RUIZ URBÓN
Universidad de Valladolid
cristina.ruiz.urbon@uva.es
ORCID: https://orcid.org/0000-0002-5535-2703
Recibido: 03.01.2023. Aceptado: 22.01.2023.
Cómo citar: Ruiz Urbón, Cristina (2023). "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*, 33: 69-96
DOI: https://doi.org/10.24197/ogigia/33.2023.69-96

Resumen: Ante las afirmaciones de varios especialistas que declaran la dificultad de aplicar análisis cuantitativos a los estudios de atribución de autoría de textos literarios breves, inferiores a las 2000-2500 palabras, en este trabajo proponemos un diseño experimental basado en la autoría de una serie de entremeses del Siglo de Oro español, con el objetivo de establecer si dicho umbral es realmente determinante para este género literario o no. Los análisis nos permitirán estudiar, asimismo, si la señal autorial o el vehículo de expresión (prosa/verso) también condicionan la fiabilidad de los resultados.
Palabras clave: atribución de autoría; entremeses; Siglo de Oro; análisis cuantitativos; idiolecto.
Abstract: Given the statements of several specialists who declare the dificulty of applying quantitative analysis to the study of authorship of short literary texts, less than 2000-2500 words, in this paper we propose an experimental design based on the authorship of a series of Spanish Golden Age entremeses. Our objective is to establish if this threshold is really decisive for this literary genre or not. The analyzes will also allow us to study whether the authorial signal or the vehicle of expression (prose/verse) also condition the reliability of the results.
Keywords: Authorship attribution; *Entremeses*; Golden Age; Quantitative analysis; Idiolect.

INTRODUCCIÓN

El principio teórico fundamental que sustenta los estudios de autoría literaria es la idea de que todo individuo posee un estilo idiolectal propio,

es decir, una realización individual e idiosincrásica del lenguaje, que permite distinguirlo de los demás. No obstante, y a diferencia de lo que ocurre con la huella dactilar o el ADN genético, las muestras lingüísticas ofrecen una información parcial, lo que nos lleva a preguntarnos qué cantidad mínima de texto necesitamos para que la atribución de una obra a un determinado autor sea realmente fiable.

En el ámbito forense, es bastante habitual que los especialistas deban enfrentarse a textos extremadamente breves, como *whatsapps*, tuits o correos electrónicos, que rara vez superan las 200 palabras (Coulthard, 2005); en esos casos, los estudios de autoría se sustentan en análisis de tipo cualitativo, ya que, para poder aplicar una metodología cuantitativa, necesitamos disponer de textos de mayor extensión. Hay estudios que han alcanzado resultados realmente prometedores con escritos breves (Sanderson y Guenter, 2006; Koppel, Schler y Bonchek-Dokow, 2007; Luyckx, 2010), pero lo cierto es que la precisión de los análisis cuantitativos decrece significativamente cuando la longitud del texto es inferior a las 1000 palabras (Hirst y Feiguina, 2007). En todo caso, y tal y como advierte Stamatatos, la longitud no es el único requisito esencial:

> It is not yet clear whether other factors (beyond text length) also affect this process. For example, let a and b be two texts of 100 words and 1,000 words, respectively. A given authorship attribution tool can easily identify the author of a, but not the author of b. What are the properties of a that make it an easy case, and what makes b so difficult, albeit much longer, than a? (2009: 553).

En el campo de la atribución de autoría literaria tampoco hay consenso sobre este punto. Uno de los primeros en reflexionar sobre esta cuestión fue Maciej Eder (2015), quien, tras realizar varios análisis estilométricos sobre textos literarios de diferentes géneros y lenguas, llegó a la conclusión de que el umbral de fiabilidad autorial se sitúa de manera general en torno a las 5000 palabras (si bien, en el caso de la poesía, este límite podría reducirse hasta las 3000); cabe señalar, no obstante, que su estudio no contempla ningún texto en lengua española. En un trabajo posterior, Eder volvió sobre este asunto, considerando entonces –en sintonía con la opinión de Statamatos– que, más que la extensión, lo importante es la fuerza idiosincrásica de la señal autorial, y que, en ciertos casos, "a sufficient amount of textual data may be as little as 2,000 words" (2017: 223).

Sea como fuere, lo cierto es que la mayor parte de los hispanistas han centrado sus esfuerzos en obras de considerable extensión –como novelas, comedias o poemarios–, y pocos se han atrevido a estudiar la autoría en conflicto de textos breves. Laura Hernández Lorenzo, tras aplicar métodos supervisados a un corpus de poemas de 53 autores del Siglo de Oro, llega a la conclusión de que "con 2000 palabras, e incluso menos, pueden realizarse análisis de atribución de autoría con resultados altamente fiables, siempre que el autor posea una señal autorial fuerte y distintiva" (2019b: 202). Y en la misma línea se manifiestan Miguel Campión Larumbe y Álvaro Cuéllar, al considerar que la aplicación de ciertos análisis estilométricos a obras literarias empieza "a ofrecer resultados fiables a partir de textos con 2000-2500 palabras" (2021: 62).

En este trabajo, y con el objetivo de seguir profundizando en esta cuestión, vamos a aplicar algunas de las técnicas cuantitativas que mejores resultados están arrojando en la atribución de autoría lingüística a una serie de entremeses del Siglo de Oro, tanto en prosa como en verso, con el objetivo de determinar si ese umbral de 2000-2500 palabras es realmente determinante o no para este tipo de género literario; y, asimismo, si hay diferencias significativas entre los diferentes autores o entre la prosa y el verso. Los resultados nos permitirán saber si dichas técnicas podrían ser aplicadas en el futuro a la larga lista de entremeses anónimos o de dudosa paternidad que conforman nuestra historia literaria, generalmente olvidados por la crítica[1].

1. El caso particular de los entremeses del Siglo de Oro

El complejo proceso de creación artística y la larga cadena de eslabones que conforman la transmisión textual del teatro del Siglo de Oro español dificultan no sólo la identificación de los autores reales, sino también la conservación fidedigna de las obras. Estas dificultades se agravan aún más en el caso de los entremeses, por el carácter secundario que tienen respecto de la obra principal, pero, sobre todo, por su naturaleza mixta y su corta extensión. Por ello nos atreveríamos a decir que, si hay un género literario realmente complejo a la hora de

[1] Mención especial merece el trabajo de Javier Blasco sobre *El entremés de los mirones*, atribuido tradicionalmente a Cervantes (Blasco, 2019a). Cabe señalar, no obstante, que esta pieza, a pesar de llevar en el título la palabra "entremés", es una especie de novela ejemplar dialogada –del tipo de *Rinconete y Cortadillo* o *El coloquio de los perros*–, con una extensión muy superior a la de las obras de teatro breve (9220 palabras).

discriminar autorías, este es sin duda alguna el entremés, ya que, a los inconvenientes generales de los estudios atributivos, hay que sumar toda una serie de limitaciones metodológicas derivadas de las especiales características de estos textos dubitados (brevedad, contaminación textual y autorial, alternancia de prosa y verso, predilección por temas y lugares comunes, etc.).

1. 1. Descripción del corpus de trabajo

Nuestro corpus de trabajo lo conforman 32 entremeses de finales del XVI y principios del XVII, tanto en prosa como en verso, de diez autores distintos (Francisco de Ávila, Gaspar de Barrionuevo y Carrión, Luis de Belmonte Bermúdez, Alonso de Castillo de Solórzano, Cervantes, Antonio Hurtado de Mendoza, Francisco de Quevedo, Luis Quiñones de Benavente, Alonso Jerónimo de Salas Barbadillo y Luis Vélez de Guevara), cuya extensión oscila entre las 871 y las 4701 palabras; si bien, 20 de los 32 entremeses tienen una extensión inferior a las 2500 palabras (Tabla 1). Hemos asignado a cada texto un tejuelo o etiqueta identificativa, formada por tres unidades: la primera hace referencia al autor; la segunda, al vehículo de expresión, distinguiendo entre prosa ("EP") y verso ("EV"); y la tercera, al título de la obra, del que tomamos como referencia la primera palabra con significado léxico (sin tildes).

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 formato 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*, *esotro*…), las desinencias verbales enclíticas (del tipo *cogióla* o *dejáronles*) y las formas verbales arcaicas en *–alle*, *–elle* e *–illa* (*satisfacelle*, *vencelle*, *decilla*…). 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– pueden resultar idiosincrásicas. El corpus de trabajo está disponible en el siguiente repositorio de *GitHub*: https://github.com/Cruizurbon/Entremeses-SdO.

