DHQ: Digital Humanities Quarterly
2018
Volume 12 Number 1

La conquista de Jerusalén ¿de Cervantes? Análisis estilométrico sobre autoría en el teatro del Siglo de Oro español [en]

José Calvo Tello <jose_dot_calvo_at_uni-wuerzburg_dot_de>, Universidad de Würzburg
Juan Cerezo Soler <juan_dot_cerezosoler_at_gmail_dot_com>, Universidad Autónoma de Madrid
Translation: Jose Calvo Tello <jose_dot_calvo_at_uni-wuerzburg_dot_de>, Universidad de Würzburg

Abstract

En este artículo aplicamos métodos estilométricos para abordar el problema de autoría de la comedia *La conquista de Jerusalén*, atribuida desde su descubrimiento a Miguel de Cervantes. Para ello hemos realizado numerosos análisis con diferentes rangos de palabras, en un total de diecisiete textos teatrales escritos, todos ellos, por los siete autores que conforman la generación teatral de 1580 y cuya actividad dramática coincide en el tiempo con la composición de *La conquista*. Hemos utilizado la unidad de distancia textual Delta para agrupar (*cluster*) los textos.

1. Introducción

La estilometría[1] es una de las metodologías de análisis cuantitativo textual más frecuentemente utilizada en las Humanidades Digitales. Su principal utilización ha sido la atribución de autoría (tanto para textos literarios como para no literarios), para lo que se ha demostrado notablemente sólida. Como ejemplos de ello se pueden citar los trabajos realizados para corpus literarios en inglés [Burrows 2002], alemán [Jannidis y Lauer 2014], francés [Schöch 2014], latín, húngaro, polaco [Rybicki y Eder 2011] u holandés [van Dalen-Oskam y van Zundert 2007]. 1

Para textos en español, se han aplicado métodos estilométricos[2] sobre la homogeneidad textual de *La Celestina* [Bernaldo de Quirós Mateo 2011] o sobre la autoría cervantina de *La tía fingida* [López 2011]. Sin embargo en nuestra percepción la estilometría se ha aplicado con menor frecuencia para textos literarios en español que para otras lenguas europeas. Esto se observa tanto en la pequeña cantidad de trabajos estilométricos que abordan un problema autorial, así como en aquellos trabajos que comprueban la validez de métodos y parámetros con corpus de diferentes lenguas (por ejemplo: [Eder 2012], [Eder 2013a], o [Jannidis et al. 2015]. Una de las posibles razones de su menor implantación en el ámbito hispánico es la falta de repertorios generales de ediciones escolares en formato electrónico en XML-TEI; la mayoría de proyectos prefieren publicar sus ediciones en HTML o PDF. A pesar de esto, recientemente se han presentado en el contexto internacional diferentes trabajos de estilometría con textos en español ([Calvo Tello et al. 2015]; [Rißler-Pipka 2016]; [Wrisley 2016][3]). Con este trabajo esperamos contribuir a la difusión y a la implantación de este método de Humanidades Digitales también en el ámbito hispánico. Su punto de inicio fue la publicación por parte de la colección *Clásicos Hispánicos*, dirigida por Pablo Jauralde Pou, de *La conquista de Jerusalén*, atribuida a Miguel de Cervantes, cuya edición ha sido preparada por Rodríguez López-Vázquez. Esta fue publicada en formato ePUB derivado de un archivo XML-TEI. 2

En la siguiente sección de este artículo comentaremos la discusión filológica sobre la autoría del texto hasta el día de hoy. Posteriormente pasaremos a detallar los diferentes aspectos metodológicos de este estudio: tanto la composición del corpus así como el método y parámetros utilizados (y la dificultad de su asignación para el español). A continuación comentaremos los resultados y la manera de validación utilizada. Finalmente llegaremos a ciertas conclusiones relacionadas tanto con el texto como con los métodos estilométricos utilizados en español. 3

2. Discusión sobre la autoría de *La conquista de Jerusalén*

La inclusión de *La conquista de Jerusalén* en el conjunto de las obras de Cervantes nace en el mismo momento de su descubrimiento en 1992 a cargo del investigador italiano, ya fallecido, Stefano Arata. Proponía en su primera publicación [Arata 1992] a Miguel de Cervantes como posible autor, considerando que *La conquista de Jerusalén* debía completar el corpus teatral cervantino hallado hasta la fecha[4]. Para respaldar su propuesta, el hispanista desplegó una primera batería de argumentos que afectaban a varios aspectos del texto: en el ámbito formal señaló numerosas semejanzas métricas, así como varias similitudes en la forma de configuración del reparto y en el tipo de acotaciones utilizadas; en cuanto al contenido, apuntó el comportamiento profundamente cervantino de las figuras alegórico-morales [Arata 1992]. Todo lo publicado por Arata ha servido de base para otros estudios que, en su mayoría, han venido a adherirse por distintas vías a la hipótesis cervantina. 4

Especialmente convincentes en este sentido resultan los trabajos de Héctor Brioso, estudios que llegan a su culmen con una edición de la comedia en 2009. En ella, amén de un detalladísimo estado de la cuestión crítica sobre la obra, se proponen nuevos datos que vinculan el nombre del complutense con la pieza custodiada en la Real Biblioteca. También el profesor José Montero Reguera ha defendido en varios lugares la primera hipótesis lanzada por el hispanista italiano y ha reforzado los argumentos existentes incidiendo particularmente en la proximidad de varios versos de la obra anónima con otros de Cervantes [Montero Reguera 1995-1997]. Se les añade, en los últimos años, la aportación de Aaron M. Kahn, que defiende la autoría cervantina a través de la lectura ideológica de *La conquista de Jerusalén,* con clave en el enfrentamiento entre la corona española y el islam, centrándose sobre todo en las figuras alegóricas que pueblan tanto la obra anónima hallada en Palacio como las cervantinas [Kahn 2010]. Moisés R. Castillo conecta hábilmente un gran número de “aspectos temáticos, dramáticos, e ideológicos” de *La conquista* con varias comedias 5

cervantinas, concretamente con las agrupadas bajo el marbete genérico de “comedias de cautivos” [Castillo 2012].

En cuanto a las similitudes halladas en el plano lingüístico entre la comedia de Palacio y la literatura de Cervantes, se ha intentado rastrear, a través del CORDE, la frecuencia con que determinados usos de *La conquista de Jerusalén* aparecen en otros autores de la generación teatral de 1580 [Rodríguez López-Vázquez 2011]. Los resultados, aunque no sean concluyentes, señalan a Cervantes como el autor más probable, con 31 coincidencias, frente a las 10 de Virués y las 8 de Juan de la Cueva y de Lobo Lasso. También se han visto fuertes convergencias en el manejo de determinadas expresiones literarias, tanto que se podría hablar sin problema de reescrituras y transliteraciones, pues Cervantes –como el resto de autores– aprovechaba, revisaba y reciclaba constantemente varios de sus pasajes y versos. Esta proximidad del material literario resulta esclarecedora al cotejar *La conquista de Jerusalén* con *La Numancia* [Baras Escolá 2010, 80]. 6

Las últimas aportaciones críticas favorables a la atribución cervantina inciden en varios aspectos de la construcción dramática, tales como la configuración de un personaje colectivo en *La conquista de Jerusalén* y cuya composición se conecta fácilmente con el quehacer teatral de Cervantes previo a Lope de Vega [Cerezo Soler 2014]. Por su parte, Fausta Antonucci aborda la atribución fijando el ojo crítico en el análisis comparado de la estructura dramática de *La conquista*, en lógica relación con *La Numancia* [Antonucci 2014] y *El trato de Argel* [Antonucci 2015]. Todo ello aparece cumplidamente recogido en la última edición de la obra a cargo de Alfredo Rodríguez López-Vázquez (2014) en la ya mencionada colección Clásicos Hispánicos. En ella se ofrecen nuevos análisis lingüísticos como refuerzo a la hipótesis cervantina y, al tiempo, se sugieren algunas enmiendas al texto fijado por Brioso Santos en 2009 [Brioso Santos 2009]. 7

Con todo, a pesar de la abundante cantidad y calidad de las aportaciones críticas sobre *La conquista de Jerusalén*, la propuesta de autoría no ha rebasado nunca el terreno de la hipótesis ni se han comparado elementos de manera coherente entre *La conquista* y el resto de obras de todos los autores posibles. Más de dos décadas de estudios sobre la comedia han conseguido que la atribución a Miguel de Cervantes sea aceptada por la mayor parte de la comunidad investigadora, más por la elocuencia y solidez parcial de los argumentos presentados que por un estudio realizado con una metodología homogénea, que reconozca la autoría de otros textos y que esté contrastado con el trabajo de otros investigadores en otras tradiciones y lenguas. Aunque nuestro análisis no sea una prueba definitiva de que Miguel de Cervantes sea el autor de *La conquista de Jerusalén*, nuestra metodología es sólida y reproducible por otros investigadores. 8

