Exploración y análisis estilométrico de autoría sobre obras teatrales de autoría discutida de Moreto

Fecha: 6 de julio de 2018

Autor: José Calvo Tello

Para: Grupo Moretianos

Fuente de corpus: Moretianos, Canon 60

Table of Contents

1. Resumen

En los últimos años un área específica de las Humanidades Digitales se ha mostrado notablemente exitosa en analizar problemas de autoría utilizando la frecuencia de rasgos estilísticos léxicos: la estilometría. En este informe analizo un conjunto de obras de autoría discutida, todas ellas con Moreto como posible candidato. Para ello se utilizan diferentes metodologías de estilometría y aprendizaje automático sobre el corpus de obras preparado por Moretianos y Canon 60.

2. Problema

La investigación en literatura ha mostrado dudas sobre la autoría de Moreto en numerosas obras de teatro. En algunos casos se duda si Moreto fue su autor, en otros casos de textos escritos a múltiples manos se duda hasta qué punto Moreto fue su autor, mientras que un tercer grupo de textos muestra ambos problemas combinados: se duda si Moreto fue su autor, y en qué medida.

En un primer momento, se decidió abordar el primer tipo de problema: analizar si Moreto es el autor de un conjunto de textos cuya autoría es dudosa pero de los que se piensa que Moreto podría ser su único autor.

3. Corpus

Las obras provienen de tres subcorpus:

- Moreto-seguro: 18 obras de teatro de cuya autoría no dudamos que sea de Moreto y solo de Moreto, editadas por Moretianos.
- Moreto-dudoso: 10 obras de teatro cuya autoría dudamos que sea de Moreto, pero pensamos que si lo fuesen serían solo de él, editadas por Moretianos, concretamente:[1] *Los engaños de un engaño, Los siete durmientes, Travesuras de Pantoja, Empezar a ser amigos, El más ilustre francés, San Luis Bertrán, El azote de su patria y renegado Abdenaga, Amor y obligación, Merecer para alcanzar* y *Cómo se vengan los nobles*.
- Otros-seguro: 59 obras de teatro cuya autoría no dudamos que no sea de Moreto (y solo fueron escritas por un autor), editadas principalmente por *Canon 60* (56 textos) y 3 editadas por Moretianos (escritas por Gómez).

De esta manera en combinación podemos tener un corpus de 28 obras de Moreto de autoría única (en duda y segura), un corpus de 77 obras de autoría segura (de Moreto o de otros) o un conjunto total de 87 obras.

De los originales fueron eliminados (desde sus diferentes formatos, Word y HTML) los encabezamientos, paratextos, acotaciones, numeración de versos y referencias al personaje que habla.

1 Estas obras fueron seleccionadas filtrando según los datos que Moretianos puso a disposición por autor: "Moreto"; cert autor: "dudosa".

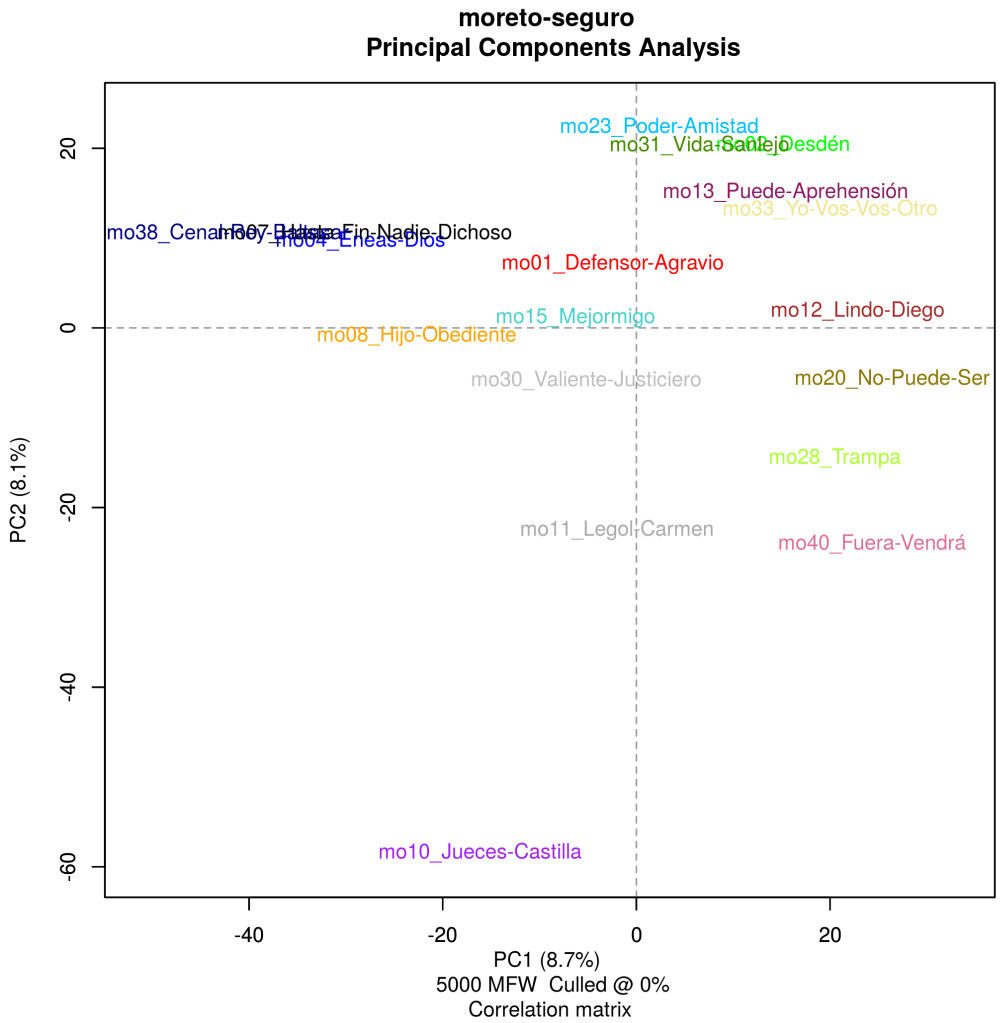
4. Metodología

Los pasos realizados son los siguientes:

1. En primer lugar se han explorado diferentes combinaciones de los subcorpus mediante técnicas de reducción de dimensionalidad y clustering
2. En segundo lugar se evaluaron los diferentes parámetros (versión de Delta, cantidad de rasgos o MFWs, algoritmo de aprendizaje)
3. En tercer lugar se predijeron las clases autoriales de las obras discutidas, tanto de manera multiclase como binaria, así como mediante Logistic Regression

4.1. Exploración de textos mediante clustering y reducción de dimensionalidad

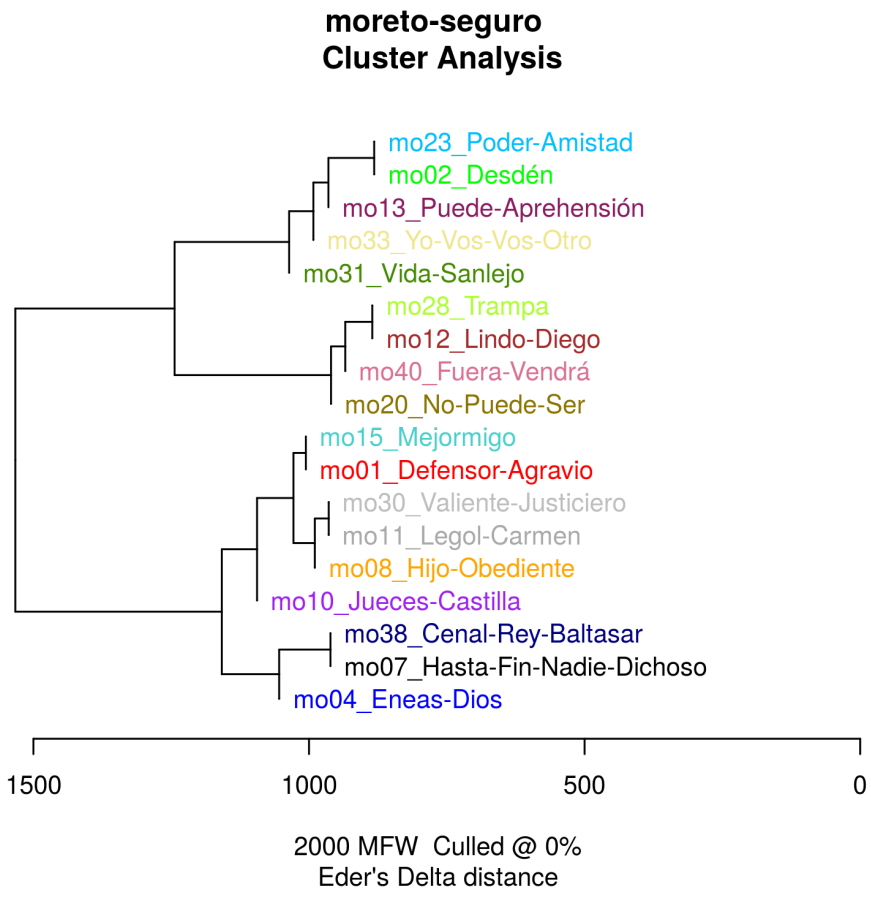
En primer lugar analicé el subcorpus Moreto-seguro mediante dos metodologías. Veamos los resultados del primero, Principal Component Analysis, una técnica explorativa de reducción de dimensionalidad muy utilizada en trabajos de lingüística computacional y estilometría (Baayen 2008, 118; Stamou 2008). Los parámetros utilizados son 1gram, 5000 MFW, correlation:



Como se puede observar, los textos se distribuyen de manera homogénea a lo largo del componente principal (eje horizontal), dejando de cierta manera aislada *Nadie-Dichosos*, *Eneas-Dios* y *Cena-Rey-Baltasar*. En el segundo compomente (el eje vertical) el único texto que aparece claramente separado del grupo es *Jueces-Castilla*.