Tabla 1. Corpus de trabajo (32 entremeses del Siglo de Oro)

AUTOR	TÍTULO	ETIQUETA	EXTENSIÓN
Francisco de 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	2 268 palabras
Gaspar de Barrionuevo y Carrión	*El triunfo de los coches*	Barr_EP_triunfo	4 701 palabras
Luis de Belmonte Bermúdez	*Sierra Morena de las mujeres*	Belm_EV_sierra	1 012 palabras
	Una rana hace ciento	Belm_EV_rana	871 palabras
Alonso de Castillo de Solórzano	*El barbador*	Cast_EV_barbador	1 294 palabras
	El comisario de figuras	Cast_EV_comisario	1 422 palabras
Miguel de Cervantes Saavedra	*El juez de los divorcios*	Cerv_EP_juez	3 004 palabras
	El rufián viudo	Cerv_EV_rufian	2 748 palabras
	La elección de los alcaldes de Daganzo	Cerv_EV_eleccion	2 386 palabras
	La guarda cuidadosa	Cerv_EP_guarda	3 684 palabras
	El vizcaíno fingido	Cerv_EP_vizcaino	4 418 palabras
	El retablo de las maravillas	Cerv_EP_retablo	3 300 palabras
	La cueva de Salamanca	Cerv_EP_cueva	3 568 palabras
	El viejo celoso	Cerv_EP_viejo	3 790 palabras
Antonio Hurtado de Mendoza	*Getafe*	Hurt_EV_getafe	1 602 palabras
	El examinador Miser Palomo	Hurt_EV_miser1	2 353 palabras
	Miser Palomo, el médico del espíritu	Hurt_EV_miser2	2 108 palabras
Francisco de Quevedo	*El marido Pantasma*	Quev_EV_marido	1 748 palabras
	El niño y Peralvillo de Madrid	Quev_EV_niño	1 390 palabras
	La vieja Mutañones	Quev_EP_vieja	2 102 palabras
Luis Quiñones de Benavente	*El ventero*	Quin_EV_ventero	1 437 palabras
	Los sacristanes	Quin_EV_sacristanes	1 824 palabras
	El barbero	Quin_EV_barbero	986 palabras
	El borracho	Quin_EV_borracho	1 859 palabras
Alonso Jerónimo de Salas Barbadillo	*El buscaoficios*	Salas_EP_buscaoficios	3 941 palabras
	Los mirones de la Corte	Salas_EP_mirones	2 254 palabras
	El caprichoso en su gusto	Salas_EV_caprichoso	3 510 palabras
	El comisario contra los malos gustos	Salas_EV_comisario	2 729 palabras
Luis Vélez de Guevara	*La burla más sazonada*	Vel_EV_burla	1 144 palabras
	Los atarantados	Vel_EV_atarantados	1 440 palabras
	Antonia y Perales	Vel_EV_antonia	1 094 palabras

Fuente: elaboración propia

Aunque no es un corpus excesivamente extenso, ya que hemos tratado de priorizar la fiabilidad de las autorías sobre la cantidad de obras, creemos que es suficiente para el fin perseguido. Los textos seleccionados, además, son un claro ejemplo de las redes temático-argumentales y las concomitancias de personajes y espacios que configuran el espectro entremesil del Siglo de Oro, y que podrían afectar, en consecuencia, a los análisis basados en cuestiones léxicas.

Así, por ejemplo, hemos compilado varios entremeses que ridiculizan el tema del matrimonio. En *La guarda cuidadosa*, *El triunfo de los coches* o *Los atarantados* se trata el tema de los acuerdos matrimoniales, en los que prima el dinero por encima del amor; y en *La cueva de Salamanca*, *El viejo celoso* o *El mortero y chistes del sacristán*, el del adulterio femenino, heredado de los antiguos *fabliaux* medievales. En *El marido pantasma* se muestra la animadversión al casamiento a través del personaje de Lobón, quien primero desaconseja a su amigo Muñoz que se case, pero que después, tras quedar viudo, le aconseja que lo haga, "que todo puede pasarse / por ver ir en procesión, / kiriada de los niños, / la mujer que nos cansó". Y en *El juez de los divorcios* se plantea la problemática de cuatro matrimonios que quieren romper su unión por distintas razones (insatisfacción sexual, desavenencias de tipo económico, celos desmedidos…), pero que terminan cantando "que vale el peor concierto / más que el divorcio mejor".

El motivo del engaño también aparece en varias de estas piezas. En *El barbador*, Piruétano y Pescaño fingen tener poderes mágicos para hacer crecer el pelo y la barba a los lampiños; en *El retablo de las maravillas*, Chanfalla y Chirinos simulan una representación teatral ante todos los habitantes de un pueblo; en *La burla más sazonada*, el estudiante Tabaco burla a un alguacil, imitando la voz de la dama a la que está persiguiendo; en *El mortero y chistes del sacristán*, Gigorro lleva a cabo la clásica burla del intercambio entre el mortero y el manteo, que ya estaba presente en uno de los cuentos del *Decamerón*; y un intercambio entre objetos vemos también en *El vizcaíno fingido*, en donde Solórzano y Quiñones engañan a Cristina, una prostituta sevillana muy famosa por su taimería, con un cambiazo entre dos cadena de oro, una verdadera y otra falsa, en un juego que presenta muchos puntos en común con uno de los capítulos de *Guzmán de Alfarache*.

En varios de los entremeses del corpus se aborda la moda de los coches de caballos y, en concreto, la pasión desmedida que sienten las mujeres por ellos; así aparece en *El triunfo de los coches*, de

Barrionuevo, en *El comisario contra los malos gustos y El buscaoficios*, de Salas Barbadillo, en *Los atarantados*, de Vélez de Guevara, en *El marido pantasma*, de Quevedo, o en *El vizcaíno fingido*, de Cervantes, en donde Brígida siente que se le "arranca el alma" al enterarse "que quitaban los coches". El contrapunto a todas estas mujeres interesadas lo encontramos en la villana Francisca, del entremés de *Getafe*, que no sucumbe ante los requiebros de Don Lucas y sus promesas de coches, joyas, vestidos y tocados.

Otro leitmotiv recurrente es el de la limpieza de sangre y la honorabilidad del linaje, que se representa a través de chistes antijudaicos y antimoriscos y de toda una serie de tópicos sobre los conversos, relacionados, principalmente, con la repugnancia por ciertos alimentos. En *La elección de los alcaldes de Daganzo* encontramos a cuatro labradores que quieren ser alcaldes y que reivindican, ante Panduro, Alonso Algarroba, Pesuña y Pedro Estornudo, su condición de cristianos viejos. En *El retablo de las maravillas*, Chirinos y Chanfalla hacen creer al gobernador y a las autoridades del pueblo que aquellos vecinos que tengan "alguna raza de confeso, o no sea habido y procreado de sus padres de legítimo matrimonio" no podrán ver las maravillas que suceden dentro del retablo. Y el tema de la calidad del cristiano viejo también se recoge en *El juez de los divorcios*, a través del personaje de Ganapán, y en *El mortero y chistes del sacristán*, donde el vejete Pero Díaz dice haber reñido "con ese barberillo / porque se ha demandado de la lengua / en vituperio de mi honrosa estirpe, / publicando en la plaza que soy moro".

Y también está presente en varias de estas piececillas el tema de las salteadoras de la calle Mayor de Madrid. A él se hace referencia en *La vieja Mutañones, El niño y Peralvillo de Madrid* y en *Sierra Morena de las mujeres*, en donde cuatro bandoleras asaltan a Garañón y Paris.

Pero las obras que representan nuestro corpus no sólo comparten unas mismas redes temáticas, sino también unos mismos personajes y espacios. En ellas encontramos mujeres interesadas o malmaridadas, vejetes, criados bobos, galanes, sacristanes, soldados, valientes, boticarios, barberos, alcaldes rurales, estudiantes, viejas, rufianes, mesoneros y venteros, etc.; y son varios los entremeses que responden a lo que se ha definido como "desfile de tipos o figuras", en donde distintos personajes destemplados van pasando ante una figura que ejerce como juez o examinador (*El juez de los divorcios*, *El buscaoficios*, *El comisario de figuras*, *El comisario contra los malos gustos* y *El*

examinador Miser Palomo). Y lo mismo ocurre con los espacios de la ficción: la villa y Corte de Madrid es uno de los escenarios más utilizados (*El triunfo de los coches*, *El caprichoso en su gusto*, *Sierra Morena de las mujeres*, *El niño y Peralvillo de Madrid*), junto a las aldeas (*La elección de los alcaldes de Daganzo*, *El retablo de las maravillas*, *Los romances…*) o las ventas y mesones (*Getafe*).

1. 2. Selección de las pruebas cuantitativas

Hasta hace relativamente poco tiempo, los estudios literarios eran eminentemente cualitativos; sin embargo, de un tiempo a esta parte, se han ido adoptando técnicas cuantitativas para analizar los textos y entender su variación en relación con factores como el periodo histórico, el contexto, el género o el autor.