3. Preparación del corpus y metodología

En esta sección especificaremos algunos aspectos de la preparación electrónica de los textos, así como la selección de los parámetros que hemos utilizado para el trabajo. 9

3.1. Diseño y preparación del corpus de textos

Aunque pueda resultar obvio, consideramos útil recordar que para aplicar una metodología cuantitativa textual es necesario disponer de los textos.[5] De los diecisiete textos que forman el corpus de este trabajo, solo conseguimos localizar un texto en XML-TEI publicado en Internet, por lo que la creación del corpus ha representado una gran cantidad de tiempo y esfuerzo. 10

Somos conscientes de que los métodos estilométricos no solo dan información sobre el autor; aunque este sea la llamada *señal más fuerte* (*strongest signal*), hay otros aspectos que se representan en los resultados como el sexo del autor [Argamon et al. 2003] [Argamon et al. 2009], el género literario [Kestemont et al. 2012], la época [Jockers 2013] o los temas tratados [Seroussi et al. 2014]. Hemos intentado anular estas señales textuales realizando un cuidado diseño del corpus, tanto en la elección de los textos como en el tratamiento estructural y ortográfico de los textos. Todos los textos utilizados son obras de teatro, en verso y todos los candidatos son hombres. Además, para la selección de los textos hemos seguido los siguientes criterios: 11

- *Tres textos por autor*: cada autor señalado como posible autor de la obra analizada esté representado por tres textos[6]
- *Subgénero*: al ser *La conquista de Jerusalén* una comedia, hemos dado preferencia a aquellas obras que pertenezcan a este mismo subgénero
- *Datación*: hemos dado prioridad a aquellas obras que se presupone que fueron escritas en la misma época que la obra analizada, es decir, alrededor de 1580
- *Digitalización*: comprensiblemente hemos utilizado aquellos textos que encontramos digitalizados a aquellos que no[7]

Tras conseguir el texto en algún formato digital, convertimos cada uno de ellos en XML-TEI siguiendo diferentes estrategias para cada formato.[8] Con estos criterios, los textos que forman nuestro corpus son: 12

Autor	Título	Año	Fuente	Formato	Editor
Leonardo de Argensola	La Isabela	1580-1590	Cervantes Virtual	HTML	
Leonardo de Argensola	La Alejandra	1580-1590	Libro escaneado	OCR	Luigi Giulinani
Rey de Artieda	Los amantes	ca. 1580	Cervantes Virtual	HTML	Teresa Ferrer Valls
Jerónimo Bermúdez	Nise lastimosa	1577	Libro escaneado	OCR	Mitchell D. Triwedi
Jerónimo Bermúdez	Nise laureada	1577	Libro escaneado	OCR	Mitchell D. Triwedi
Miguel de Cervantes	La Numancia	1585	Clásicos Hispánicos	XHTML	Gastón Gilabert
Miguel de Cervantes	Los tratos de Argel	ca. 1580	Cervantes Virtual	HTML	Florencio Sevilla Arroyo
Miguel de Cervantes	El gallardo español	1615	Cervantes Virtual	HTML	Florencio Sevilla Arroyo
Juan de la Cueva	El Saco de Roma	1582	Cervantes Virtual	HTML	
Juan de la Cueva	El infamador	1582	Cervantes Virtual	HTML	
Juan de la Cueva	Príncipe Tirano (I)	1582	TESO	HTML	M. C. Simón Palmer
Lobo Lasso de la Vega	Constantinopla	1587	TESORO	XML-TEI	Alfredo Hermenegildo
Lobo Lasso de la Vega	Dido restaurada	1587	Cervantes Virtual	HTML	
Cristobal de Virués	Semiramis	1609	Comedias.org	HTML	Vern Williamsen
Cristobal de Virués	Atila el furioso	1609	Libro escaneado	OCR	Alfredo Hermenegildo
Cristobal de Virués	Marcela	1609	Biblioteca.org.ar	PDF	
¿Miguel de Cervantes?	La Jerusalem	ca. 1580	Clásicos Hispánicos	XML-TEI	Alfredo Rodríguez López-Vázquez

Figure 1. Tabla resumen del corpus

13 Aquellos textos que provienen de un proceso de OCR (ya sea del escaneado nuestro, como lo fueron cuatro de los textos, o de Google Books) han sido especialmente corregidos de erratas, errores de lectura y otras inconsistencias del proceso. Cada versión electrónica se cotejó con su misma edición impresa. Una vez tuvimos los textos en XML-TEI, se sometió todo el corpus a un proceso de unificación ortográfica con el objetivo de lograr homogeneidad léxica.[9] La razón para hacer esto era anular la diferencia de edición de los textos: los textos modernizados podrían tender a agruparse juntos frente a los modernizados al reconocerse el agrupamiento automático como diferencias léxicas aquello que en realidad son meras diferencias de modernización. Esto podría llevar a malintepretaciones sobre los datos sobre autoría. Dado que todos los textos manejados fueron extraídos de fuentes diversas con diferentes procedimientos de modernización, al unificar el corpus conseguíamos que esa diferencia no se reproduzca en los resultados. Esta unificación se ha llevado a cabo conforme a los siguientes criterios:

- Se han actualizado todos los grupos consonánticos cultos, tales como –ct– (*auctor*), –sc– (*esclarescer*), –pr– (*propia*), –ch– (*christianos*), –ph– (*esphera*), –pt– (*captivo*).
- Se ha corregido según norma actual el uso vacilante de –b– y –v–. De la misma forma, se ha actualizado el uso indiscriminado tanto de –u– con valor consonántico (*tuuieren*), como de –v– con valor vocálico (*avto*).
- Simplificación de reduplicaciones gráficas que no respondan a necesidades ortográficas actuales, tales como –ss– (*tuviesse*), –cc– (*succeso*) y –rr– (*honrra*).
- Sustitución de la grafía –ç– por –c– o –z– según norma ortográfica actual (*coraçón*).
- Se ha sustituido –q– por –c– según precisa la norma ortográfica actual (*quanto*).
- Se ha sustituido –x– por –j– según precisa la norma ortográfica actual (*lexos*).
- Actualización gráfica de nombres propios (*Ynés*, *Portogal*, *Galiçia*, etc.).
- Se han corregido todas las interjecciones exclamativas (*O - Oh; Ai - Ay*).
- Se han respetado las formas contraídas *dello*, *desto* y *aquesto* con el fin de no afectar al comportamiento métrico de la obra.
- Asimismo se ha respetado la forma antigua infinitivo+pronombre (*decilla*, *matallo*).

14 De este corpus en XML-TEI se derivó un corpus análogo en formato texto plano codificado en UTF-8 que contiene exclusivamente los parlamentos pronunciados por los personajes. Es decir, se eliminaron automáticamente los metadatos y el *teiHeader* por completo; paratextos (*front,* listados de personajes, *back* y acotaciones); nombres de personajes (en elemento *speaker*) y encabezamientos (en elemento *head*); etiquetas, comentarios, atributos y valores (como la numeración de los versos) XML. El texto más breve, *El saco de Roma*, de Juan de la Cueva, contiene más de 7000 palabras, por lo que todas las obras se encuentran por encima del mínimo de 5000 palabras señaladas como necesarias por Eder [Eder 2012] para estudios estilométricos de autoría.

3.2. Discusión sobre el método y los parámetros

15 En cuanto a los parámetros concretos, creemos útil señalar que hemos intentado sostener cada uno de los parámetros en estudios empíricos anteriores, aspecto que no siempre ha sido posible. Hemos utilizado la versión clásica de Delta (Burrows 2002; Argamon 2008) para poder comparar mejor nuestros resultados con otros trabajos anteriores.

16 Rybicki y Eder investigan de manera meticulosa en diferentes lenguas europeas el rango de palabras que optimiza los resultados para autoría literaria [Rybicki y Eder 2011]. Aunque sus conclusiones no deben tomarse como definitivas para toda la literatura de esas lenguas[10], representan un punto de partida excelente. Lamentablemente no utilizaron corpus en español.[11] Veamos cuáles son sus resultados resumidos en la siguiente tabla. En ella el eje horizontal señala la cantidad de palabras tenidas en cuenta para el análisis; el eje vertical señala el punto de partida siguiendo el orden de palabras más frecuentes. En la tabla aparecen aquellas lenguas (señaladas por sus dos primeras letras en inglés) que los autores señalan

como segmentos con un reconocimiento de autoría óptima. Hemos colocado colores a cada lengua para que el sea más sencillo reconocerlas[12]:

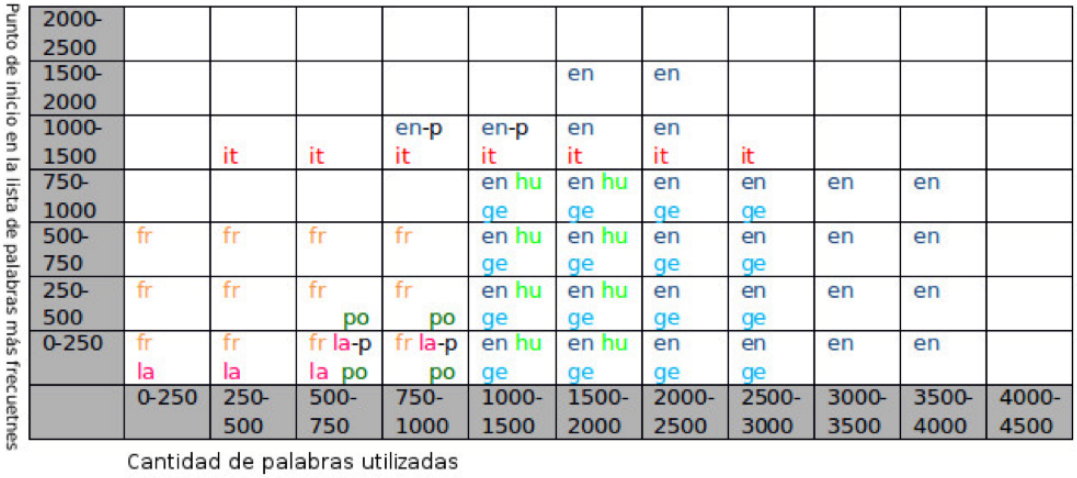

Figure 2. Corpus que obtuvieron mejores resultados en Rybicki y Eder [Rybicki y Eder 2011]

Hay que tener en cuenta que esta tabla es una simplificación a partir del trabajo de Rybicki y Eder [Rybicki y Eder 2011] ya que en ella aparecen agrupados bloques de un mínimo de 250 palabras, mientras que en el trabajo original se puede apreciar las diferencias palabra por palabra. Podemos observar que mientras que hay áreas donde solo una lengua tiene resultados óptimos (como el inglés y el francés en los extremos horizontales), hay otras donde numerosas y diferentes lenguas europeas muestran resultados óptimos. La siguiente tabla ilustra la cantidad de lenguas que muestran resultados óptimos por cada rango de palabras y nos servirá para comparar nuestros resultados[13]: 17

	0-250	250-500	500-750	750-1000	1000-1500	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500
2000-2500	0	0	0	0	0	0	0	0	0	0	0
1500-2000	0	0	0	0	0	1	1	0	0	0	0
1000-1500	0	1	1	2	2	2	2	1	0	0	0
750-1000	0	0	0	0	3	3	2	2	1	1	0
500-750	1	1	1	1	3	3	2	2	1	1	0
250-500	1	1	2	2	3	3	2	2	1	1	0
0-250	2	2	4	3	3	3	2	2	1	1	0

Figure 3. Cantidad de corpus con sus mejores resultados en Rybicki y Eder [Rybicki y Eder 2011]

Para el análisis de los datos electrónicos hasta ahora explicados, hemos utilizado agrupación mediante aprendizaje automático no supervisado (o *clustering*) a través de medidas de distancia textuales y su visualización mediante dendrogramas. El software utilizado ha sido el paquete de R diseñado por Eder, Kestemont y Rybicki llamado *stylo* [Eder et al. 2016].[14] 18

4. Resultados y comparación de resultados

Al no poder saber cuál es el rango de palabras óptimo para el español, hemos decidido probar todos los posibles (en total 77) con una granularidad mínima de 250 (ampliando a 500 una vez pasadas las primeras 1000 palabras más frecuentes) con un rango entre 0 y 4500 en cuanto a la cantidad de palabras, y entre 0 y 2000 en cuanto al punto de partida en la lista de palabras más frecuentes. Por cada uno de estos rangos hemos realizado un dendrograma. Por ejemplo, la siguiente imagen muestra el dendrograma resultante de utilizar las 750 palabras más frecuentes sin eliminar ninguna de las más frecuentes (rango en el que la figura 3 muestra resultados óptimos en cuatro corpus): 19

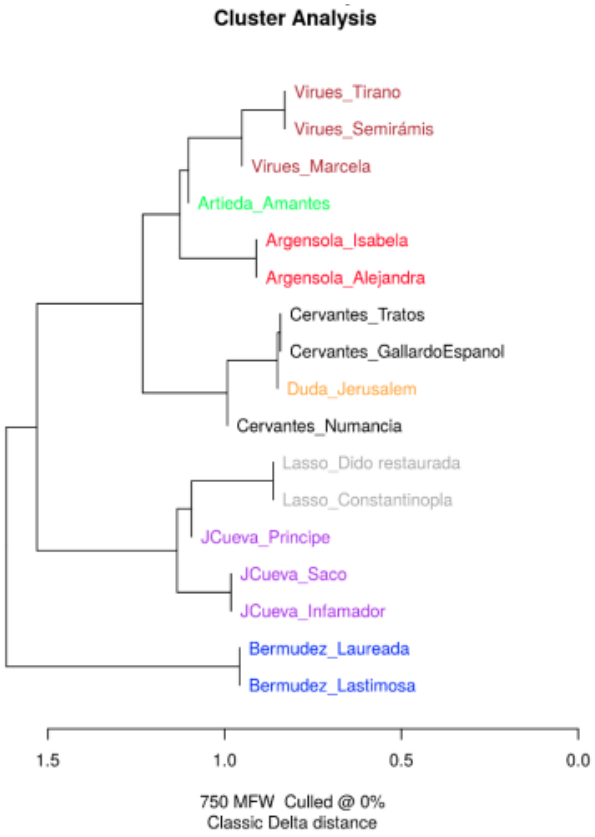

Figure 4. dendrograma realizado de las 750 MFW

Como podemos observar, el texto de *La conquista de Jerusalén* aparece agrupado con el resto de textos de Cervantes. El resto de textos aparecen organizados por autoría con una excepción, *Príncipe* de Juan de la Cueva. Es decir, el método ha cometido un error al intentar reconocer la autoría de uno de los textos por lo que puede estar cometiendo un error similar con la *Jerusalén*. Veamos ahora otro ejemplo de dendograma utilizando otro rango para el que tres de los corpus de la figura 3 mostraban resultados óptimos; en concreto las 2000 palabras más frecuentes habiendo eliminado las primeras 250 palabras más frecuentes[15]:

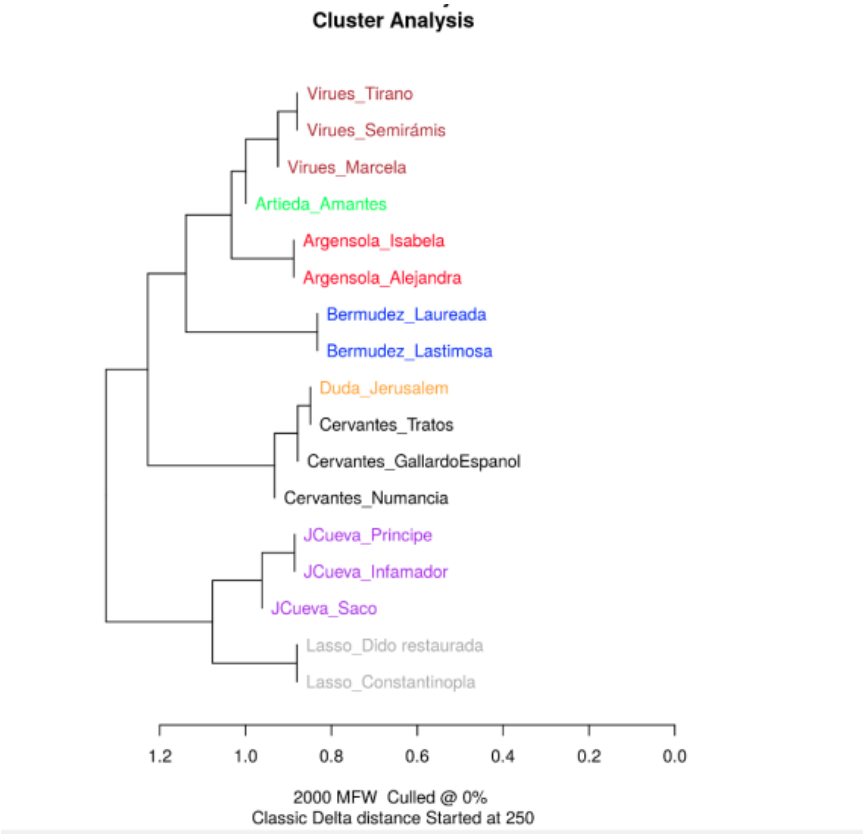

Figure 5. dendograma de las 2000 MFW, habiendo eliminado las 250 primeras

En este caso no solo la *Jerusalén* vuelve a estar agrupada con los textos de Cervantes, sino que se presenta en relación muy cercana a los *Tratos de Argel*, seguramente, por la coincidencia de contenido. Pero lo más importante: el resto de textos aparecen organizados correctamente según sus correspondientes autores. 21

Vemos en estos dendogramas lo que Rybicki y Eder [Rybicki y Eder 2011] ya habían constatado: la diferencia de rangos de palabras modifica los resultados. Una de las soluciones propuestas para este problema es utilizar *bootstrap consensus tree* [Eder 2013b] que aúna en un análisis los resultados de numerosos agrupamientos. En la siguiente imagen se observan los resultados de los nueve dendogramas creados desde las 500 palabras más frecuentes hasta las 4500 aumentando cada vez en 500 palabras: 22

