ISSN: 2254-4496 / Vol. 5 Nº 2 / Noviembre de 2016

{CARACTERES}

Estudios culturales y críticos de la esfera digital

En este número participan María Luisa Bellido-Gant, Anabel Fernández-Moreno, José Manuel Fradejas Rueda, Ricardo González García, David Hallberg, Almudena Mangas-Vega, Sergio Martínez Luna, Javier Merchán Sánchez-Jara, Alessandro Mistrorigo, David Ruiz-Torres, Vega Sánchez-Aparicio, María Simarro Vázquez.

EDITORIAL DELIRIO

Caracteres. Estudios culturales y críticos de la esfera digital

Caracteres es una revista académica interdisciplinar y plurilingüe orientada al análisis crítico de la cultura, el pensamiento y la sociedad de la esfera digital. Esta publicación prestará especial atención a las colaboraciones que aporten nuevas perspectivas sobre los ámbitos de estudio que cubre, dentro del espacio de las Humanidades Digitales. Puede consultar las normas de publicación en la web (http://revistacaracteres.net/normativa/).

ISSN: 2254-4496

Editorial Delirio (www.delirio.es)

EL ANÁLISIS ESTILOMÉTRICO APLICADO A LA LITERATURA ESPAÑOLA: LAS NOVELAS POLICIACAS E HISTÓRICAS

STYLOMETRIC ANALYSIS APPLIED TO SPANISH LITERATURE: HISTORICAL AND CRIME FICTION

JOSÉ MANUEL FRADEJAS RUEDA
UNIVERSIDAD DE VALLADOLID

ARTÍCULO RECIBIDO: 14-07-2016 | ARTÍCULO ACEPTADO: 15-10-2016

RESUMEN:
En este artículo se trata de mostrar si un ordenador es capaz de determinar la autoría de un texto. Para ello se ha creado un corpus de 122 novelas contemporáneas (69 de tema histórico, 50 policiacas y 3 del oeste) y se han analizado con el paquete de análisis estilométrico *stylo*. De todos los análisis que ofrece este paquete, escrito en R, se ha utilizado el más sencillo: el análisis de grupos. Los resultados han sido muy interesantes ya que con un mínimo de 100 palabras (las más frecuentes) el ordenador ha sido capaz de agrupar, sin error alguno, las distintas obras de cada autor y ha sabido asignar al autor real aquellas que se publicaron bajo seudónimo.

ABSTRACT:
This paper demonstrates that a computer can determine the authorship of a text. To this end we created a corpus of 122 contemporary novels written in Spanish (69 historical novels, 50 crime novels, and 3 westerns). The corpus was then studied using *stylo*, a stylometric analysis package written in the programming language R. We chose to apply the simplest of the multiple types of analysis offered by this package: cluster analysis. The results are very interesting: by taking into account just the 100 most frequently used words (MFW), the computer was able to group the different works of each author as well as

assigning those published under a pseudonym to the true author without incurring in any errors.

PALABRAS CLAVE:
Estilometría, atribución de autoría, R, novela histórica, novela policiaca

KEYWORDS:
Stylometry, authorship attribution, R, historical fiction, crime fiction

---

José Manuel Fradejas Rueda. Catedrático de Filología Románica en la Universidad de Valladolid. Su investigación se ha centrado en la crítica textual neolachmanniana, la lingüística histórica y las tecnologías de la información y comunicación aplicadas a la investigación de la lengua y literatura españolas, en especial en la edición digital de textos medievales, los lenguajes de marcas (XML-TEI) y la estilometría (en R).

1. Introducción

Los problemas de autoría, bien porque una obra nos ha llegado anónima, bien porque se ha publicado bajo seudónimo, es una de las tareas a las que los especialistas en literatura se han aplicado desde la noche de los tiempos. Usualmente, la identificación se ha realizado por medio de métodos cualitativos, hermenéuticos, como es determinarlo por cuestiones de índole textual como pueden ser la semejanza y giros peculiares de la lengua, por el tema tratado, por la métrica, por las figuras retóricas que emplea, por el estilo, como por circunstancias extratextuales como puede ser la historia del texto, fuentes secundarias, citas, cartas, estudios de archivo y biblioteca, contexto de la época… (Gil-Albarellos, 2010: 343) y ahí es donde se han originado ciertas disputas académicas[1]. Dentro de la literatura española hay dos casos paradigmáticos: el autor del *Lazarillo* y quién es realmente Alonso Fernández de Avellaneda. No se ha llegado a ninguna conclusión aún y, quizá, nunca se llegue puesto que no tenemos acceso a más pruebas que los textos mismos y en ellos se buscan con lupa y gran cuidado expresiones que tiene el texto en cuestión y que se documentan en otros textos del autor al que se le quiere atribuir o, en algunos casos, para rechazarla. Así, por ejemplo, Madrigal (2003) establece que el autor de *La tía fingida* es Miguel de Cervantes, y para tal demostración se basa en unas pocas expresiones que solo encuentra en otras obras de Cervantes, es decir, que tienen «una equivalencia verbal con la obra de Cervantes». Es el mismo tipo de análisis que hace Martín

[1] Una relación y breve exposición de algunos de los métodos más usuales se puede ver en Gil-Albarellos (2011).