Los primeros intentos por cuantificar el estilo de escritura se remontan al siglo XIX, cuando el físico estadounidense Thomas Mendenhall midió la extensión de varios cientos de miles de palabras de las obras de Bacon, Marlowe y Shakespeare, tratando de demostrar que la longitud de palabra no es una variable discriminatoria entre autores. Ya en el siglo XX, George Unde Yule en Inglaterra y George Kingsley Zipf en Estados Unidos intentaron establecer fórmulas matemáticas para discriminar autorías, basadas en la frecuencia léxica, como la "Característica K de Yule", ineficaz según los estudios posteriores, y la "Ley Zipf", que a día de hoy se sigue utilizando. Pero no fue hasta el trabajo de los estadísticos estadounidenses Mosteller y Wallace sobre la autoría de los *Papeles Federalistas*, del año 1964, cuando los análisis computacionales empezaron a ganar adeptos entre los lingüistas: su método, basado en el análisis estadístico bayesiano de la frecuencia de palabras funcionales, ofreció resultados de discriminación significativos entre los distintos candidatos a autor. Otras investigaciones estilométricas posteriores, sin embargo, dieron lugar a múltiples controversias, como el estudio sobre la autoría de *El Libro del Mormón*, que generó un intenso debate entre los defensores de la autoría de Joseph Smith, su firmante, y los defensores de una autoría compartida por al menos cuatro personas; un debate que puso sobre la mesa la necesidad de testar algunos de los métodos propuestos. Desde finales de los años 80 del siglo XX, y más intensamente en el siglo XXI, los esfuerzos de la Estilometría han ido encaminados en esta dirección; así, se han ido descartando algunos parámetros propuestos para discriminar autorías, como la "Característica

K de Yule", antes comentada, y validando otros nuevos, como la distribución de la frecuencia de las palabras más utilizadas (*Delta distance*), el análisis de componentes principales (PCA) de las palabras de función, el estudio de las clases de palabras mediante etiquetadores gramaticales automáticos (*Pos tagging*) o la detección de largas cadenas de palabras coincidentes (*verbatim*), entre otros.

En los siguientes apartados, y con el objetivo de testar la validez de los análisis cuantitativos en los estudios de atribución de autoría de los entremeses auriseculares, vamos a llevar a cabo tres comprobaciones relacionadas esencialmente con el léxico, por ser los análisis que, a priori, están más condicionados por la extensión de los textos y los clichés propios del género entremesil. Analizaremos, en primer lugar, la distribución de la frecuencia de las palabras más empleadas en los textos; a continuación, aplicaremos esta misma medida (*Delta distance*) a las secuencias de cinco caracteres; y, por último, estudiaremos la longitud de palabras idénticas entre los distintos pares de textos (*verbatim*). No se trata de testar estos tres parámetros, suficientemente probados por la comunidad científica, sino su validez para los textos seleccionados.

1. 3. Análisis de la distribución de la frecuencia de las palabras más empleadas en los textos (*Delta distance*, 1-grama de palabra)

Una de las medidas mejor acogidas hasta la fecha para discriminar autorías es la del algoritmo Delta, propuesto por John Burrows (2002), que parte de la idea de que la variación en las frecuencias de las palabras más empleadas en un texto (*tokens*) permiten reconocer su autoría. El algoritmo de Burrows (*Classic Delta*) ha sido actualizado más tarde por Hoover (*Hoover Delta*) (2004), Argamon (*Delta Lineal*) (2008), Smith y Aldridge (*Cosine Delta*) (2011) y Maciej Eder (*Eder's Delta*) (2013), que propone una modificación pensada específicamente para las lenguas flexivas. Aunque la mayor parte de los estudios con Delta se han hecho con textos en lengua inglesa, alemana, francesa, polaca, húngara o latina, también han probado su efectividad para el español Rißler-Pipka (2016), Wrisley (2016), Calvo Tello (2016), Fradejas Rueda (2016 y 2019), Blasco (2016, 2019a, 2019b y 2022), De la Rosa y Suárez (2016), Rojas Castro (2017), Cerezo Soler y Calvo Tello (2019), García-Reidy (2019),

Hernández Lorenzo (2019a), Vega García-Luengos (2021) y Campión Larumbe y Cuellar (2021)[2].

Delta trabaja del siguiente modo: tokeniza los textos del corpus, calcula diferentes datos de frecuencia de los *tokens*, los estandariza mediante democratización léxica (*z-scores*) y compara los resultados de la suma de cada palabra en cada pareja de textos (*delta-scores*) para agruparlos jerárquicamente (Calvo Tello, 2016). Delta, al igual que otros métodos estilométricos, basa sus resultados en la comparación de un número *n* de textos; en consecuencia, si añadimos o quitamos uno o varios de esos textos, los nodos de ramas pueden variar.

Para testar la validez del análisis de la distribución de la frecuencia de las palabras más empleadas en los 32 entremeses de nuestro corpus (*Delta distance*), utilizamos el paquete Stylo (Eder, Rybicki y Kestemont, 2016), seleccionando el idioma español, el algoritmo Delta de Eder (*Eder's Delta*) y la opción "0% *culled*", ya que no consideramos necesario que las palabras más frecuentes aparezcan en un número determinado de textos. Aunque Stylo permite acceder a los archivos de cada proceso (lista de palabras, tabla de frecuencias, valores delta para cada pareja de textos, etc.), también ofrece los resultados en forma de gráfico, como dendrogramas (*Cluster Analysis*) o árboles de consenso (*Bootstrap Consensus tree*), lo que resulta ciertamente útil para su comprensión en el ámbito de las Humanidades digitales.

Mostramos el resultado en forma de dendrograma[3], considerando las 800 (Gráfico 1) y las 1000 (Gráfico 2) palabras más frecuentes (MFW):

[2] Los estudios empíricos han demostrado que la eficacia de todos estos algoritmos varía no sólo de unas lenguas a otras, sino también en función del tipo de texto (Rybicki y Eder, 2011). En español, se han empleado diferentes algoritmos: Blasco, en su trabajo sobre Avellaneda, combina Classic Delta con Eder's Delta (2016); Fradejas Rueda elige Classic Delta para su propuesta con textos medievales; (2019); García- Reidy opta por Cosine Delta para estudiar la autoría de la comedia *Siempre ayuda a la verdad* (2019); en su estudio de los textos poéticos de Fernando Herrera, Hernández Lorenzo, tras testar diferentes medidas y parámetros, llega a la conclusión de que "la mejor combinación es la aplicación de la Delta de Eder sobre las 4000 palabras más frecuentes con una media de éxito o aciertos superior al 85%" (2019a: 314); y en el proyecto *ETSO. Estilometría aplicada al Teatro del Siglo de Oro*, están usando Classic Delta (Vega García-Luengos, 2021; Campión Larumbe y Cuellar, 2021). Nosotros, al igual que Cerezo Soler y Tello (2019), optamos por Eder's Delta, que, al menos para nuestro corpus de entremeses indubitados, arroja los resultados más fiables.

[3] Hemos optado por el análisis de conglomerados para poder testar varias MFW, de cara a futuros estudios de atribución de autoría de entremeses del Siglo de Oro español, si bien es cierto que otra opción interesante –y que evita el *cherry picking*– es crear