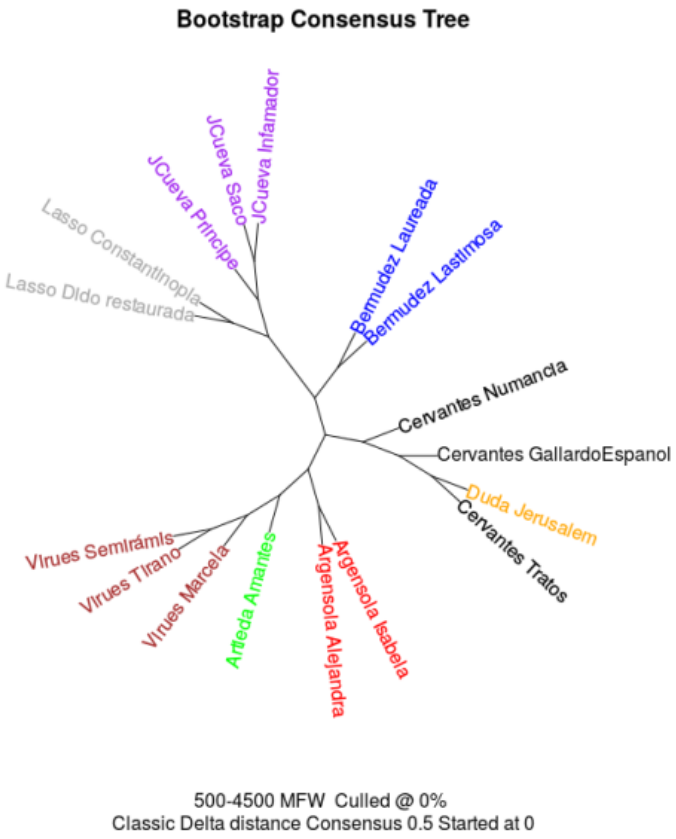

Figure 6. *consensus tree* desde las 500 MFW hasta 4500

23 En la figura 6 volvemos a observar que el texto de *La conquista de Jerusalén* aparece agrupada con el resto de textos de Cervantes.[16] Este mismo análisis se realizó con la variante de Eder de Delta y los resultados son idénticos. La figura 6 proporciona claves para entender la relación de *La conquista* con el resto de obras escritas por el complutense: los resultados nos muestran que la obra cuestionada establece conexiones más o menos cercanas con las otras obras cervantinas, conforme a la siguiente formulación: (((*La Jerusalén*+*Los tratos de Argel*)+*El gallardo español*)+*La Numancia*). Las dos primeras podrían, perfectamente, guardar relación tanto cronológica (fueron compuestas en fechas muy cercanas) como temática (ambas desarrollan su argumento en un contexto de enfrentamiento religioso contra el islam). Por su lado, *El gallardo español* compartiría con ellas ese motivo temático de enfrentamiento bélico de tipo religioso, si bien su composición se llevó a cabo en fechas más tardías. Quedaría, en fin, la *Numancia* en un lugar algo más apartado del resto, pues pese a que su composición tuvo que ser temprana (es una obra cercana a la década de 1580), el contenido alberga pocas relaciones temáticas con el resto, con lo que queda en nuestra clasificación estilométrica más desplazada del resto.

24 Sin embargo ese *consensus tree* muestra solo los resultados partiendo del primer puesto en la lista de palabras más frecuentes. El problema es que si observamos la figura 2, el corpus del italiano muestra los mejores resultados al eliminar entre las 1000 y 1500 palabras más frecuentes. Es fácil aceptar la intuición de que un corpus español debe comportarse de una manera similar a como lo hace un corpus en italiano. Por lo que nuestro *consensus tree* no estaría recogiendo los rangos donde mejor se analiza la autoría.

25 Por esto nos hemos hecho dos preguntas[17]:

1. ¿Cuántos de los rangos organizan *La conquista de Jerusalén* con los textos cervantinos?
2. Ignorando el texto de *La conquista de Jerusalén*, ¿cuántos autores son correctamente reconocidos en cada rango?

26 Comparando ambas respuestas, podremos llegar a un nivel alto de seguridad sobre si el método está organizando correctamente los textos por autoría, y por lo tanto podemos pensar que lo hace correctamente con *La conquista de Jerusalén*. Para responder a ambas preguntas realizamos los dendogramas y sintentizamos los resultados en las figuras 7 y 8. En la figura 7 se observa en cuántos de los dendogramas aparecen los textos de Cervantes organizados. Los valores siguen los siguientes criterios:

- 3: los textos de Cervantes y el discutido aparecen en una rama juntos y aislados
- 2: los textos de Cervantes y el discutido aparecen en una rama juntos pero otro texto de otro autor aparece en la rama
- 1: algunos textos de Cervantes y el discutido aparecen en una rama (aislados de otros autores o no)
- 0: el texto discutido aparece relacionado con un autor diferente a Cervantes

2000	1	2	3	3	3	3	3	3	3	3	3
1500	3	3	3	3	3	3	3	3	3	3	3
1000	3	3	3	3	3	3	3	3	3	3	3
750	3	3	3	3	3	3	3	3	3	3	3
500	1	3	3	3	3	3	3	3	3	3	3
250	0	3	3	3	3	3	3	3	3	3	3
0	1	3	3	3	3	3	3	3	3	3	3
	250	500	750	1000	1500	2000	2500	3000	3500	4000	4500

Figure 7. rangos y resultados de asignación a Cervantes de *La conquista de Jerusalén*

Como se puede observar, de manera sistemática los tres textos de Cervantes y *La conquista de Jerusalén* aparecen agrupados en una rama aislada, aunque no siempre ocurre esto. También es llamativo el hecho de que en un único caso (250 palabras, habiendo eliminado las primeras 250 palabras más frecuentes)[18] el texto de *La conquista de Jerusalén* aparezca relacionado con otro autor. El caso concreto es el siguiente: 27

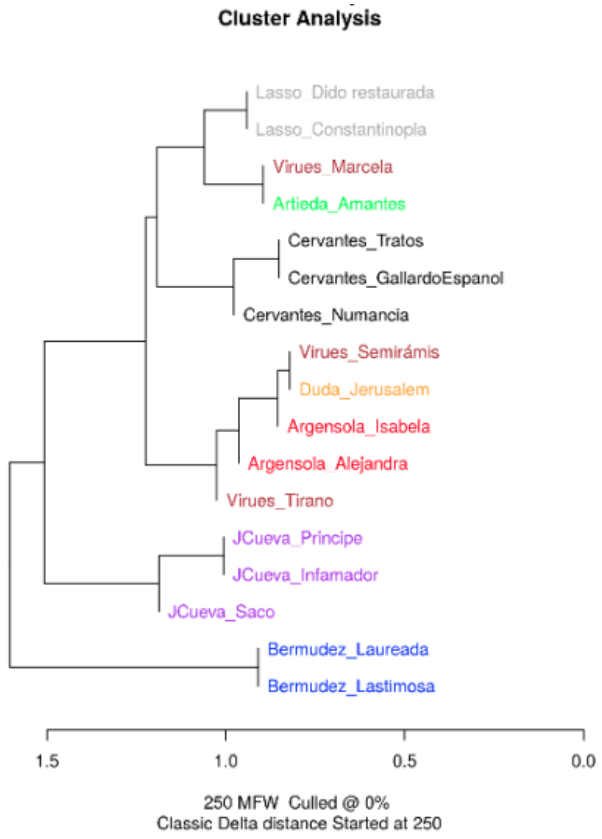

Figure 8. Único análisis que asignó *La conquista* a otro autor

Como se puede observar, en este caso *La conquista de Jerusalén* aparece en una rama en la que se encuentran textos de varios autores (Virués y Argensola). Es decir, el único dendograma que no muestra a Cervantes como autor de la *La conquista de Jerusalén* no señala de manera clara otro autor. 28

Para completar la respuesta, queríamos observar cuántos de los autores aparecen correctamente reconocidos en cada rango (ignorando el caso de *La conquista de Jerusalén)*. El resultado es el siguiente: 29

2000	1	4	5	2	5	6	5	4	5	3	3
1500	4	4	6	6	2	4	4	5	4	5	4
1000	4	4	6	4	6	4	4	4	4	4	3
750	4	5	6	4	6	6	4	6	4	4	5
500	4	5	6	6	4	6	6	6	6	4	4
250	4	5	6	5	4	6	6	6	6	6	4
0	2	5	5	6	6	6	6	6	6	6	6
	250	500	750	1000	1500	2000	2500	3000	3500	4000	4500

Figure 9. cantidad de autores reconocidos por cada rango

Como se puede observar, hay una gran cantidad de rangos (en total 29) que organizan correctamente a los seis autores que están representados en el corpus con varios textos. Estos mismos rangos asignan en la figura 8 el texto de *La conquista de Jerusalén* a Cervantes. Es interesante observar que los rangos de la figura 8 que no asignan *La conquista de Jerusalén* a Cervantes son aquellos que en la figura 9 tienen resultados más deficientes, consiguiendo entre 1 y 4 autores correctamente reconocidos. 30

Si se compara la figura 9 con la figura 3[19], se observa ciertas similitudes: 31

- Se observan resultados pobres al eliminar las 2000 palabras más frecuentes.
- Los rangos hasta 250 y 500 palabras (eliminando o no algunas de las palabras más frecuentes) no parecen suficientes para analizar autoría.
- Se observan rangos óptimos utilizando las palabras más frecuentes sin eliminar ninguna.
- También aparecen rangos óptimos entre 1500 y 4000 palabras, eliminando hasta 500 palabras más frecuentes.