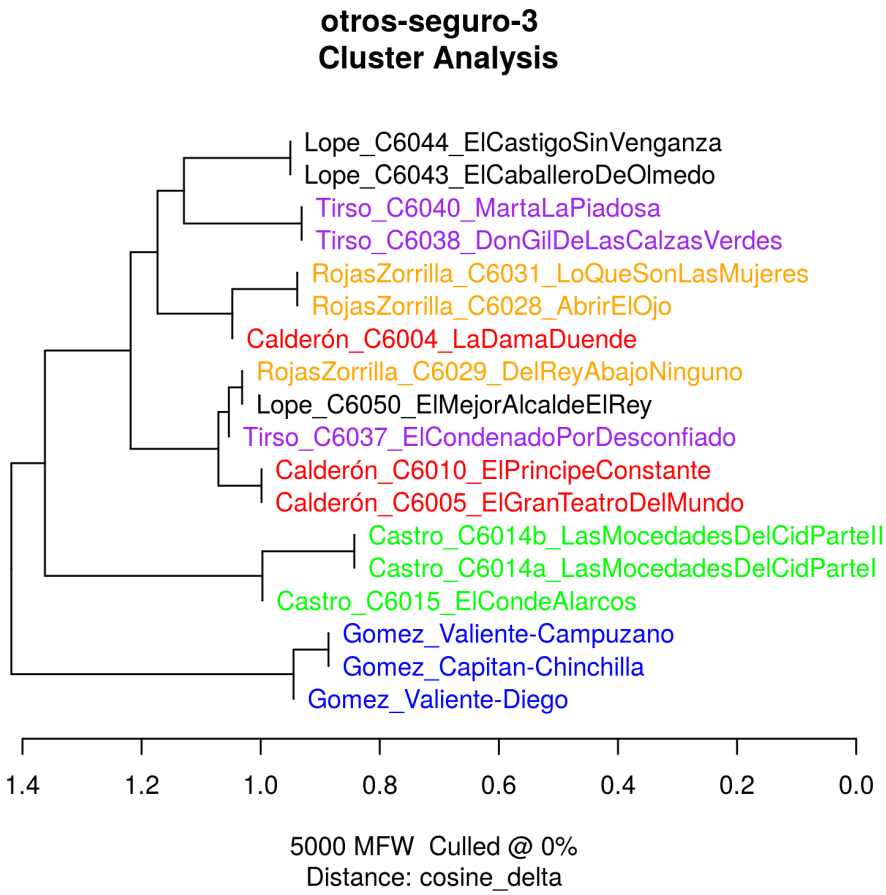
Veamos ahora el resultado del análisis de cluster utilizando la medida de distancia textual propuesta por John Burrows (2002) pero modificada por diferentes autores como Smith y Aldridge (2011, Cosine Delta) o Eder (Eder, Kestemont, and Rybicki 2016). Los paramétros utilizados son Eders Delta, 1gram, 2000 MFW,[2] Ward:

2 Esta combinación de parámetros es la que mejores resultados tendrá en la sección 4.3. sobre evaluación. En los documentos pueden observarse el resto de resultados.



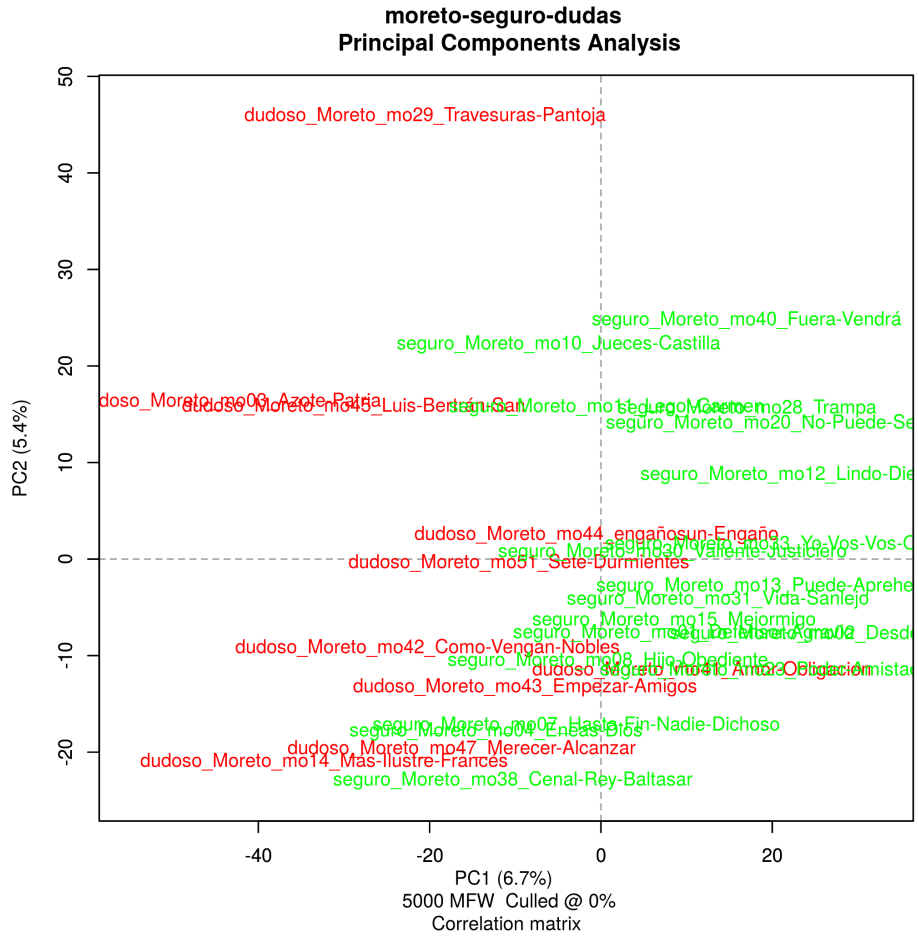
De nuevo observamos que *Jueces-Castilla* es el texto más apartado del resto de las obras (la obra con una posición más a la izquierda en el eje horizontal). ¿Hay algún aspecto de *Jueces-Castilla* que explique la posición aislada en ambos análisis? La posible razón de esta separación es el género literario de esta obra, comedia, único representante de este subgénero en este corpus; sin embargo hay otros subgéneros literarios también representados por un único caso (bíblico, drama histórico, figurón, política-religiosa), por lo que no parece que este dato sea suficiente para explicar su aislamiento.