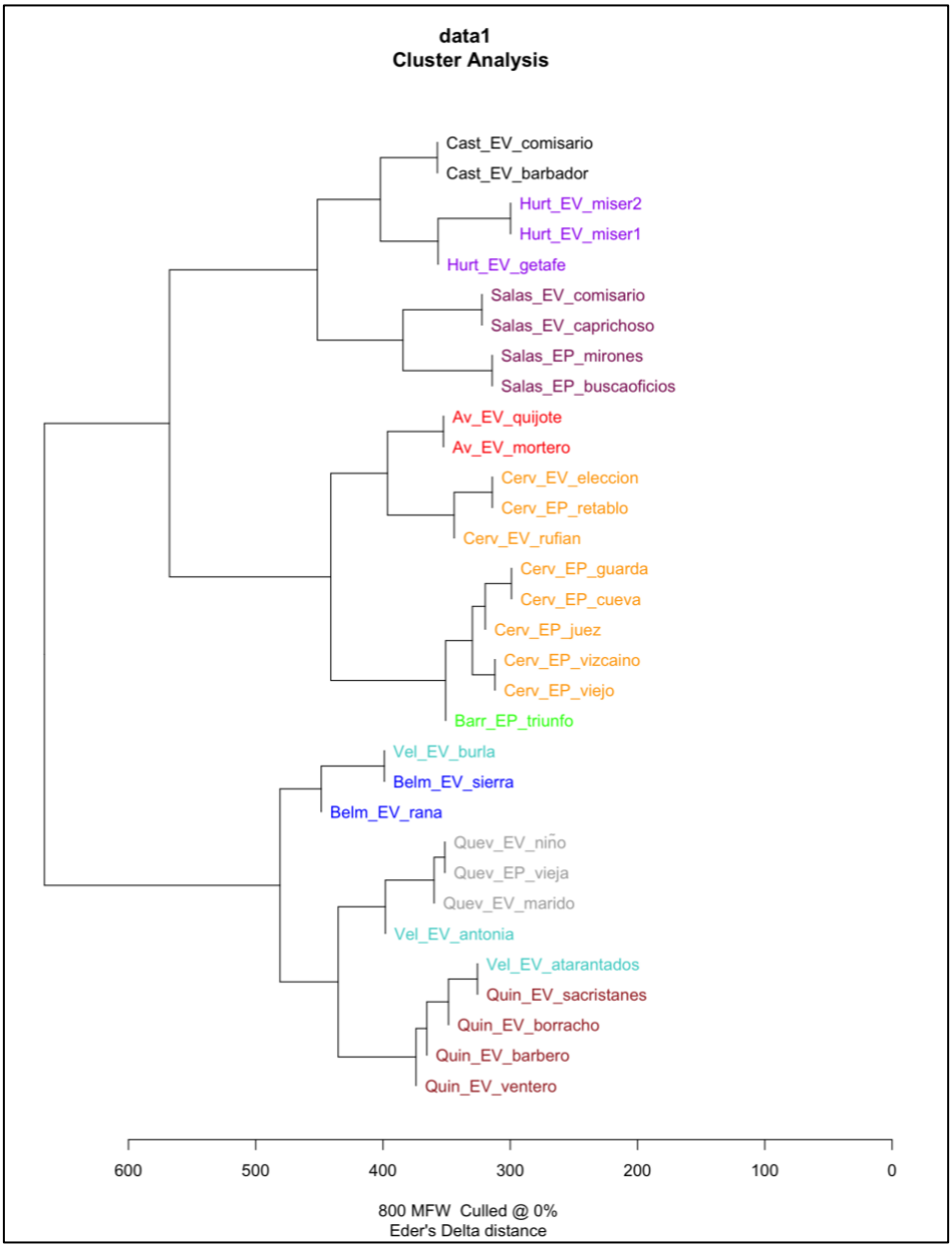

Gráfico 1. Clúster basado en la frecuencia de las palabras más empleadas (RStudio. Stylo. Cluster Analysis. Eder's Delta. MFW 800. Culled 0%)

árboles de consenso, ya que estos ofrecen la agrupación más estable, es decir, aquélla que permanece a lo largo de un porcentaje mínimo de iteraciones (Eder, 2013).

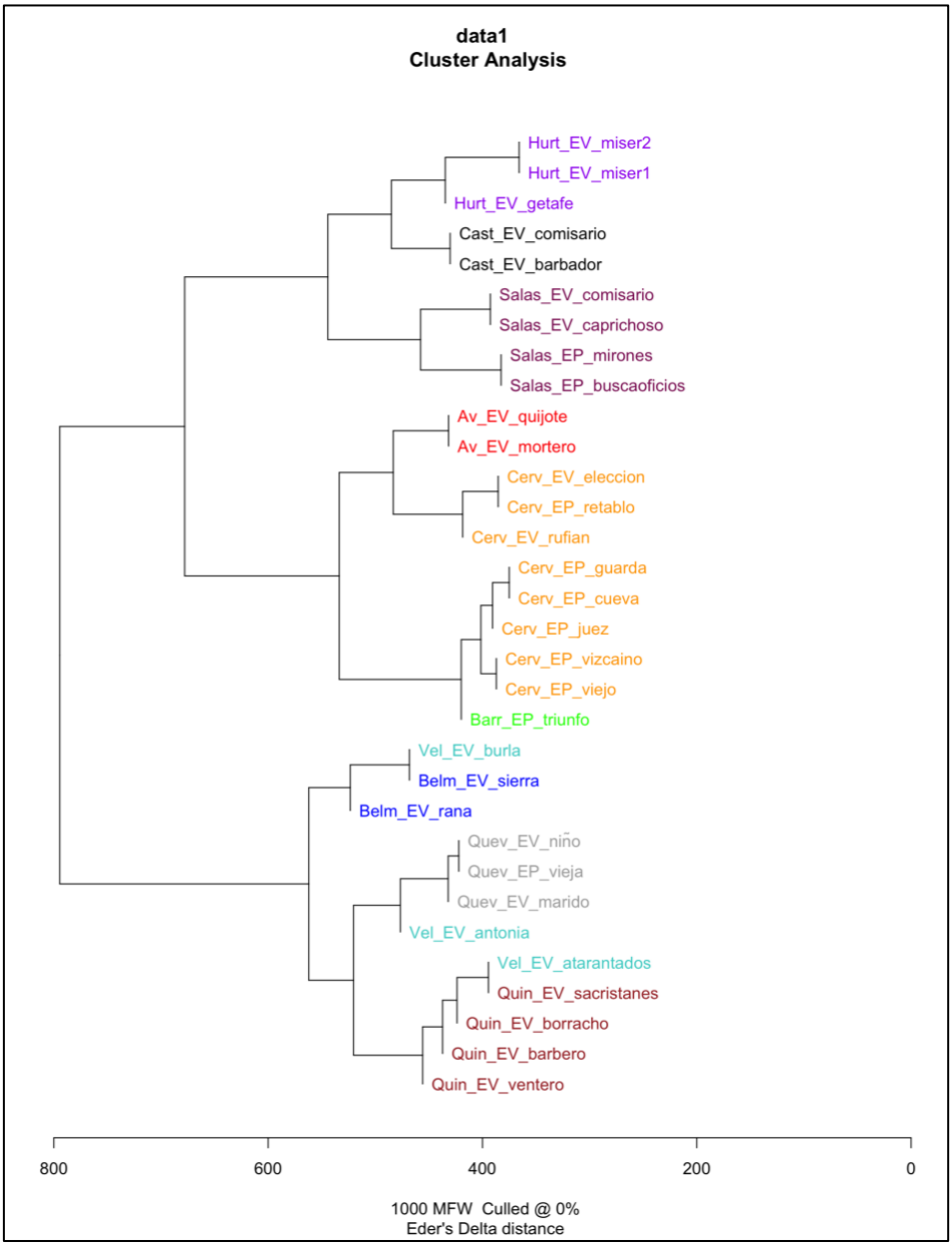

Gráfico 2. Clúster basado en la frecuencia de las palabras más empleadas (RStudio. Stylo. Cluster Analysis. Eder's Delta. MFW 1000. Culled 0%)

Estos dos gráficos nos llevan a las siguientes consideraciones[4]:

a) A pesar de la brevedad de las muestras analizadas, el algoritmo Eder's Delta presenta un resultado de agrupación altamente fiable (tanto con 800 como con 1000 MFW). En un primer nivel de agrupación, se distribuyen las obras en dos grandes ramas: por un lado, las de Hurtado de Mendoza, Castillo de Solórzano, Salas Barbadillo, Ávila, Cervantes y Barrionuevo; y, por otro, las de Vélez de Guevara, Belmonte, Quevedo y Quiñones. En los siguientes niveles de agrupación, todos los textos de autor conocido se han identificado con otros textos del mismo autor, a excepción de los tres entremeses de Vélez de Guevara, que no se asocian nunca directamente entre sí. Este resultado confiere al análisis de la distribución de la frecuencia relativa de las palabras más empleadas en los textos un porcentaje de acierto del 100% en primer nivel y superior al 90% en el resto de niveles (29 de 32 asociaciones).

b) No podemos profundizar en las causas que ocasionan que los tres entremeses de Vélez de Guevara no se asocien entre sí. Es cierto que son tres de los textos más breves (1094, 1144 y 1440 palabras) y que, como ya hemos explicado, cuanto menor es la extensión, menor es también la fiabilidad de los resultados obtenidos; pero también es cierto que otros textos de similar extensión sí se han asociado correctamente (véanse, por ejemplo, *El barbero* de Quiñones, con 986 palabras, los dos textos de Belmonte, con 871 y 1012 palabras, respectivamente, y los dos textos de Castillo de Solórzano, con 1294 y 1422). Las causas, posiblemente, estén relacionadas con la baja calidad de las muestras (contaminación autorial y/o textual) o con las prácticas pseudoplagiarias de Vélez[5], que implicarían, en consecuencia, una baja señal autorial.

[4] Para poder interpretar correctamente los resultados es importante advertir que el 0 indica la identidad total y que, por tanto, cuanto más alejada del 0 se produzca la asociación entre dos textos dados, menos significativa será dicha asociación (o, dicho de otro modo, las asociaciones serán más sólidas y fiables cuanto más a la derecha se produzcan); asimismo, es importante tener en cuenta que, si hay varias subramas dentro de una misma rama, las que están situadas en el centro serán más similares entre sí que las dos de los extremos (arriba y abajo).