5. Conclusión y futuro trabajo

La utilización de métodos de agrupamiento de aprendizaje automático con rangos de palabras más frecuentes, método comprobado para textos literarios en numerosas lenguas europeas y que reconoce correctamente la autoría del resto de textos de nuestro corpus, agrupa *La conquista de Jerusalén* de manera sistemáticamente con otros textos de Cervantes. Ninguno de los textos escritos por Lasso, Virués, Bermúdez, Argensola o Juan de la Cueva analizados en este estudio aparecen como textos similares a *La conquista*. Estos resultados consolidan la teoría de la autoría cervantina de *La conquista de Jerusalén*. 32

Además observamos en los diferentes análisis relaciones de cercanía tanto cronológicas como temáticas entre diferentes obras de un mismo autor, así como similitudes entre diferentes autores que pueden ser interpretadas a la luz de datos filológicos y de la historia de la literatura, relaciones que necesitan de más investigación. 33

Los rangos desde 1000 hasta 4000 palabras más frecuentes utilizando la lista desde su comienzo se muestran en este estudio como los rangos más estables para estudiar autoría estilométricamente. Comparando estos rangos con los resultados en otras lenguas europeas, los rangos entre 1000 y 3000 palabras más frecuentes parecen los más seguros. Estos resultados también confirman la intuición de que las palabras más frecuentes son las más útiles para reconocimiento de autoría. Sin embargo estos rangos no son los únicos que aparecen con buenos resultados en nuestros corpus, por lo que es necesaria más investigación. 34

La dificultosa tarea de codificar este corpus en XML-TEI nos anima a continuar trabajando en estudios estilométricos sobre autoría y confiamos publicar en breve nuevos resultados. Como continuación de este estudio nos gustaría explorar este mismo caso con un corpus más amplio, procedimientos desarrollados en recientes publicaciones y otros sistemas de evaluación. También nos gustaría continuar analizando estilométricamente otros aspectos relacionados con la obra cervantina y el teatro del Siglo de Oro, por lo que nos gustaría invitar a todos los investigadores que encuentren esta aportación interesante a colaborar con nosotros y ampliar, así, el campo de la discusión sobre autorías en textos escritos en lengua española. 35

Notes

[1] Queremos agradecer a Maciej Eder sus comentarios y opiniones sobre varios detalles de este estudio. También queremos agradecer a Christof Schöch sus comentarios y su ayuda en la construcción del corpus.

[2] También se ha utilizado la palabra *estilometría* en la tradición española para realizar estudios sobre la frecuencia de *hapax*, n-gramas o fragmentos concretos y compararlos con corpus lingüísticos equilibrados, como en Madrigal [Madrigal 2008] [Madrigal 2009]. Sin embargo, los estudios de Madrigal se diferencian notablemente de lo que hoy en día se considera estilometría, tanto por los tipos de unidades analizadas, su cantidad y su elección como por el corpus de comparación y su evaluación.

[3] Para ver una conferencia del mismo autor centrado en los *Bocados de Oro*: http://djwrisley.com/?p=207

[4] Además de las *Ocho comedias y ocho entremeses nunca representados*, que vieron publicación en vida del autor, el corpus de obras dramáticas de Cervantes se amplía a *El cerco de Numancia* y a *Los tratos de Argel,* obras conocidas desde el siglo XVIII, conservadas en copia manuscrita y compuestas originalmente en torno a la década de 1580. Hay, además, un buen número de obras de las que solo conocemos el título gracias a las menciones que Miguel de Cervantes hizo en la *Adjunta* al *Viaje de Parnaso* y que permanecen, todavía, perdidas.

[5] Esto puede tenderse a olvidar si se trabaja con textos en inglés o alemán, donde proyectos como *TextGrid* u *Oxford Text Archive* ponen a disposición de cualquier usuario miles de textos en varios formatos, entre ellos XML-TEI. La situación es radicalmente opuesta para el español, donde el XML-TEI ha sido menos utilizado y, de los principales proyectos que lo han utilizado, ninguno ha puesto a disposición de la comunidad investigadora el código originario (como es el caso del proyecto TESO, Cervantes Virtual o Biblioteca Digital Artelope).

[6] Esto no ha sido posible en algunos casos debido a que algunos de estos autores escribieron solo una o dos obras teatrales. El trabajo se preguntó también cómo proceder con aquellos autores como Cervantes, Francisco de la Cueva o Juan de la Cueva que escribirieron más obras teatrales. Se decidió en este caso también utilizar solamente tres textos de estos autores por varias razones: en primer lugar porque eso hubiese multiplicado el trabajo de preparación del corpus; en segundo lugar porque al representar a algunos autores con una cantidad mucho mayor de textos que los otros autores desequilibraría el corpus y por lo tanto podría afectar a los resultados estilométricos. Es un aspecto que nos gusaría estudiar en mayor detalle en el futuro.

[7] Hemos preferido, en este orden: XML-TEI, XHTML (ePUB), HTML, PDF, imagen.

[8] Para las fuentes en HTML hemos transformado el texto de una manera similar a la que el grupo *Computergestützte literarische Gattungsstilistik* trabaja [Schöch et al. 2014].

[9] Las ediciones del Cervantes Virtual ya habían sido modernizadas de esta manera.

[10] Muestran diferencias importantes entre géneros y es de esperar que las novelas del siglo XXI no se comporten exactamente de la misma manera que las novelas de siglos anteriores.

[11] No es demasiado sorprendente si se tiene en cuenta lo complicado que sigue resultando acceder a corpus literarios óptimos para atribución de autoría. Un intento reciente en esta dirección es el corpus publicado por Calvo Tello y Henny en 2015: https://github.com/cligs/textbox/tree/master/es.

[12] En caso de que haya varios corpus de una lengua (inglés y latín), el de prosa no está marcado y el de poesía está marcado con un guión y la letra *p*. Estamos abiertos a sugerencias sobre la manera de visualizar estos datos sin perder información.

[13] Además del dato numérico, colocamos colores para que sea más fácilmente reconocible al ojo humano

[14] Hemos utilizado la versión 0.6.2.4. Queremos agradecer a los diseñadores de este software su trabajo, documentación, matenimiento y docencia. *Stylo* permite que filólogos y humanistas en general puedan utilizar complejos procedimientos estadísticos para responder a preguntas básicas sobre los textos, sin tener que afrontar el desarrollo de cientos de líneas de código de tratamiento estadístico de textos. Nos gustaría reconocer y valorar su enorme aportación a las Humanidades Digitales, así como recomendar su uso a otros investigadores.

[15] Es decir, que de la lista original de palabras más frecuentes se utilizan aquellas que estarían entre la posición 251 y 2250.

[16] Como explicación de esta visualización, observamos tres ramas principales: Una en la esquina superior izquierda que engloba los textos de Lasso y Juan de la Cueva; una superior derecha donde aparecen los textos de Bermúdez; una tercera inferior que engloba al resto de autores: en ella Cervantes forma su propia subrama; Virués Artieda y Argensola aparecen en la otra subrama.

[17] Los diferentes métodos de aprendizaje automático utilizan técnicas de evaluación que suelen requerir la división de los datos en varios sets: de aprendizaje, de prueba y de implementación. Nuestro corpus no contiene la suficiente cantidad de texto como para permitir que el sistema aprenda de varios textos los rasgos de un autor, probarlo con otro y finalmente aplicarlo. Ni siquiera si tuviésemos todos los textos escritos por estos autores podríamos utilizar una metodología así ya que algunos de los autores escribieron, únicamente, un par obras teatrales.

[18] Es decir, que utiliza las palabras que estarían en los rangos desde el 251 hasta el 500 en la lista de palabras más frecuentes.

[19] Sabemos que nuestro corpus es mucho menor y que es teatro en verso. Pero teniendo en cuenta la total falta de este tipo de trabajos para el español, consideramos que otros investigadores pueden utilizar algunos de los rangos que aparecen como óptimos en la anterior tabla.

Works Cited

Antonucci 2014 Antonucci, Fausta. “La estructura dramática de *La conquista de Jerusalén* por Godofre de Bullón: un análisis comparado con *La Numancia*”. En *Desde Artife. Estudios dedicados a Aldo Ruffinatto en el IV Centenario de las Novelas Ejemplares*, 97-108. Alessandria: Edizioni dell'Orso, 2014.