Como siguiente paso, he analizado mediante clustering los resultados al seleccionar de manera aleatoria tres textos de autoría indiscutida de los otros autores:

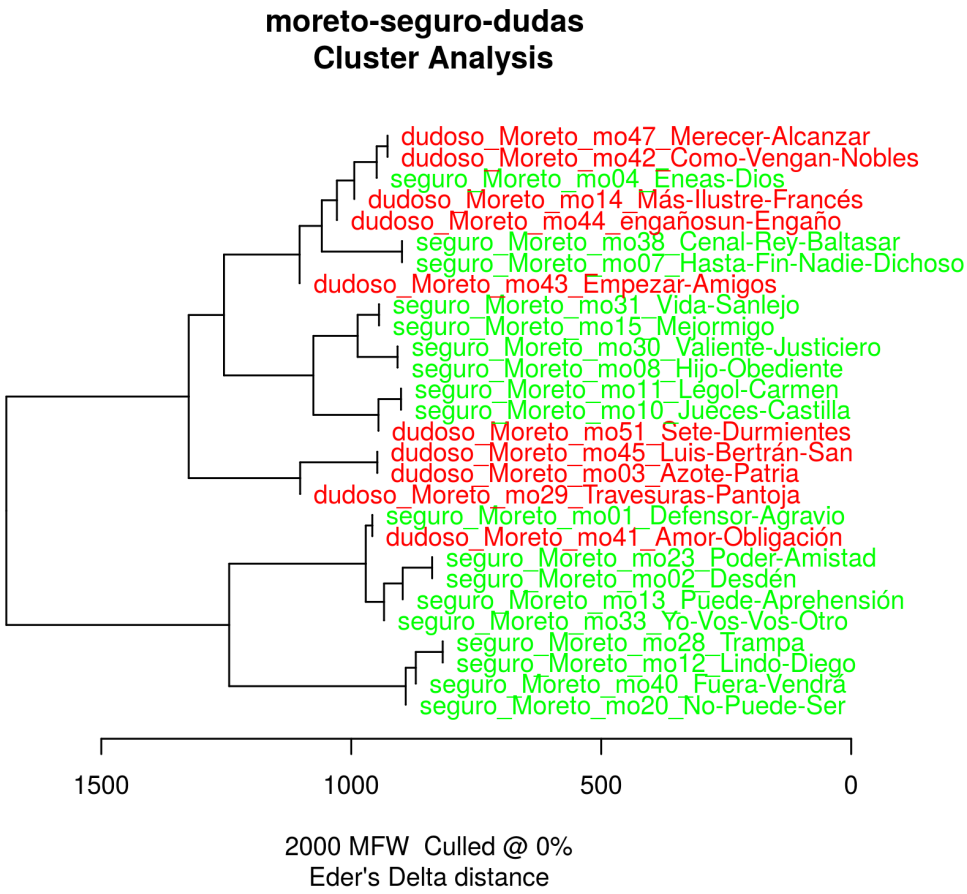


Dependiendo de los parámetros que se elijan (cantidad de palabras, versión de Delta; en este caso 5000 MFW y cosine Delta), los resultados son más o menos satisfactorios. Como se observa, algunos de los textos no aparecen claramente organizados por autor (Calderón, Lope, Tirso, Zorrilla). La clusterización tiene dos desventajas: en primer lugar, se elimina información ya que observamos las relaciones más estrechas entre los textos, pero hemos perdido el resto de relaciones con los otros textos. En segundo lugar, la evaluación resulta más complicada que con otro tipo de tareas que veremos más adelante como clasificación o aprendizaje automático supervisado.

Veamos ahora cuál es el resultado si colocamos solo los textos tanto seguro como discutidos de Moreto, en primer lugar en PCA:



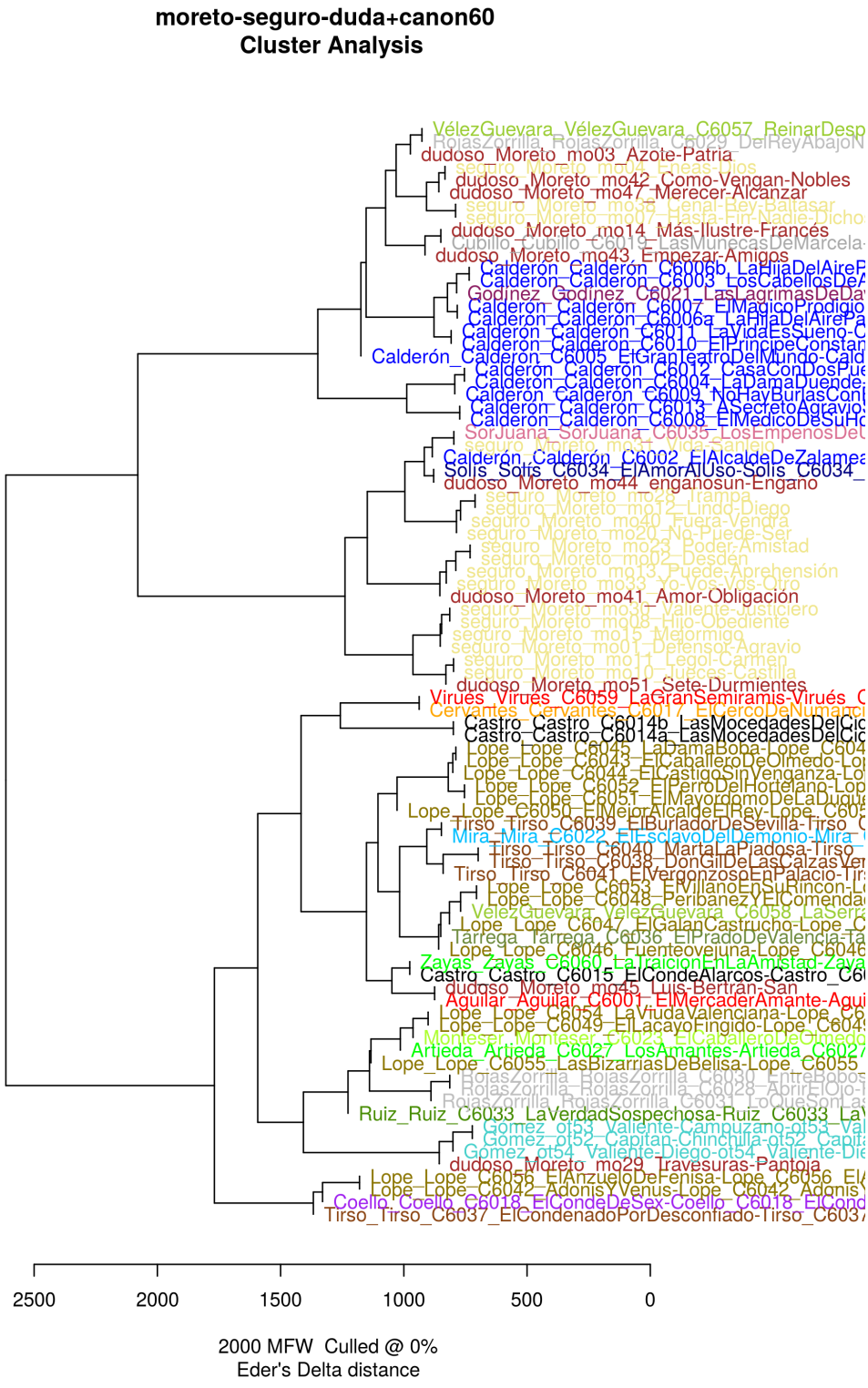
Los textos *Azote-Patria* y *Luis-Betrán* aparecen aislados del resto, aunque con valores similares a los que tienen otros ~~tedtos~~ como *Merecer-Alcanzar* o *Mas-Ilustre-Frances* en la dimensión primera u horizontal. En la segunda dimensión, la vertical, aparece claramente aislado *Travesuras-Pantoja*. Tratemos de ver ahora si se observan los mismos resultados en el cluster:



Las tres obras dudosas que comentábamos anteriormente como aisladas (*Azote-Patria*, *Luis-Betrán* y *Travesuras-Pantoja*) aparecen aquí también como aisladas en un único subcluster, hecho que va sosteniendo la idea de que esas tres obras no serían de Moreto.

En la parte superior de la anterior imagen se encuentra un subcluster con 8 obras en las que 5 serían dudosas y otras 3 serían seguras de Moreto. Especialmente llamativa es la posición en él de *Eneas-Dios*. Recordemos que esta obra no tenía ninguna posición extrema en el PCA o el cluster de las obras seguras de Moreto. *Empezar-Amigos* tienen una posición llamativa como uno de los valores más extremos en el eje horizontal. Finalmente las obras de *Amor-Obligación* y *Siete-Durmientes* quedarían estrechamente relacionada con otras obras seguras de Moreto, por lo que su autoría parecería confirmada.

Finalmente podemos colocar todos los textos que tenemos, discutidos y no discutidos, de Moreto o de otros autores, y visualizarlo mediante clustering:



De las 18 obras seguras de Moreto, solo *Vida-Sanalejo* aparece de manera claramente mal ordenada (junto con obras de Sor Juana, Calderón, Solís y una dudosa). *Amor-obligación* y *Siete-Durmientes* aparecen de nuevo claramente entre las obras de Moreto. Por contra, algunas obras dudosas aparecen claramente alejadas de las obras de Moreto, como *Travesuras-Pantoja* o *Luis-Bertran-San*.

4.2. Evaluación de parámetros mediante clasificación

Como he señalado antes, la modificación de los parámetros como cantidad de palabras o versión de Delta alteran los resultados. Por eso es necesario realizar un trabajo de evaluación de los paramétros sobre los textos cuya autoría sea segura. En este caso no estamos ante un caso de aprendizaje automático no supervisado cuyo resultado debe ser interpretado como en los casos anterior, sino ante un caso de aprendizaje automático supervisado. Aunque esta metodología es hasta ahora poco utilizada en los trabajos de estilometría en español, está siendo cada vez más asumida como el procedimiento a realizar en contexto internacional (Juola 2015; Kestemont et al. 2016).

Para este objetivo he creado un corpus lo más amplio posible que contuviese por lo menos tres textos por cada autor; el resultado ha sido un corpus de los siguientes autores: Calderón, Castro, Gómez, Lope, Moreto, Rojas Zorrilla, Tirso. De este corpus se han ido probando en bucle la cantidad de palaras (de 1000 a 5000), la versión de Delta (*Cosine*, *Eder* o *Classic*) y el algoritmo de clasificación (*Support Vector Machines*, *k-nearest Neighbours*, *Random Forest y Logistic Regression*) utilizando cross validation. Veamos las diez combinaciones de parámetros con mejores resultados para la distinción de autores:

autor-mean accuracy	autor-std accuracy	MFWs	classifiers	distances
0.90	0.07	1000	SVC	cosine_delta
0.86	0.12	1000	LR	cosine_delta
0.86	0.12	2000	SVC	cosine_delta
0.86	0.12	2000	LR	cosine_delta
0.86	0.12	3000	SVC	cosine_delta
0.86	0.12	1000	LR	dist.eder
0.86	0.12	2000	SVC	dist.eder
0.86	0.12	2000	LR	dist.eder
0.86	0.12	3000	SVC	dist.eder
0.86	0.12	3000	LR	dist.eder

Como se puede observar, cosine Delta con 1000 palabras (clasificado con Support Vector Machines) consigue los mejores resultados para distinguir los diferentes autores, con un valor de 0.9 de media sobre el cross validation (desviación estándar de 0.07), es decir, la clasificación es típicamente correcta entre el 97% y el 83% de las veces.