[5] *La burla más sazonada, Los atarantados* y *Antonia y Perales* fueron publicados con el nombre de Vélez de Guevara en la *Flor de entremeses y sainetes de diferentes autores* (Madrid, 1657). Menéndez Pelayo es el primero en desconfiar de los "nombres ilustres" que figuran en esta colección (1903: VII) y Urzáiz Tortajada, por su parte, nos recuerda que Vélez se valió frecuentemente de títulos, argumentos y versos de autores ajenos (1997).

c) Todos los entremeses de Cervantes parten de un tronco común, si bien se advierten dos ramas diferenciadas: una compuesta por todos los entremeses en prosa, a excepción de *El retablo de las maravillas*, y otra compuesta por este entremés y los dos en verso. Aunque necesitaríamos realizar un estudio mucho más exhaustivo para establecer la posible causa de esta diferenciación, nos aventuramos a presuponer que está relacionada con dos épocas o momentos en la redacción de los entremeses[6].

d) *El triunfo de los coches*, de Barrionuevo, no se asocia directamente con ningún entremés (algo esperado, pues es el único entremés conservado de este autor)[7], pero muestra cierta proximidad –es decir, cierta identificación léxica– con una de las dos subramas cervantinas (la integrada por *La guarda cuidadosa*, *La cueva de Salamanca*, *El juez de los divorcios*, *El vizcaíno fingido* y *El viejo celoso*). El hecho de no poseer otro entremés indubitado de este autor nos imposibilita saber si la cercanía intraescritor (entre dos entremeses de Barrionuevo) sería mayor que la que detectamos interescritor (entre el entremés de Barrionuevo y varios de los entremeses de Cervantes) (*nearest neighbor problem*).

e) Los dos entremeses de Ávila se asocian entre sí, aunque guardan también cierta cercanía léxica con la subrama cervantina formada por *La elección de los alcaldes de Daganzo*, *El retablo de las maravillas* y *El rufián viudo llamado Trampagos*.

f) Parece haber mayor relación de proximidad entre los textos escritos en verso, por un lado, y los escritos en prosa, por otro; así ocurre en los entremeses de Salas Barbadillo y en la mayoría de los de Cervantes (pero no en los de Quevedo).

[6] Son muchos los autores que han expuesto su teoría sobre la fecha de composición de los ocho entremeses publicados en 1615, aunque ninguno lo ha hecho basándose en análisis estilométricos. La gran mayoría cree que los dos entremeses en verso fueron los de composición más tardía, en consonancia con la evolución general del género; otros, sin embargo, como Cotarelo Valledor, se decantan por el orden en que están impresos. Creemos que sería interesante utilizar análisis semiautomáticos para ofrecer una teoría sobre este asunto, basada en la evolución del idiolecto de una persona a lo largo del tiempo.

[7] *El triunfo de los coches* es el único entremés indubitado de Gaspar de Barrionuevo. Algunos le han atribuido *Los habladores*, por su estrecha relación con Toledo, o *El toquero*, pero no hay unanimidad al respecto (Huerta Calvo, coord., 2008: 214).

A pesar de las dificultades metodológicas de estudiar la autoría de unas piezas literarias tan breves y volátiles como son los entremeses barrocos, el análisis basado en el estudio de la variación de las palabras más frecuentes muestra una fiabilidad del 100% en un primer nivel y de más del 90% en los siguientes niveles, sin variaciones significativas entre las 800 y las 1000 MFW.

1. 4. Análisis de la distribución de la frecuencia relativa de caracteres (*Delta distance,* 5-gramas de caracteres)

Otra variante utilizada en los estudios de atribución de autoría es la frecuencia de n-gramas de grafemas o de caracteres[8], aunque lo cierto es que su fiabilidad no siempre alcanza los mismos niveles de identificación en español que en inglés. Tal y como determinamos en un estudio de hace unos años, "la fiabilidad de los perfiles de n-gramas de grafemas [en español] crece conforme aumenta el número de grafemas analizados (23% para 2, 30% para 3, 36% para 4 y 50% para 5), algo que contrasta con lo que ocurría en lengua inglesa, en donde el perfil más fiable era 2-gram profile (79%) y el menos 9-gram profile (18%)" (Blasco y Ruiz Urbón, 2009: 43).

De un tiempo a esta parte se tiende a trabajar con caracteres en lugar de grafemas, pues son capaces de capturar matices a nivel léxico, sintáctico y estructural (Houvardas y Stamatatos, 2006). En los dos siguientes gráficos mostramos los dendrogramas resultantes de aplicar la distancia Delta al estudio de la distribución de la frecuencia relativa de los 5-gramas de caracteres en nuestro corpus de entremeses indubitados, primero con 800 (Gráfico 3) y después con 1000 MFW (Gráfico 4), seleccionando, al igual que en el apartado anterior, el idioma español, el algoritmo mejorado de Eder (*Eder's Delta*) y el total de secuencias de 5 caracteres (0% *culled*). En esta ocasión, hemos optado por eliminar las unidades gráficas que se corresponden con signos de puntuación (coma, punto, punto y coma, dos puntos, paréntesis, etc.), ya que estos signos son marcas de los editores modernos y, en caso alguno, de los autores primigenios.

[8] Siguiendo a Grieve (2007), usamos *grafemas* para referirnos sólo a las letras del alfabeto y *carácter* para cualquier tipo de unidad gráfica (letras, dígitos, espacios en blanco y signos de puntuación).

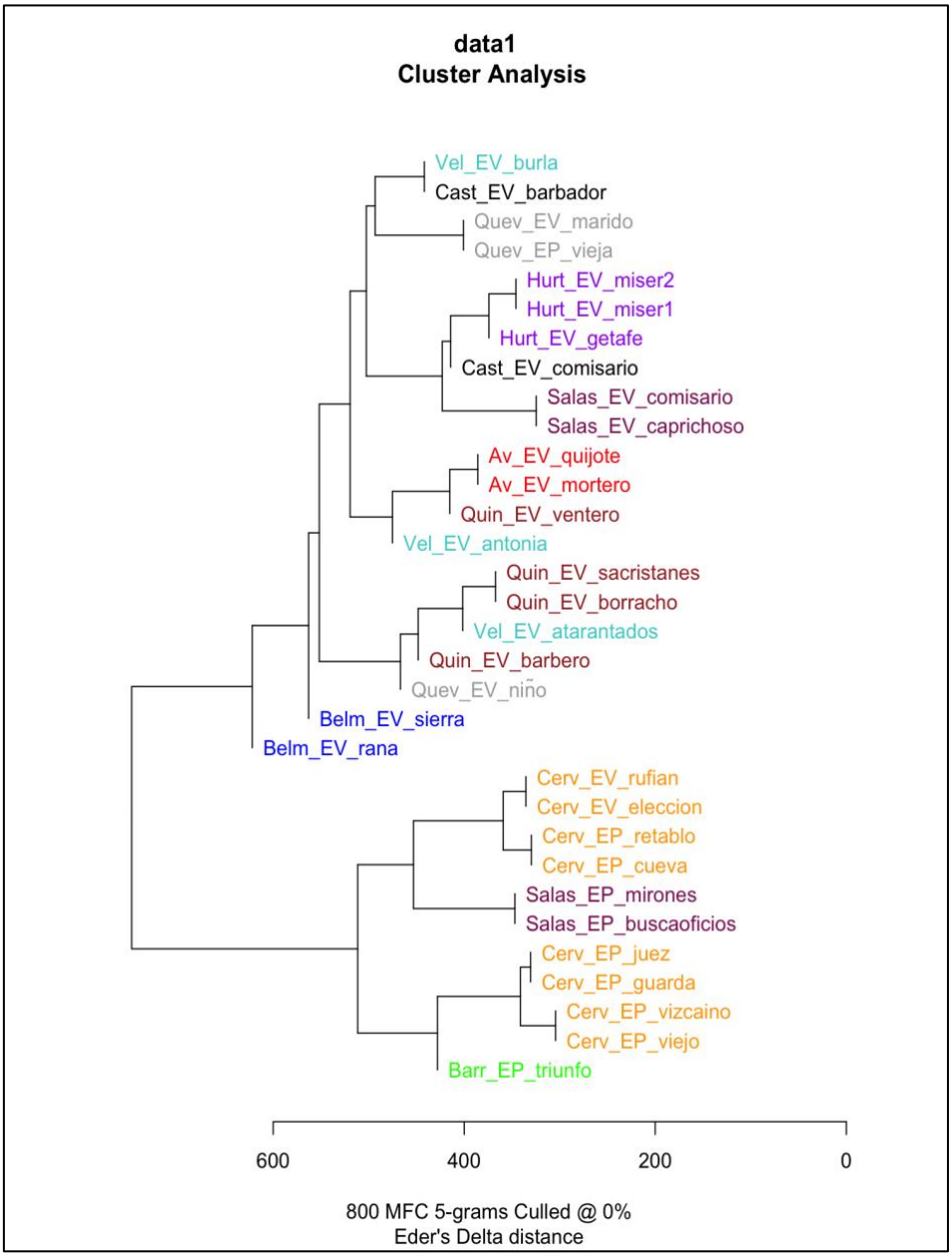

Gráfico 3. Clúster basado en la frecuencia de los 5-gramas de caracteres más empleados
(RStudio. Stylo. Cluster Analysis. Eder's Delta. MFC 800 5-grams. Culled 0%)