Antonucci 2015 Antonucci, Fausta. “La estructura dramática del teatro cervantino de la primera “época”: una propuesta de análisis”. *Cuadernos AISPI* 5 (2015): 131-46.

Arata 1989 Arata, Stefano. *Los manuscritos teatrales (siglos XVI y XVII) de la Biblioteca de Palacio*. Pisa: Giardino, 1989.

Arata 1991 Arata, Stefano. “Loyola y Cepeda: Dos dramaturgos del Siglo de Oro en la Biblioteca de Palacio”. *Manuscrt.Cao* IV (1991): 3-15.

Arata 1992 Arata, Stefano. “La conquista de Jerusalén, Cervantes y la generación teatral de 1580”. *Criticón* 54 (1992): 9-112.

Arata 1996 Arata, Stefano. “Teatro y coleccionismo teatral a finales del siglo XVI (el conde de Gondomar y Lope de Vega)”. *Anuario de Lope de Vega* 2 (1996): 7-24.

Arata 1997 Arata, Stefano. “Notas sobre La conquista de Jerusalén y la transmisión manuscrita del primer teatro cervantino”. *Edad de Oro* 16 (1997): 53-66.

Argamon 2008 Argamon, Shlomo. “Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations”. En *Literary and Linguistic Computing* 23 (2) (2008): 131-47.

Argamon et al. 2003 Argamon, Shlomo, Moshe Koppel, Jonathan Fine, y Shimoni Anat Rachel. “Gender, Genre, and Writing Style in Formal Written Texts”, En *Text and Talk*, no. 23 (2003): 321-346.

Argamon et al. 2009 Argamon, Shlomo, Jean-Baptiste Goulain, Russell Horton, y Mark Olsen. “Vive La Différence! Text Mining Gender Difference in French Literature”. En *Digital Humanities Quarterly* 3 (2) (2009). http://www.digitalhumanities.org/dhq/vol/3/2/000042.html.

Baras Escolá 2010 Baras Escolá, Alfredo. “Los textos de Cervantes. Teatro”. *Anales Cervantinos* 42 (2010): 73-88.

Bernaldo de Quirós Mateo 2011 Bernaldo de Quirós Mateo, José Antonio. “La Celestina: Adiciónes primeras amplificadas con adiciónes secundas. Consequencias para la atributión de la autoría”. *Etiópicas*, no. 7 (2011): 87-104.

Brioso Santos 2009 Brioso Santos, Héctor. “A propósito de la historicidad de La conquista de Jerusalén: los cuatro milagros de la primera cruzada”. *Anuario de Estudios Cervantinos* 5 (2009): 101-24.

Brioso Santos 2010 Brioso Santos, Héctor. “Análisis métrico de La conquista de Jerusalén por Godofre de Bullón de... ¿Miguel de Cervantes?” *Cuatrocientos años del Arte Nuevo de hacer comedias de Lope de Vega [Actas]* 2 (2010): 287-94.

Brioso Sánchez y Brioso Santos 2007 Brioso Sánchez, Máximo, y Brioso Santos, Héctor. “De Heliodoro a Tasso y a ¿Cervantes?” *Philología Hispalensis* 21 (2007): 155-72.

Burguillo 2013 Burguillo, Francisco Javier. “Guerra y milicia en los albores del “Arte nuevo”: la “Comedia del saco de Roma” (1579) de Juan de la Cueva”. En *Del pensamiento al texto. Textualización del saber en el Renacimiento español*, 23-60. Madrid: Academia del Hispanismo, 2013.

Burrows 2002 Burrows, John. “‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship”. En *Literary and Linguistic Computing* 17 (3) (2002): 267-87.

Calvo Tello et al. 2015 Calvo Tello, José, Christof Schöch, Nanete Rißler-Pipka, y Tobias Kraft. 2015. “Humanidades Digitales y estudios hispánicos en Alemania”. *Voy y Letra* 26 (1) (2015): 45-61.

Camamis 1977 Camamis, George. *Estudios sobre el cautiverio en el Siglo de Oro*. Madrid: Editorial Gredos, 1977.

Canavaggio 2000 Canavaggio, Jean. “De un Lope a otro Lope: Cervantes ante el teatro de su tiempo”. *Anuario de Lope de Vega* 6 (2000): 51-60.

Canavaggio 2005 Canavaggio, Jean. *Cervantes*. Espasa. Madrid, 2005.

Castillo 2012 Castillo, Moisés R. “Espacios de ambigüedad en el teatro cervantino: La conquista de Jerusalén y los dramas de cautiverio”. *Cervantes: Bulletin of the Cervantes Society of America* 32, 2 (2012): 123-42.

Cerezo Soler 2013 Cerezo Soler, Juan. ““La Conquista de Jerusalén” y la literatura de Cervantes. Nuevas semejanzas que respaldan su autoría”. En *Festina lente. Actas del II congreso internacional Jóvenes Investigadores del Siglo de Oro (JISO 2012)*, editado por Carlos Mata Induráin, Adrián J. Sáez, y Ana Zúñiga Lacruz. Pamplona: Servicio de Publicaciones de la Universidad de Navarra, 2013. http://dadun.unav.edu/handle/10171/29457.

Cerezo Soler 2014 Cerezo Soler, Juan. ““La Conquista de Jerusalén” en su contexto: sobre el personaje colectivo y una vuelta más a la atribución”. *Dicenda: cuadernos de filología hispánica* 32 (2014): 33-49.

Eder 2012 Eder, Maciej. “Mind Your Corpus: Systematic Errors in Authorship Attribution”. En *Digital Humanities 2012: Conference Abstracts*, Hamburg, Hamburg Univ. Press (2012): 181-85.

Eder 2013a Eder, Maciej. “Does Size Matter? Authorship Attribution, Small Samples, Big Problem”. En *Digital Scholarship in the Humanities* 30 (2) (2013): 167-182.

Eder 2013b Eder, Maciej. “Bootstrapping Delta: a safety-net in open-set authorship attribution”. En *Digital Humanities 2013: Conference Abstracts*, Lincoln: University of Nebraska-Lincoln (2013): 169-172.

Eder et al. 2016 Eder, Maciej, Kestemont, Mike, y Rybicki, Jan. *Stylometry with R: A package for computational text analysis*. En *R Journal*, 16 (1), 2016.https://journal.r-project.org/archive/accepted/.

Eisenberg 2003 Eisenberg, Daniel. “¿Qué escribió Cervantes?” En *Sobre Cervantes*, editado por Martínez Torrón, Diego, 9-26. Alcalá de Henares: Centro de Estudios Cervantinos, 2003.

García-Bermejo Giner 2013 García-Bermejo Giner, Miguel. “Estando letras y armas en su punto: el teatro y los aledaños del poder en España a fines del siglo XVI”. En *Del pensamiento al texto. Textualización del saber en el Renacimiento español*, 85-122. Madrid: Academia del Hispanismo, 2013.

Jannidis et al. 2015 Jannidis, Fotis, Steffen Pielström, Christof Schöch, y Thorsten Vitt. “Improving Burrows’ Delta – An Empirical Evaluation of Text Distance Measures.” En *Digital Humanities 2015 Conference Abstracts*. ADHO: Sydney 2015. http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html.

Jannidis y Lauer 2014 Jannidis, Fotis, y Gerhard Lauer. “Burrows’s Delta and Its Use in German Literary History”. En *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, Rochester: Camden House (2014): 29-54.

Jockers 2013 Jockers, Matthew L. *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press (2013).

Kahn 2010 Kahn, Aaron M. “Towards a theory of attribution: Is La conquista de Jerusalén by Miguel de Cervantes?” *Journal of European Studies* 40 (2) (2010): 99-128.

Kestemont et al. 2012 Kestemont, Mike, Kim Luyckx, Walter Daelemans, y Thomas Crombez. “Cross-Genre Authorship Verification Using Unmasking.” En *English Studies* 93 (3) (2012): 340-56.

López 2011 López, Freddy. “Donde se muestran algunos resultados de atribución de autor en torno a la obra cervantina (Wherein are Shown some Results of Autorship Attribution to Cervantes’ Work).” En *Revista Colombiana de Estadística* 34 (1) (2011): 15-37.

Madrigal 2008 Madrigal, José Luis. “Notas sobre la autoría del Lazarillo”. En *Revista de Literatura Española Medieval y del Renacimiento (LEMIR)*, 12

(2008): 137-236.

Madrigal 2009 Madrigal, José Luis. “Tirso, Lope y el Quijote de Avellaneda”. *Revista de Literatura Española Medieval y del Renacimiento (LEMIR),* 13 (2009): 191-250.

Montero Reguera 1994-1995 Montero Reguera, José. “Reseña a Stefano Arata, “La conquista de Jerusalén [...]””. *Manuscrt.Cao* VI (1994-1995): 83-87.