En este caso en realidad el verdadero foco de nuestro análisis no es la atribución de autoría de un texto concreto cuyos autores candidatos hayamos listado (clasificación multiclase), sino verificación de autoría: ¿fueron los textos discutidos escritos por Moreto? Por eso podemos binarizar la autoría señalando dos únicos valores: autor_Moreto y autor_no_Moreto. En el corpus anterior estos son los resultados de la evaluación:

Moreto-mean	Moreto-std	MFWs	classifiers	distances
0.95	0.07	2000	RF	dist.eder
0.95	0.07	3000	RF	dist.eder
0.90	0.07	1000	RF	dist.delta
0.90	0.07	2000	RF	dist.delta
0.86	0.12	5000	RF	cosine_delta
0.86	0.00	1000	KNN	cosine_delta
0.86	0.00	1000	SVC	cosine_delta
0.86	0.00	1000	LR	cosine_delta
0.86	0.00	2000	KNN	cosine_delta
0.86	0.00	2000	SVC	cosine_delta

Como se puede observar, en este caso la media del cross validation se encuentra en 0.95, con una desviación estándar de 0.07, por lo que el rango de 90% a 100% queda cubierto en los casos típicos. Los parámetros en este caso son Eder's Delta con 2000 palabras, mediante algoritmo Random Forest.

Aún así, nos interesa además tener los resultados para un corpus mayor. En él he colocado todos los textos de Moreto cuyos subgéneros literarios son compartidos por los textos discutidos (en concreto: capa y pesada, comedia, hagiográfico, palatina y semihistórico; un total de 14 obras), junto con el resto de las obras seguras de otros autores (59 obras). Los resultados de la clasificación binaria en este caso son los siguientes:

Moreto-mean	Moreto-std	MFWs	classifiers	distances
0.88	0.11	3000	KNN	cosine_delta
0.86	0.08	2000	SVC	dist.eder
0.86	0.11	2000	KNN	cosine_delta
0.86	0.11	4000	KNN	cosine_delta
0.85	0.07	2000	SVC	cosine_delta
0.85	0.07	1000	SVC	dist.eder
0.85	0.07	1000	LR	dist.eder
0.85	0.07	2000	LR	dist.eder
0.85	0.07	3000	SVC	dist.eder
0.85	0.07	3000	LR	dist.eder

Aunque la media más alta es para cosine Delta, su desviación estándar es mayor, sus resultados fluctúan más. El siguiente mejor resultado es Eder's Delta (2000 MFW), con dos décimas menos de media de *accuracy* pero tres décimas menos de desviación estándar. Es decir, es la misma combinación de parámetros que nos encontramos en el caso anterior. Por eso esta selección de parámetros se han utilizado en este informe. Es decir, en las tres evaluaciones de las clasificaciones nos movemos en los rangos de 0.97 y 0.78 de precisión.

4.3. Predicción de autoría

Una vez los métodos han sido evaluados, podemos utilizar los algoritmos para que predizcan la categoría autorial de los casos en duda. En primer lugar he planteado la preguntado de manera multiclase, es decir, el algoritmo tiene a disposición los diferentes nombres de los autores y puede seleccionar a cada uno como posible autor de cada obra discutida. Para ello hemos utilizado los parámetros que han surgido en la evaluación como los mejores, es decir Eder's Delta, 2000 MFW, utilizando el algoritmo Support Vector Machines Classifiers (SVC). Veamos los resultados:

Autor predecido	Texto discutido
Tirso	mo03_Azote-Patria
Calderón	mo14_Más-Ilustre-Francés
Gómez	mo29_Travesuras-Pantoja
Moreto	mo41_Amor-Obligación
Moreto	mo42_Como-Vengan-Nobles
Calderón	mo43_Empezar-Amigos
Moreto	mo44_engañosun-Engaño
Lope	mo45_Luis-Bertrán-San
Moreto	mo47_Merecer-Alcanzar
Moreto	mo51_Sete-Durmientes

De las 10 obras discutidas, 5 han sido asignadas a Moreto, 2 a Calderón y el resto a Tirso, Gómez y Lope. Hay que tener en cuenta que estos resultados no significan que estos autores realmente escribieran esas obras, sino que sus frecuencias léxicas son más similares a las de las obras de estos autores que se encuentran en el corpus de comparación. Además, solo de Moreto tenemos uno de los mayores corpus de comparación hasta ahora posibles, mientras que de los demás solo tenemos corpus oportunísticos obtenidos de *Canon 60*. Aunque los resultados puedan estar señalando pistas interesantes, de ninguna manera debemos entender estas atribuciones a otros autores que no sean Moreto como seguras.

En cuanto a las obras atribuidas a Moreto, *Siete-Durmientes* y *Amor-Obligación* han aparecido hasta ahora en todos los análisis como de Moreto, por lo que los resultados siguen siendo coherentes. Las otras tres obras ya habían aparecido como obras cercanas a *Eneas-Dios* y *Hasta-Fin-Nadie-Dichoso* de Moreto, ambas en el corpus de aprendizaje de este paso.