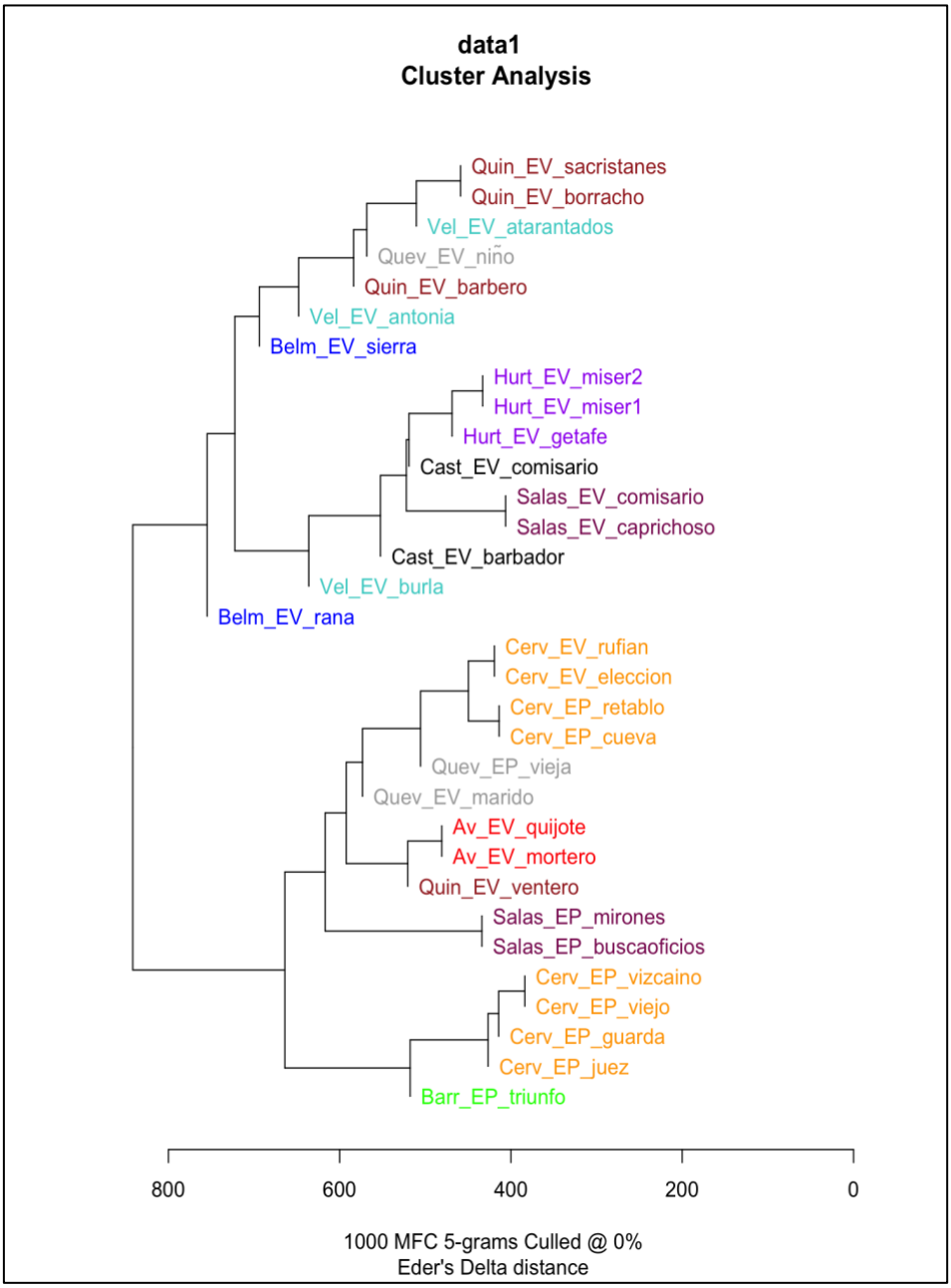

Gráfico 4. Clúster basado en la frecuencia de 5-gramas de caracteres más empleados (RStudio. Stylo. Cluster Analysis. Eder's Delta. MFC 1000 5-grams. Culled 0%)

Dejando a un lado el caso de Vélez de Guevara, el resultado vuelve a ser altamente satisfactorio: todas las obras que en el último nivel se asocian directamente con otra son del mismo autor. Sorprende, sin embargo, que los cuatro entremeses de Salas Barbadillo se distribuyan por pares en dos ramas distintas (prosa *vs* verso) o que alguno de los entremeses de Quevedo, Quiñones de Benavente o Castillo de Solórzano parezcan estar más próximos a los de otros autores que a los de su verdadero autor.

En todo caso, y aunque con un porcentaje de efectividad algo más bajo que con 1-grama de palabra, el estudio de la frecuencia de 5-gramas de caracteres también resulta bastante efectivo para discriminar autorías en los entremeses del Siglo de Oro.

1. 5. Análisis de secuencias de palabras iguales y en el mismo orden (*verbatim*)

El número de palabras iguales y en el mismo orden (*verbatim*) que pueden producir dos escritores distintos se reduce drásticamente en función de la longitud de la frase (Coulthard, 2004). En español se tiene como referencia el estudio realizado por Marquina y Queralt (2014), en el que se concluye que es poco probable que dos personas distintas produzcan de forma independiente secuencias idénticas de cinco o más palabras, siempre y cuando dichas secuencias no sean una expresión común de la lengua, compartida por multitud de hablantes. En consecuencia, la detección de casos significativos de *verbatim* se ha convertido en una prueba indiscutible de autoría (o plagio) en el ámbito forense.

Es evidente que cuanto mayor sea la longitud de los textos, más significativo resultará encontrar o no cadenas de palabras iguales y en el mismo orden; pero es importante advertir que los casos significativos de *verbatim* pueden darse también en textos muy breves.

En todos los entremeses analizados, hay un larguísimo listado de *verbatim* de dos ítems, que son compartidos por la mayoría de autores ("por qué", "para que", "las fiestas", "la causa", "los hombres", "mi casa", "para bailar", "¡Vive Dios!", etc.).

Son también muy frecuentes los *verbatim* de tres ítems, que pueden aparecer en la obra de un mismo autor, como es el caso de Quevedo con "aparécese lleno de" (Quev_EV_marido y Quev_EV_niño) o de Salas Barbadillo con "me admiro que", "que se pierda", "vive el cielo" o "qué

os parece" (Salas_EV_caprichoso y Salas_EV_comisario), pero que, por lo general, son compartidos por distintos autores: "de la vista" (Belm_EV_rana y Salas_EP_mirones), "de mis ojos" (Quev_EV_niño y Vez_EV_atarantados), "de mi casa" (Belm_EV_rana y Quev_EP_vieja), "de mi pluma" (Quev_EV_niño y Salas_EV_caprichoso), "en los corrales" (Belm_EV_rana y Vez_EV_burla), "luego al momento" (Av_EV_quijote, Av_EV_mortero y Barr_EP_triunfo), "de una vez" (Cast_EV_barbador y Salas_EP_mirones), "para que digan" (Quin_EV_sacristanes y Vélez Vez_EV_antonia), "Dios os haga" (Av_EV_quijote y Belm_EV_rana), "un hombre que" (Quev_EV_marido y Vez_EV_burla), "cuerpo de Dios" (Cerv_EP_retablo y Quin_EV_barbero y Quin_EV_ventero), "váyase con Dios" (Cerv_EP_cueva, Cerv_EP_guarda y Barr_EP_triunfo), "pernil de tocino" (Cerv_EP_juez y Av_EV_quijote), "¿hay tal disparate?", (Cerv_EP_guarda y Av_Ev_mortero), etc.