Montero Reguera 1995-1997 Montero Reguera, José. “¿Una nueva obra teatral cervantina? Notas en torno a una reciente atribución”. *Anales Cervantinos* 33 (1995-1997): 355-66.

Rey Hazas 1992 Rey Hazas, Antonio. “Cervantes y Lope ante el personaje colectivo: La Numancia frente a Fuenteovejuna”. *Cervantes y el teatro. Cuadernos de Teatro Clásico* 7 (1992): 69-91.

Rey Hazas 1994 Rey Hazas, Antonio. “Las comedias de cautivos de Cervantes”. *Los imperios orientales en el teatro del Siglo de Oro [Actas de las XVI Jornadas de Teatro Cásico]*, 1994, 29-56.

Rey Hazas 1999 Rey Hazas, Antonio. “Cervantes se reescribe: Teatro y Novelas Ejemplares”. *Criticón* 76 (1999): 119-64.

Rey Hazas 2005 Rey Hazas, Antonio. *Poética de la libertad y otras claves cervantinas*. Madrid: Eneida, 2005.

Rißler-Pipka 2016 Rißler-Pipka, Nanette. “Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales”. En Ehrlicher, Hanno. El otro Quijote. La continuación de Avellaneda y sus efectos. Mesa Redonda-Universität Augsburg, Augsburg (2016). (Manuscrito)

Rodríguez López-Vázquez 2011 Rodríguez López-Vázquez, Alfredo. “La Jerusalén de Cervantes: Nuevas pruebas de su autoría”. *Artifara: Revista de Lenguas y Literaturas ibéricas y latinoamericanas* 11 (2011).

Rojo Alique 1996-1998 Rojo Alique, Pedro C. “Notas acerca del Catálogo de manuscritos de la Biblioteca del Palacio Real de Madrid”. *Manuscrt.Cao* VII (1996-1998): 83-131.

Rybicki y Eder 2011 Rybicki, Jan, y Maciej Eder. “Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?” En *Literary and Linguistic Computing* 26 (3) (2011): 315-21.

Schöch 2014 Schöch, Christof. “Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik”. *Literaturwissenschaft im digitalen Medienwandel*. Beihefte von Philologie im Netz 7, 2014: 130-157. http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf.

Schöch et al. 2014 Schöch, Christof, et al. Toolbox. Universität Würzburg, Würzburg (2014). https://github.com/cligs/toolbox/graphs/contributors

Seroussi et al. 2014 Seroussi, Yanir, Ingrid Zukerman, y Fabian Bohnert. “Authorship Attribution with Topic Models”. En *ACL Anthology* 40 (1) (2014): 269-310.

Vaccari 2006 Vaccari, Debora. “Aproximación al contenido de una carpeta inédita de la Biblioteca Nacional de Madrid (Ms/14612/9)”. En *Campus stellae: haciendo camino en la investigación literaria*, editado por Fernández López, Dolores, Domínguez Pérez, Mónica, y Rodríguez-Gallego, Fernando, 1:466-74. Santiago de Compostela, 2006.

Wrisley 2016 Wrisley, David Joseph. “Modeling the Transmission of Al-Mubashshir Ibn Fātik’s Mukhtār Al-Ḥikam in Medieval Europe: Some Initial Data-Driven Explorations”. En *Journal of Religion, Media and Digital Culture* Special Issue “Digital Humanities in Jewish, Christian and Arabic/Islamic Ancient Traditions”. (5) (2016).

Zimic 1992 [Zimic 1992] Zimic, Stanislav. *El teatro de Cervantes*. Madrid: Castalia, 1992.

van Dalen-Oskam y van Zundert 2007 van Dalen-Oskam, Karina, y Joris van Zundert. “Delta for Middle Dutch: Author and Copyist Distinction in “Walewein””. En *Literary and Linguist Computing* 22 (3) (2007): 345-62.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.

DHQ: Digital Humanities Quarterly

2018
Volume 12 Number 1

The Conquest of Jerusalem: by Cervantes? Styometric analysis on authorship in the Golden Age Spanish theater[es]

José Calvo Tello <jose_dot_calvo_at_uni-wuerzburg_dot_de>, Universidad de Würzburg
Juan Cerezo Soler <juan_dot_cerezosoler_at_gmail_dot_com>, Universidad Autónoma de Madrid
Translation: Jose Calvo Tello <jose_dot_calvo_at_uni-wuerzburg_dot_de>, Universidad de Würzburg

Abstract

In this article we apply stylometric methods to approach the authorship problem of the comedy *La conquista de Jerusalén*, attributed since its discovery to Miguel de Cervantes. For this purpose we have performed numerous analyses with different range of most frequent words in a total of seventeen theater plays, all of them written by the seven authors that define the *generación teatral de 1580* and who wrote plays actively when *La conquista* was composed. We have used the distant measure Delta to cluster the text.

Note on Translation

For articles in languages other than English, DHQ provides an English-language abstract to support searching and discovery, and to enable those not fluent in the article's original language to get a basic understanding of its contents. In many cases, machine translation may be helpful for those seeking more detailed access. While DHQ does not typically have the resources to translate articles in full, we welcome contributions of effort from readers. If you are interested in translating any article into another language, please contact us at editors@digitalhumanities.org and we will be happy to work with you.

Works Cited

Antonucci 2014 Antonucci, Fausta. “La estructura dramática de *La conquista de Jerusalén* por Godofre de Bullón: un análisis comparado con *La Numancia*”. En *Desde Artife. Estudios dedicados a Aldo Ruffinatto en el IV Centenario de las Novelas Ejemplares*, 97-108. Alessandria: Edizioni dell'Orso, 2014.

Antonucci 2015 Antonucci, Fausta. “La estructura dramática del teatro cervantino de la primera “época”: una propuesta de análisis”. *Cuadernos AISPI* 5 (2015): 131-46.

Arata 1989 Arata, Stefano. *Los manuscritos teatrales (siglos XVI y XVII) de la Biblioteca de Palacio*. Pisa: Giardino, 1989.

Arata 1991 Arata, Stefano. “Loyola y Cepeda: Dos dramaturgos del Siglo de Oro en la Biblioteca de Palacio”. *Manuscrt.Cao* IV (1991): 3-15.

Arata 1992 Arata, Stefano. “La conquista de Jerusalén, Cervantes y la generación teatral de 1580”. *Criticón* 54 (1992): 9-112.

Arata 1996 Arata, Stefano. “Teatro y coleccionismo teatral a finales del siglo XVI (el conde de Gondomar y Lope de Vega)”. *Anuario de Lope de Vega* 2 (1996): 7-24.

Arata 1997 Arata, Stefano. “Notas sobre La conquista de Jerusalén y la transmisión manuscrita del primer teatro cervantino”. *Edad de Oro* 16 (1997): 53-66.

Argamon 2008 Argamon, Shlomo. “Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations”. En *Literary and Linguistic Computing* 23 (2) (2008): 131-47.

Argamon et al. 2003 Argamon, Shlomo, Moshe Koppel, Jonathan Fine, y Shimoni Anat Rachel. “Gender, Genre, and Writing Style in Formal Written Texts”, En *Text and Talk*, no. 23 (2003): 321-346.

Argamon et al. 2009 Argamon, Shlomo, Jean-Baptiste Goulain, Russell Horton, y Mark Olsen. “Vive La Différence! Text Mining Gender Difference in French Literature”. En *Digital Humanities Quarterly* 3 (2) (2009). http://www.digitalhumanities.org/dhq/vol/3/2/000042.html.

Baras Escolá 2010 Baras Escolá, Alfredo. “Los textos de Cervantes. Teatro”. *Anales Cervantinos* 42 (2010): 73-88.

Bernaldo de Quirós Mateo 2011 Bernaldo de Quirós Mateo, José Antonio. “La Celestina: Adiciónes primeras amplificadas con adiciónes secundas. Consequencias para la atributión de la autoría”. *Etiópicas*, no. 7 (2011): 87-104.

Brioso Santos 2009 Brioso Santos, Héctor. “A propósito de la historicidad de La conquista de Jerusalén: los cuatro milagros de la primera cruzada”. *Anuario de Estudios Cervantinos* 5 (2009): 101-24.

Brioso Santos 2010 Brioso Santos, Héctor. “Análisis métrico de La conquista de Jerusalén por Godofre de Bullón de... ¿Miguel de Cervantes?” *Cuatrocientos años del Arte Nuevo de hacer comedias de Lope de Vega [Actas]* 2 (2010): 287-94.

Brioso Sánchez y Brioso Santos 2007 Brioso Sánchez, Máximo, y Brioso Santos, Héctor. “De Heliodoro a Tasso y a ¿Cervantes?” *Philología Hispalensis* 21 (2007): 155-72.

Burguillo 2013 Burguillo, Francisco Javier. “Guerra y milicia en los albores del “Arte nuevo”: la “Comedia del saco de Roma” (1579) de Juan de la Cueva”. En *Del pensamiento al texto. Textualización del saber en el Renacimiento español*, 23-60. Madrid: Academia del Hispanismo, 2013.