Como siguiente paso he binarizado las autorías en dos únicos valores posibles como ya se había realizado en la evaluación. En este caso el algoritmo agrupa al resto de autores en una única clase y lo único que intenta realizar es la división entre los textos de Moreto y los del resto. De nuevo utilizo los parámetros Eder's Delta, 2000 MFWs, SVC:

Autor predecido	Texto discutido
No-Moreto	mo03_Azote-Patria
No-Moreto	mo14_Más-Ilustre-Francés
No-Moreto	mo29_Travesuras-Pantoja
Moreto	mo41_Amor-Obligación
No-Moreto	mo42_Como-Vengan-Nobles
No-Moreto	mo44_engañosun-Engaño
No-Moreto	mo45_Luis-Bertrán-San
No-Moreto	mo47_Merecer-Alcanzar
Moreto	mo51_Sete-Durmientes
No-Moreto	mo43_Empezar-Amigos

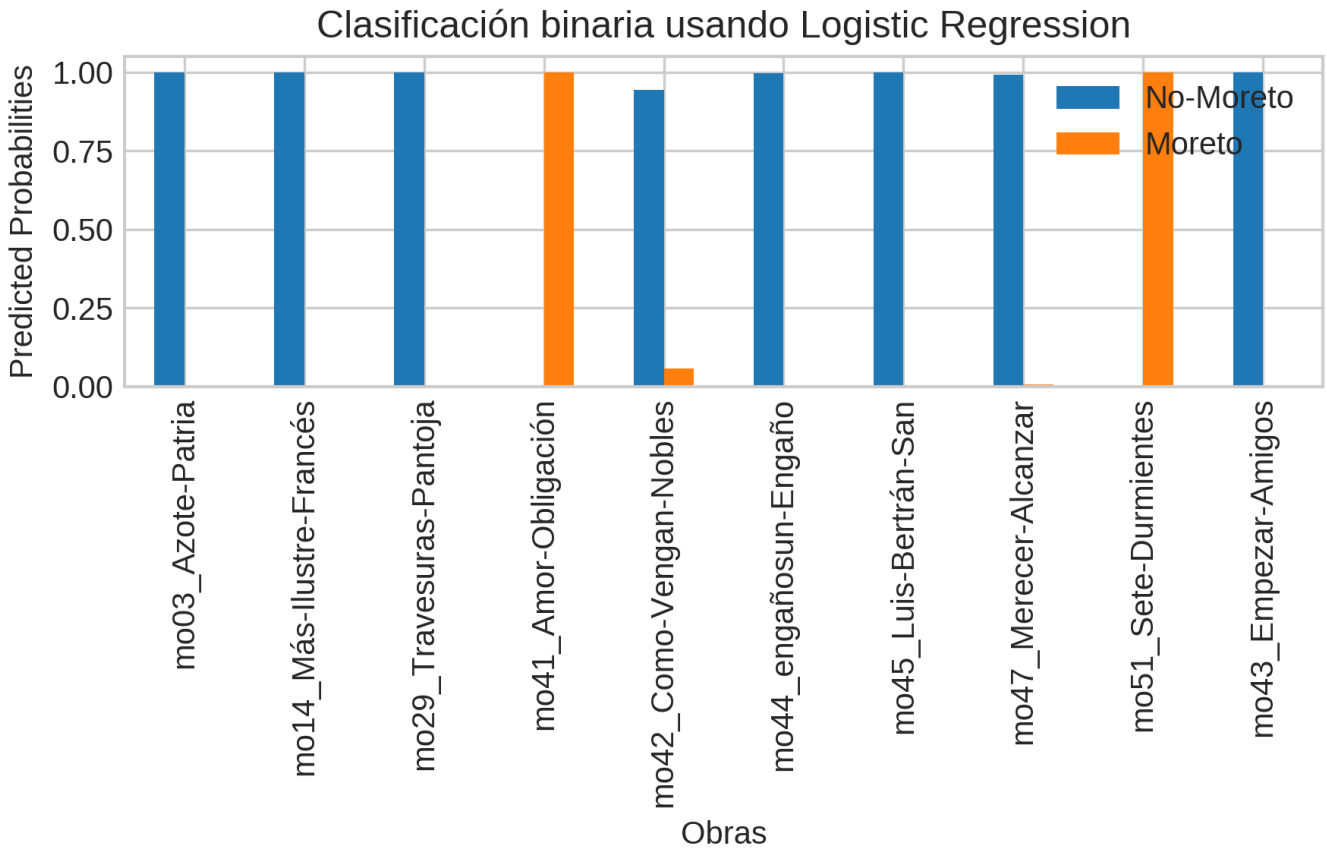
En este caso los resultados se modifican: solo dos obras aparecen clasificadas como de Moreto: *Amor-Obligación* y *Siete-Durmientes*. El resto de obras aparecen asignadas como No-Moreto. Como vemos los resultados de este paso no son del todo coherentes con los del anterior: tres textos han dejado de ser señalados como de Moreto. La razón posiblemente está en que en el anterior paso el algoritmo no tenía suficiente material como para aprender las características exactas de todos los autores, mientras que Moreto está representado por un conjunto amplio de textos. Por eso he realizado dos pasos más.