Menor es el número de *verbatim* de cuatro ítems[9]. En estos casos seguimos encontrando secuencias compartidas por distintos autores, como "ay, desdichada de mí" (Cerv_EP_guarda y Vel_EV_burla), "¿qué voces son éstas?" (Cerv_EP_vizcaino y Barr_EP_triunfo)[10], "bueno, por vida mía" (Cerv_EV_rufian, Salas_EP_mirones y Barr_EP_triunfo), "no falta más que" (Cerv_EP_retablo y Av_EV_quijote), "¿qué hemos de hacer?" (Salas_EV_comisario y Vel_EV_atarantados), "¿qué te ha / habemos hecho?" (Av_EV_mortero y Vel_EV_atarantados) o "pero, ¿qué es esto?" (Cerv_EP_juez y Quin_EV_sacristanes), pero también hay ya un número significativamente alto de secuencias de cuatro ítems iguales compartidas en obras de un mismo autor; así, en Ávila encontramos "a manos de un" (Av_EV_mortero y Av_EV_quijote) o "que soy peor que" (Av_EV_mortero y Av_EV_quijote)[11]; en Cervantes, "Dios que va de" (Cerv_EV_rufian y Cerv_EV_eleccion), "no es / será posible que" (Cerv_EP_viejo, Cerv_EP_vizcaino y Cerv_EP_viejo),

[9] No hemos considerado los casos en los que hay una alteración del orden de las palabras, ya que no son propiamente *verbatim*. Por ejemplo: "hallo por mi cuenta" (Salas_EP_buscaoficios) / "por mi cuenta hallo" (Quev_EV_marido), "he de perder el juicio" (Quin_EV_barbero) / "el juicio he de perder" (Cast_EV_comisario) o "ya no lo puedo sufrir" (Cerv_EP_cueva) / "no lo puedo ya sufrir" (Quin_Ev_sacristanes).

[10] En Hurtado de Mendoza encontramos esta secuencia, pero en otro orden: "¿qué voces éstas son?" (Hurt_EV_miser2).

[11] En este caso la similitud va más allá: "que soy peor que Judas si me enojo" (Av_EV_mortero) y "que soy peor que el diablo si me enojo" (Av_EV_quijote).

"aquí como de molde" (Cerv_EP_vizcaino y Cerv_EV_rufian) o "yo así lo creo" (Cerv_EP_cueva y Cerv_EP_viejo); y en Hurtado de Mendoza, "voto a Cristo que" (Hurt_EV_getafe y Hurt_EV_miser2) y "vase y sale él" (Hurt_EV_miser 1 y Hurt_EV_miser 2).

Pero, como señalábamos antes, lo verdaderamente relevante es detectar casos de *verbatim* de cinco o más ítems y comprobar si su uso es idiosincrásico o no, mediante su búsqueda en un corpus de referencia de la época. En la siguiente tabla mostramos los casos detectados en nuestro corpus de entremeses indubitados –tras su rastreo con el programa de análisis textual CopyCatch Gold– y su posible presencia en la base de datos CORDE, entre 1550 y 1650 (Tabla 2). En el resultado del número de concordancias, no contabilizamos los casos en los que CORDE remite a alguno de los entremeses de la columna "TEXTOS", pues lo que tratamos de probar es el grado de rareza de ese *verbatim*; por ello, indicamos el número total de concordancias y, dado el caso, entre paréntesis, el número de ellas que pertenecen a alguno de los autores de los textos en los que hemos detectado ese caso de *verbatim*. Por ejemplo, en "otra vez torno a decir", CORDE sólo detecta una concordancia en esas fechas (además de la del *El viejo celoso* y *El vizcaíno fingido*), y esa concordancia es también de Cervantes (*La ilustre fregona*).

Tabla 2. Estudio de los *verbatim* de 5 o más ítems

VERBATIM	ÍTEMS	TEXTOS	CORDE [12]
a pagar de mi dinero	5	Cast_EV_comisario Vel_EV_antonia	10 (1 Cast)
otra vez torno a decir	5	Cerv_EP_vizcaino Cerv_EP_viejo	1 (1 Cerv)
pueda / puedo ir a la mano	5	Cerv_EV_eleccion Vel_EV_atarantados	16 (2 Cerv)
que me parece a mí	5	Cerv_EV_eleccion Salas_EV_comisario	14
que es lo mismo que	5	Salas_EV_caprichoso Vel_EV_antonia	271 Salas (2)
qué es lo que queréis	5	Cerv_EP_guarda Cerv_EP_retablo	17 (1 Cerv)
quién es aquí el señor	5	Cerv_EP_retablo Cast_EV_barbador	2 (1 Cerv)

[12] Hay que advertir que en las búsquedas en CORDE hemos utilizado comodines ("?" y "*") en aquellos casos en los que puede haber flexión verbal o vacilación ortotipográfica.

vuesa merced, por su vida	5	Cerv_EP_vizcaino Cerv_EP_retablo	0
quién te mete a ti en	6	Cerv_EV_rufian Cerv_EV_eleccion	5 (2 Cerv)
beso a vuesas mercedes las manos	6	Cerv_EP_retablo Cerv_EP_viejo[13]	0[14]
llevar / llevarán un pan como unas nueces	6	Vel_EV_atarantados Vel_EV_burla	1
que se me arranca el alma	6	Cerv_EP_vizcaino Cerv_EP_viejo	1 (1 Cerv)
eso haré yo de muy buena gana, y	8	Cerv_EP_vizcaino Cerv_EP_juez	0
pero, ¿por qué me lo pregunta vuesa merced?	8	Cerv_EP_vizcaino Cerv_EP_guarda	0
yo desde luego te doy mi bendición para que, con la de dios y con ella, salgas bien de todas las aventuras cortesanas	23	Salas_EP_mirones Salas_EP_buscaoficios	0

Fuente: elaboración propia

Tal y como se desprende de la tabla anterior, los *verbatim* de 5 ítems, aunque significativos, no son casi nunca una marca indiscutible de autoría; sí lo son, sin embargo, todos los casos de más de 5 ítems, a excepción de la frase hecha "llevar / llevarán un pan como unas nueces"[15]. De este modo, la localización de casos de *verbatim* de más de 5 ítems puede convertirse en una prueba irrefutable de autoría, válida para textos breves, como los entremeses; su no localización, no obstante, no serviría para negar una autoría común.

CONCLUSIONES

La utilización de metodología cuantitativa en los estudios de atribución de autoría de textos breves sigue siendo un reto, pero este estudio demuestra que el límite de 2000-2500 palabras marcado por

[13] En Salas Barbadillo encontramos esta misma secuencia, pero en otro orden: "Beso las manos a vuesas mercedes" (Salas_EP_buscaoficios).

[14] El *verbatim* "beso a vuesas mercedes las manos" sólo aparece en Cervantes, pero en *La pícara Justina* encontramos "besa a vuesas mercedes las manos".

[15] Este *verbatim* de 6 ítems es bastante significativo, pero no determinante, pues el CORDE identifica su uso en otro texto de la época (*Sucesos de Sevilla de 1592 a 1604*, de Francisco de Ariño). En el *Vocabulario* de Correas aparece "dar un pan como unas nueces" con el significado de dar "palos, golpes y pesadumbres" (1924).

algunos especialistas no es ciertamente determinante. En el caso particular de los entremeses del Siglo de Oro español, y a pesar de las dificultades derivadas de las características propias del género (brevedad, reutilización de temas, personas y escenarios, alternancia de prosa y verso) y de su compleja transmisión (contaminación autorial y textual), los análisis realizados han resultado ser significativamente efectivos.

El estudio de la distribución de la frecuencia relativa de las palabras más empleadas en los textos muestra una fiabilidad superior al 90% en el último nivel, sin variaciones significativas entre las 800 y las 1000 palabras más frecuentes. Cabe señalar, asimismo, que esa falta de fiabilidad no está relacionada con la longitud de los textos, ya que los tres errores detectados afectan a Vélez de Guevara, cuyos entremeses no son los más breves del corpus. Mención especial merece el caso de Barrionuevo, del que sólo disponemos de un entremés, que se asocia (aunque no directamente) con los de Cervantes, en lo que se ha venido a llamar el "problema del vecino más cercano" (*nearest neighbor problem*).

El análisis de la distribución de la frecuencia relativa de las 800 y 1000 cadenas de 5 caracteres más utilizados en los textos –obviando los signos de puntuación– también ofrece resultados fiables para discriminar autorías, aunque en menor grado que con palabras independientes; cabe señalar, no obstante, que la causa de esta menor efectividad no reside en la extensión de los textos, ya que la experiencia nos demuestra que ocurre de igual modo con textos mucho más largos. Aun así, todas las asociaciones directas –a excepción del caso particular de Vélez de Guevara– se producen entre obras de un mismo autor.

El estudio de los *verbatim* demuestra que –como establece la Lingüística forense– los casos compartidos de más de 5 ítems sólo se dan entre obras de unos mismos autores. Hemos detectado secuencias de 6 palabras iguales y en el mismo entre obras de Vélez de Guevara y de Cervantes; de 8 entre obras de Cervantes; y hasta de 23 entre obras de Salas Barbadillo.