Burrows 2002 Burrows, John. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship". En *Literary and Linguistic Computing* 17 (3) (2002): 267-87.

Calvo Tello et al. 2015 Calvo Tello, José, Christof Schöch, Nanete Rißler-Pipka, y Tobias Kraft. 2015. "Humanidades Digitales y estudios hispánicos en Alemania". *Voy y Letra* 26 (1) (2015): 45-61.

Camamis 1977 Camamis, George. *Estudios sobre el cautiverio en el Siglo de Oro*. Madrid: Editorial Gredos, 1977.

Canavaggio 2000 Canavaggio, Jean. "De un Lope a otro Lope: Cervantes ante el teatro de su tiempo". *Anuario de Lope de Vega* 6 (2000): 51-60.

Canavaggio 2005 Canavaggio, Jean. *Cervantes*. Espasa. Madrid, 2005.

Castillo 2012 Castillo, Moisés R. "Espacios de ambigüedad en el teatro cervantino: La conquista de Jerusalén y los dramas de cautiverio". *Cervantes: Bulletin of the Cervantes Society of America* 32, 2 (2012): 123-42.

Cerezo Soler 2013 Cerezo Soler, Juan. ""La Conquista de Jerusalén" y la literatura de Cervantes. Nuevas semejanzas que respaldan su autoría". En *Festina lente. Actas del II congreso internacional Jóvenes Investigadores del Siglo de Oro (JISO 2012)*, editado por Carlos Mata Induráin, Adrián J. Sáez, y Ana Zúñiga Lacruz. Pamplona: Servicio de Publicaciones de la Universidad de Navarra, 2013. http://dadun.unav.edu/handle/10171/29457.

Cerezo Soler 2014 Cerezo Soler, Juan. ""La Conquista de Jerusalén" en su contexto: sobre el personaje colectivo y una vuelta más a la atribución". *Dicenda: cuadernos de filología hispánica* 32 (2014): 33-49.

Eder 2012 Eder, Maciej. "Mind Your Corpus: Systematic Errors in Authorship Attribution". En *Digital Humanities 2012: Conference Abstracts*, Hamburg, Hamburg Univ. Press (2012): 181-85.

Eder 2013a Eder, Maciej. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem". En *Digital Scholarship in the Humanities* 30 (2) (2013): 167-182.

Eder 2013b Eder, Maciej. "Bootstrapping Delta: a safety-net in open-set authorship attribution". En *Digital Humanities 2013: Conference Abstracts*, Lincoln: University of Nebraska-Lincoln (2013): 169-172.

Eder et al. 2016 Eder, Maciej, Kestemont, Mike, y Rybicki, Jan. *Stylometry with R: A package for computational text analysis*. En *R Journal*, 16 (1), 2016.https://journal.r-project.org/archive/accepted/.

Eisenberg 2003 Eisenberg, Daniel. "¿Qué escribió Cervantes?" En *Sobre Cervantes*, editado por Martínez Torrón, Diego, 9-26. Alcalá de Henares: Centro de Estudios Cervantinos, 2003.

García-Bermejo Giner 2013 García-Bermejo Giner, Miguel. "Estando letras y armas en su punto: el teatro y los aledaños del poder en España a fines del siglo XVI". En *Del pensamiento al texto. Textualización del saber en el Renacimiento español*, 85-122. Madrid: Academia del Hispanismo, 2013.

Jannidis et al. 2015 Jannidis, Fotis, Steffen Pielström, Christof Schöch, y Thorsten Vitt. "Improving Burrows' Delta – An Empirical Evaluation of Text Distance Measures." En *Digital Humanities 2015 Conference Abstracts*. ADHO: Sydney 2015. http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html.

Jannidis y Lauer 2014 Jannidis, Fotis, y Gerhard Lauer. "Burrows's Delta and Its Use in German Literary History". En *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, Rochester: Camden House (2014): 29-54.

Jockers 2013 Jockers, Matthew L. *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press (2013).

Kahn 2010 Kahn, Aaron M. "Towards a theory of attribution: Is La conquista de Jerusalén by Miguel de Cervantes?" *Journal of European Studies* 40 (2) (2010): 99-128.

Kestemont et al. 2012 Kestemont, Mike, Kim Luyckx, Walter Daelemans, y Thomas Crombez. "Cross-Genre Authorship Verification Using Unmasking." En *English Studies* 93 (3) (2012): 340-56.

López 2011 López, Freddy. "Donde se muestran algunos resultados de atribución de autor en torno a la obra cervantina (Wherein are Shown some Results of Autorship Attribution to Cervantes' Work)." En *Revista Colombiana de Estadística* 34 (1) (2011): 15-37.

Madrigal 2008 Madrigal, José Luis. "Notas sobre la autoría del Lazarillo". En *Revista de Literatura Española Medieval y del Renacimiento (LEMIR)*, 12 (2008): 137-236.

Madrigal 2009 Madrigal, José Luis. "Tirso, Lope y el Quijote de Avellaneda". *Revista de Literatura Española Medieval y del Renacimiento (LEMIR),* 13 (2009): 191-250.

Montero Reguera 1994-1995 Montero Reguera, José. "Reseña a Stefano Arata, "La conquista de Jerusalén [...]"". *Manuscrt.Cao* VI (1994-1995): 83-87.

Montero Reguera 1995-1997 Montero Reguera, José. "¿Una nueva obra teatral cervantina? Notas en torno a una reciente atribución". *Anales Cervantinos* 33 (1995-1997): 355-66.

Rey Hazas 1992 Rey Hazas, Antonio. "Cervantes y Lope ante el personaje colectivo: La Numancia frente a Fuenteovejuna". *Cervantes y el teatro. Cuadernos de Teatro Clásico* 7 (1992): 69-91.

Rey Hazas 1994 Rey Hazas, Antonio. "Las comedias de cautivos de Cervantes". *Los imperios orientales en el teatro del Siglo de Oro [Actas de las XVI Jornadas de Teatro Cásico]*, 1994, 29-56.

Rey Hazas 1999 Rey Hazas, Antonio. "Cervantes se reescribe: Teatro y Novelas Ejemplares". *Criticón* 76 (1999): 119-64.

Rey Hazas 2005 Rey Hazas, Antonio. *Poética de la libertad y otras claves cervantinas*. Madrid: Eneida, 2005.

Rißler-Pipka 2016 Rißler-Pipka, Nanette. "Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales". En Ehrlicher, Hanno. El otro Quijote. La continuación de Avellaneda y sus efectos. Mesa Redonda-Universität Augsburg, Augsburg (2016). (Manuscrito)

Rodríguez López-Vázquez 2011 Rodríguez López-Vázquez, Alfredo. "La Jerusalén de Cervantes: Nuevas pruebas de su autoría". *Artifara: Revista de Lenguas y Literaturas ibéricas y latinoamericanas* 11 (2011).

Rojo Alique 1996-1998 Rojo Alique, Pedro C. “Notas acerca del Catálogo de manuscritos de la Biblioteca del Palacio Real de Madrid”. *Manuscrt.Cao* VII (1996-1998): 83-131.

Rybicki y Eder 2011 Rybicki, Jan, y Maciej Eder. “Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?” En *Literary and Linguistic Computing* 26 (3) (2011): 315-21.

Schöch 2014 Schöch, Christof. “Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik”. *Literaturwissenschaft im digitalen Medienwandel*. Beihefte von Philologie im Netz 7, 2014: 130-157. http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf.

Schöch et al. 2014 Schöch, Christof, et al. Toolbox. Universität Würzburg, Würzburg (2014). https://github.com/cligs/toolbox/graphs/contributors

Seroussi et al. 2014 Seroussi, Yanir, Ingrid Zukerman, y Fabian Bohnert. “Authorship Attribution with Topic Models”. En *ACL Anthology* 40 (1) (2014): 269-310.

Vaccari 2006 Vaccari, Debora. “Aproximación al contenido de una carpeta inédita de la Biblioteca Nacional de Madrid (Ms/14612/9)”. En *Campus stellae: haciendo camino en la investigación literaria*, editado por Fernández López, Dolores, Domínguez Pérez, Mónica, y Rodríguez-Gallego, Fernando, 1:466-74. Santiago de Compostela, 2006.

Wrisley 2016 Wrisley, David Joseph. “Modeling the Transmission of Al-Mubashshir Ibn Fātik’s Mukhtār Al-Ḥikam in Medieval Europe: Some Initial Data-Driven Explorations”. En *Journal of Religion, Media and Digital Culture* Special Issue “Digital Humanities in Jewish, Christian and Arabic/Islamic Ancient Traditions”. (5) (2016).

Zimic 1992 [Zimic 1992] Zimic, Stanislav. *El teatro de Cervantes*. Madrid: Castalia, 1992.

van Dalen-Oskam y van Zundert 2007 van Dalen-Oskam, Karina, y Joris van Zundert. “Delta for Middle Dutch: Author and Copyist Distinction in “Walewein””. En *Literary and Linguist Computing* 22 (3) (2007): 345-62.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.