Jiménez (2007) para establecer que Alonso Fernández de Avellaneda es Jerónimo de Pasamonte. No discuto ninguna de estas dos atribuciones, no soy quién para ello, pero ciertos análisis parecen poner el «Avellaneda» más cerca de Tirso de Molina (Madrigal, 2005) que de la *Pícara Justina* (Blasco Pascual, 2005) y de la *Vida y trabajos de Jerónimo de Pasamonte* (Martín Jiménez, 2007); lo más curioso es que esos análisis sitúan lingüísticamente el «Avellaneda» con *Los trabajos de Persiles y Sigismunda*, las *Novelas ejemplares* y *La Galatea*, es decir, lo alinea con Cervantes. ¿Qué pasaría si Cervantes hubiera escrito el «Avellaneda»?, que es la sorpresa que López expresa al final de su trabajo sobre *La tía fingida* (López, 2011: 36).

Insisto, no es el objetivo de este trabajo demostrar quién pudiera estar tras el nombre de Alonso Fernández de Avellaneda. Nada más lejos. Tan solo experimentar con unas técnicas de investigación que son productivas y eficaces y que pueden ser de gran valor a la hora de determinar el probable autor de un texto dado.

2. La estilometría

Uno de los grandes debates de la literatura inglesa es la llamada *Shakespeare Authorship Question*, o lo que es lo mismo: la disputa de si William Shakespeare escribió a William Shakespeare y cuál es, en realidad, el canon de las obras shakespearianas. Es una debate que surgió a mediados del siglo XIX y cuenta con numerosos candidatos, pero de entre todos ellos destacan tres: Edward de Vere, XVII Conde de Oxford; Francis Bacon y Christopher Marlowe (Craig, 2009). Este debate se sitúa dentro de los problemas de atribución de autoría y desde su

concepción a mediados del siglo XUX se ha tratado de resolver por medio de técnicas estilométricas, es decir, por medio del análisis estadístico del estilo literario (Holmes, 1998: 111; 1999: 8378) y para Juola *estilometría* es un *near-synonymous* de atribución de autoría (2006: 238).

En la bibliografía científica española el término *estilometría* aparece por primera vez en Montoya Martínez y Rubio Flores (1994), pero no tiene el mismo sentido, pues ese trabajo, sobre la metáfora y la comparación en la segunda partida alfonsí, lo único que ha hecho ha sido "analizar y cuantificar las comparaciones […] por medio de medios electrónicos, [de] todos los elementos comparativos más comunes de la época […]: *asi*, *tanto*, *tan*, *tal* y formas del verbo *ser*, combinadas con la partícula *como*" (Montoya Martínez & Rubio Flores, 1994: 158). En verdad, lo único que consiguieron fue localizar los datos con un programa llamado WordCruncher y presentar algunas estadísticas descriptivas básicas: tokens (palabras del texto), tipos (palabras únicas), caracteres totales, tamaño total –no indican la magnitud– y el número de frases localizadas y en qué título de la segunda partida (Montoya Martínez & Rubio Flores, 1994: 160). Muchos años más tarde, la emplea Frías Delgado (2009), quien utiliza NLTK (Natural Language Tool Kit –(Loper, Bird & Klein, 2009)–) y algunos *scripts* en Python para mostrar que la longitud de las palabras (por número de letras) es un factor estable en la lengua española desde el punto de vista diacrónico, que podría (no lo demuestra) servir para discriminar géneros pero que no es suficiente para discriminar autorías. Posteriormente aparece como palabra clave en Troya Déniz (2015) en un análisis de variación ideolectal –aparición y uso

de *quizá(s)* y *tal vez*– en novelas escritas en español a principios del siglo XXI (entre 2001 y 2014)[2].

La primera aproximación al problema fue la de T. Mendenhall (1901) que se reflejó en «A Mechanical solution of a literary problem». La solución mecánica consistió en contar las letras de cada una de las palabras de las obras de Shakespeare y comparar la longitud de las palabras de las obras atribuidas a Shakespeare con los sospechosos habituales: de Vere, Bacon y Marlowe[3].

El procedimiento era muy sencillo: una persona leía una palabra, contaba el número de letras y lo anunciaba en voz alta; otra apretaba el botón adecuado (uno para cada número) en una máquina registradora construida al efecto. Procedieron así a lo largo de dos millones de palabras (400 000 eran de Shakespeare). La conclusión a la que llegaron era que la longitud de palabra más usual en Shakespeare era de cuatro letras, «a thing never met with before» (Mendenhall, 1901: 102).