Estos tres análisis resultan por tanto efectivos para los estudios de atribución de autoría de textos breves (y, en particular, de los entremeses del Siglo de Oro español), aunque con ciertas limitaciones. Es cierto que la longitud de un texto puede condicionar la fiabilidad que conferimos a un resultado, pero también hay otros factores que son incluso más determinantes; como, por ejemplo, la baja señal autorial (Vélez de Guevara) o la errónea propuesta de candidatos a autor (Barrionuevo).

Este trabajo despeja dudas, pero crea también nuevas incógnitas. A partir de estos resultados, sería interesante revisar la autoría de los tres entremeses de Vélez de Guevara; certificar que *El triunfo de los coches* no es de Cervantes, sino simplemente que Cervantes es el autor que más se aproxima a Barrionuevo de todo el corpus contemplado, mediante técnicas de verificación de autoría como *General Imposters*, del paquete Stylo; o estudiar si las diferencias entre los ocho entremeses cervantinos se corresponden con dos momentos distintos de escritura (variación intraescritor).

FINANCIACIÓN
Esta investigación no recibió ninguna financiación externa.

BIBLIOGRAFÍA

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

Argamon, Shlomo (2008), "Interpreting Burrows's Delta: Geometric and probabilistic foundations", *Literary and Linguistic Computing*, 23 (2), pp. 131-147.

Asensio, Eugenio (1971), *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.

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

Blasco, Javier (2019a), "Atribuciones cervantinas desde la estilometría. *El entremés de Los mirones*", en Guillermo Laín Corona, Rocío Santiago Nogales y José Romera Castillo (coords.), *Cartografía teatral en homenaje al profesor José Romera Castillo*, Madrid, Visor Libros, pp. 151-168.

Blasco, Javier (2019b), "La graciosa y gratuita disputa sobre la autoría de la *Historia verdadera del inconfundible Bernal Díaz del Castillo*", *Boletín de la Real Academia Española (BRAE)*, 99 (319), pp. 5-44.

Blasco, Javier (2022), "La «boutade» de la muerte del autor: El caso de Carmen Mola", *Anales de literatura española contemporánea (ALEC)*, 47 (3), pp. 249-266.

Blasco, Javier y Cristina Ruiz Urbón (2009), "Evaluación y cuantificación de algunas técnicas de 'Atribución de autoría' en textos españoles", *Castilla. Estudios de literatura*, 0, pp. 27-47.

Burrows, John Frederick (2002), "«Delta»: A measure of stylistic difference and a guide to likely authorship", *Literary and Linguistic Computing*, 17 (3), pp. 267-287.

Calvo Tello, José (2016), "Entendiendo Delta desde las Humanidades", *Caracteres. Estudios culturales y críticos de la esfera digital*, 5 (1), pp. 140-176.

Campión Larumbe, Miguel y Álvaro Cuéllar (2021), "Discernir entre original y refundición en el teatro del Siglo de Oro a través de la estilometría: el caso de *El mejor amigo, el muerto*", *Talía. Revista de estudios teatrales*, 3, pp. 59-69.

Cerezo Soler, Juan y José Calvo Tello (2019), "Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de *La conquista de Jerusalén*", *Anales Cervantinos*, 51, pp. 231-250.

Correas, Gonzalo (1924), *Vocabulario de refranes y frases proverbiales y otras fórmulas comunes de la lengua castellana*, Madrid, Tipografía de la Revista de Archivos, Bibliotecas y Museos.

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

Coulthard, Malcolm (2004), "Author identification, idiolect and linguistic uniqueness", *Applied Linguistics*, 25 (4), pp. 431-447.

Coulthard, Malcolm (2005), "The Linguistic as Expert Witness", *Linguistics and the Human Sciences*, 1 (1), pp. 39-58.

Eder, Maciej (2013), "Bootstrapping Delta: A safety net in open-set authorship attribution", *Digital Humanities 2013. Conference abstracts. University of Nebraska/Lincoln, USA. 16-19 July 2013*, pp. 169-172.

Eder, Maciej (2015), "Does size matter? Authorship attribution, small samples, big problem", *Digital Scholarship in the Humanities*, 30 (2), pp. 167-182.

Eder, Maciej (2017), "Short samples in authorship attribution: A new approach", *Digital Humanities 2017: Conference abstracts.* Montreal, Canada, pp. 221-224. Disponible en: https://dh2017.adho.org/abstracts/341/341.pdf (fecha de consulta: 28/11/2022).

Eder, Maciej, Mike Kestemont y Jan Rybicki (2016), "Stylometry with R: A Package for Computational Text Analysis", *The R Journal*, 8 (1), pp. 107-121.

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

Fradejas Rueda, José Manuel (2019), "Estilometría y Edad Media castellana", *Romanische Studien*, 6, pp. 49-74.

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

Graeme, Hirst y Ol'ga Feiguina (2007), "Authorship attribution for small texts: Literary and forensic experiments", *International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, 30th Annual Internacional ACM SIGIR Conference* (SIGIR '07), Amsterdam, Netherlands. Disponible en: http://ceur-ws.org/Vol-276/paper3.pdf (fecha de consulta: 28.11.2022).

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

Hernández Lorenzo, Laura (2019a), *Los textos poéticos de Fernando de Herrera: Aproximaciones desde la Estilística de corpus y la Estilometría*, tesis doctoral, Universidad de Sevilla. Disponible en: https://idus.us.es/handle/11441/93465 (fecha de consulta: 28.11.2022).

Hernández-Lorenzo, Laura (2019b), "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), pp. 189-228.

Hoover, David L. (2004), "Testing Burrows's Delta", *Literary and Linguistic Computing*, 19 (4), pp. 453-475.

Houvardas, John y Efstathios Stamatatos (2006), "N-gram feature selection for authorship identification", en. J. Euzenat y J. Domingue (eds.), *Proceedings of Artificial Intelligence: Methodologies, Systems, and Applications*, Springer Verlag, pp. 77-86.

Huerta Calvo, Javier (coord.) (2008), *Historia del teatro breve en España*, Madrid/Frankfurt am Main, Iberoamericana/Vervuert.

Koppel, Moshe, Jonatan Schler y Elisheva Bonchek-Dokow (2007), "Measuring differentiability: Unmasking pseudonymous authors", *Journal of Machine Learning Research*, 8, pp. 1261-1276.

Luyckx, Kim (2010), *Scalability Issues in Authorship Attribution*, Brussels, University Press Antwerp.

Marquina, Montse y Sheila Queralt (2014), "Similarity threshold to detect plagiarism in Spanish", *RAEL: revista electrónica de lingüística* aplicada, 13 (1), pp. 79-95.

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

Rißler-Pipka, Nanete (2016), "Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales»", en *El otro Quijote. La continuación de Avellaneda y sus efectos*, Hanno Ehrlicher (ed.), mesa redonda. Augsburgo: Universität Augsburg, pp. 27-51.

Rojas Castro, Antonio (2017), "Luis de Góngora y la fábula mitológica del Siglo de Oro: Clasificación de textos y análisis léxico con métodos informáticos", *Studia Aurea*, 11, pp. 111-142.

Rosa, Javier de la y Juan Luis Suárez (2016), "The Life of *Lazarillo de Tormes* and of his Machine Learning Adversities. Non-Traditional Authorship Attribution Techniques in the Context of the Lazarillo", *Lemir*. 20, pp. 373-438.

Rybicki, Jan y Maciej Eder (2011), "Deeper Delta across genres and languages: Do we really need the most frequent words?", *Literary and Linguistic Computing*, 26 (3), pp. 315-321.

Sanderson, Conrad y Simon Guenter (2006), "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation", *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australia, pp. 482-491.

Smith, Peter W. H. y W. Aldridge (2011), "Improving authorship attribution: Optimizing Burrows' delta method", *Journal of Quantitative Linguistics*, 18 (1), pp. 63-88.

Statamatos, Efstathios (2009), "A Survey of Modern Authorship Attribution Methods", *Journal of the American Society for Information Science and Technology*, 60 (3), pp. 538-556.

Urzáiz Tortajada, Héctor (1997), "Un entremés olvidado de Luis Vélez de Guevara: Los atarantados", *Criticón*, 71, pp. 127-157.

Vega García-Luengos, Germán (2021), "Las comedias de Lope de Vega: confirmaciones de autoría y nuevas atribuciones desde la estilometría", *Talía. Revista de estudios teatrales,* 3, pp. 91-108.

Wrisley, David Joseph (2016), "Modeling the Transmission of al-Mubashshir Ibn Fātik's Mukhtār al-Ḥikam in Medieval Europe: Some Initial Data-Driven Explorations", *Journal of Religion, Media and Digital Culture*, 5 (1), pp. 228-257.