En primer lugar he analizado de nuevo la pregunta con un corpus de nuevo balanceado por autores en el que cada uno aparece representado solo por tres textos. Estos son los resultados tanto para la clasificación multiclase como binaria:

Texto discutido	Clase binaria predecida	Autor predecido
mo03_Azote-Patria	No-Moreto	Tirso
mo14_Más-Ilustre-Francés	No-Moreto	Calderón
mo29_Travesuras-Pantoja	No-Moreto	Gómez
mo41_Amor-Obligación	**Moreto**	**Moreto**
mo42_Como-Vengan-Nobles	No-Moreto	**Moreto**
mo43_Empezar-Amigos	No-Moreto	Calderón
mo44_engañosun-Engaño	No-Moreto	Tirso
mo45_Luis-Bertrán-San	No-Moreto	Tirso
mo47_Merecer-Alcanzar	No-Moreto	Lope
mo51_Sete-Durmientes	**Moreto**	**Moreto**

De nuevo en este corpus dos de los textos vuelven a ser categorizados como de Moreto en ambos análisis. De otras tres obras que habían sido categoriazadas como Moreto en el análisis multiclase, ahora solo una continúa como de Moreto: *Como-Vengan-Nobles*; las otras dos han pasado a ser señaladas como de otros autores.

Finalmente he utilizado otro algoritmo de clasificación, *Logistic Regression* (usado en Riddell and Schöch 2014), que no solamente señala la categoría que predice, sino además también señala las probabilidades por cada categoría. De esta manera podemos observar la diferencia de probabilidades por cada clase:



Como se puede observar, el único textos cuyas probabilidades no están de manera absolutamente contundente a favor o en contra de ambas clases es *Como-Vengan-Nobles*, cuyas probabilidades de que no sea de Moreto es de 0.94, mientras que las probabilidades de que sea de Moreto son de 0.06. Obsérvese que esta misma obra es la que en ambos intentos de clasificación multiclase han señalado como de Moreto, aunque los binarios lo hayan descartado. Los otros 9 textos discutidos son claramente clasificados ya sean como de Moreto (*Amor-Obligación* y *Siete-Durmientes*) o no (resto). El único texto cuyas probabilidades se sitúan por encima del intervalo de confianza (estándar en ciencias sociales y humanas) de 0.05 es *Como-Vengan-Nobles*, aunque todas las evidencias que tenemos señalan que no sería Moreto su autor.

5. Conclusiones

En este informe se han analizado mediante diferentes técnicas la verificación de autoría de 10 obras discutidas de Moreto, una cantidad excepcional de obras para un único trabajo. Para ello no solo se han realizado las tareas habituales de clustering y reducción de dimensionalidad, además se han evaluado los parámetros, se han realizado tareas de clasificación (multiclase y binaria) y finalmente se han observado las probabilidades de clasificación binaria utilizando Logistic Regression. Metodológicamente se ha observado que Eder's Delta tiene resultados ligeramente mejores que Cosine Delta y ambas mejores que la versión clásica de Delta para textos españoles teatrales del Siglo de Oro. Rangos de palabras entre 1000 y 3000 son en general los que mejores resultados aportan.

En cuanto a los resultados de autoría, de las 10 obras analizadas dos aparecen de manera coherente y continúa como de Moreto: *Amor-Obligación* y *Siete-Durmientes*. El resto de obras aparecen claramente como que no son de Moreto con una excepción: *Como-Vengan-Nobles*. Aunque la clasificación binaria lo señala como que no es de Moreto, las probabilidades aportadas por Logistic Regression, aunque notablemente altas, son más bajas que el resto y se encuentran por debajo del intervalo de confianza de 0.05. Sería necesario continuar investigando este caso, quizás con un estudio que busque los autores candidatos y que los represente de manera coherente.

Este trabajo no agota todas las posibilidades de análisis ni puede predecir las nuevas metodologías que se desarrollen en el futuro, mediante las cuales podrían ser necesario revisar este trabajo. Un aspecto que tanto el grupo de Moretianos como la comunidad de filólogos trabajando en el Siglo de Oro, especialmente en teatro, deben asumir es la creación de corpus homogéneos en los que se apliquen técnicas y estándares que permitan un trabajo limpio y sencillo al analizar los textos cuantitativamente: codificación de los textos en XML-TEI, utilización de identificadores compartidos (VIAF, Wikidata o BNE), identificadores unívocos propios para textos y autores, estandarización de metadatos en el encabezamiento de las obras, utilización de herramientas de control de versiones y de archivación, etcétera. Las ediciones de los textos son la base de nuestros trabajos, por lo que debemos buscar tener unas bases sólidas accesibles al resto de la comunidad investigadora.

6. Bibliografía

Baayen, R Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. 1. publ. Cambridge: Cambridge University Press.

Burrows, John. 2002. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (3): 267–87. https://doi.org/10.1093/llc/17.3.267.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 16 (1): 1–15.

Juola, Patrick. 2015. "The Rowling Case: A Proposed Standard Protocol for Authorship Attribution." *Digital Scholarship in the Humanities* 30 (suppl. 1): 100–113. https://doi.org/10.1093/llc/fqv040.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. "Authenticating the Writings of Julius Caesar." *Expert Systems with Applications* 63: 86–96. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2016.06.029.

Riddell, Allen, and Christof Schöch. 2014. "Progress through Regression." In *Digital Humanities 2014: Conference Abstracts*. Lausanne: UNIL/EPFL. http://dharchive.org/paper/DH2014/Paper-60.xml.

Smith, Peter W. H., and W. Aldridge. 2011. "Improving Authorship Attribution: Optimizing Burrows' Delta Method." *Journal of Quantitative Linguistics* 18 (1): 63–88. https://doi.org/10.1080/09296174.2011.533591.

Stamou, Constantina. 2008. "Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating." *Literary and Linguistic Computing* 23 (2): 181–99. https://doi.org/10.1093/llc/fqm029.