El laborioso procedimiento de Mendenhall debió influir en que nadie se ocupara de los problemas de estilometría, aunque tuvo una secuela en el análisis en el que Brinegar (1963) trató de establecer si Mark Twain escribió las diez cartas Snodgrass que se publicaron en el *New Orleans Daily Crescent* en 1861 bajo la firma de Quintus Curtius Snodgrass y que describían, de manera humorística, las aventuras y experiencias del autor en Nueva

---

[2] En Dialnet (https://dialnet.unirioja.es/servlet/articulo?codigo=4533246, acceso 17-06-2016) se localiza otro artículo en el que aparece el término estilometría pero que no es sino la versión al español (Merriam, 2013) del artículo Thomas Merriam publicado en *Notes and Queries*, 241 (1997), por lo que no se puede tener en cuenta. Asimismo, se mencionan dos tesis doctorales realizadas en las universidades politécnicas de Cataluña y Valencia.

[3] Estos recuentos también los hicieron con M. Cervantes, A. Dumas, J. V. von Scheffel, A. Boito, Julio César y Eurípides (Mendenhall, 1901: 100-101, figs. 2 y 3).

Orleans, Baton Rouge y Washington como soldado confederado durante la Guerra Civil americana (1861–1865). Este problema de autoría se planteó porque entre los muchos seudónimos de Mark Twain —ya de por sí un seudónimo de Samuel Langhorne Clemens (1835–1910)— estaba el de Thomas Jefferson Snodgrass. La conclusión a la que llegó es que Twain no fue el autor.

Sin embargo, la verdadera aplicación de la estilometría y su primer gran éxito surgió a principios de la década de 1960 con el llamado caso de los *Federalist Papers*.

Los *Federalist Papers* es una serie de 85 artículos y ensayos que fueron publicados bajo el seudónimo Publius en los periódicos de Nueva York en 1787 para persuadir a los norteamericanos de que ratificaran la nueva constitución. Se sabe que los escribieron Alexander Hamilton (51), James Madison (14) y John Jay (5)[4] y otros tres fueron coescritos por Madison y Hamilton. Sin embargo, doce de ellos los reclamaban como propios tanto Hamilton como Madison. Mosteller y Wallace (1964), basándose en la frecuencia de uso de palabras gramaticales, *function words*, es decir de artículos, conjunciones, preposiciones, pronombres y ciertos adverbios, adjetivos y verbos auxiliares, como discriminadores de estilo establecieron la autoría de cada uno de los doce ensayos disputados. Así, pudieron ver, por ejemplo, que la preposición *upon* aparecía 3.24 veces por cada 1 000 palabras en los escritos de Hamilton frente a 0.23 en los de Madison. En cambio, este prefería

[4] Alexander Hamilton (1755–1804) es uno de los padres fundadores de los Estados Unidos de América y creador del sistema financiero estadounidense. James Madison (1751-1836) fue el cuarto presidente de Estados Unidos (1809–1817) y un teórico político. John Jay (1745–1829) fue uno de los signatarios del Tratado de París (1783) por medio del cual se dio por concluida la guerra de independencia americana y también fue el primer presidente del Tribunal Supremo de Estados Unidos.

la palabra *whilst* frente *while*, que era la favorecida por Hamilton. La conclusión a la que llegaron es que esos doce artículos fueron escritos por James Madison.

Esto llevó a la constatación de que pueden ser mucho más interesantes las palabras gramaticales, las *function words*, para establecer la huella lingüística de un autor que las palabras semánticas ya que las gramaticales no dependen del contexto, del tema ni del género y, además, las palabras gramaticales se usan de manera inconsciente, con lo que son más capaces de atrapar las selecciones estilísticas de los diferentes autores (Stamatatos, 2009: 540), aunque se puede *jugar* con muchos marcadores de estilo como la longitud de palabras[5], la de las oraciones, el número de sílabas por palabra, la distribución de letras, la distribución de *n-grams*, las colocaciones, la distribución de las partes de la oración, la ratio tipo-token, la distribución del vocabulario, etc. (Holmes, 1994).

En todos los trabajos que sobre estilometría se han realizado desde Mosteller y Wallace hasta principios del siglo XXI los investigadores se han ayudado de los ordenadores (*computer-assisted*), pero no los han realizado con ordenadores (*computer-based*), con lo que ha habido grandes limitaciones (Stamatatos, 2009: 538). A pesar de ello, ha habido grandes éxitos como fue la identificación del autor de la novela *Primary Colors: A Novel of Politics*, que se publicó 1996 anónimamente. Esta novela es un *roman a clef* en el que se describe la primera campaña presidencial de Bill Clinton en 1992. Posteriormente, gracias al análisis estilométrico realizado por Donald Foster, especializado en el

[5] De acuerdo con el análisis de Frías Delgado (2009) la longitud de palabras no es un discriminador de autoría.

debate de la *Shakespeare Authorship Question*[6], se estableció que el autor era Joe Klein, columnista del *Newsweek*.

En el año 2006 se presentó un prototipo computerizado de atribución de autoría, el *Java Graphical Authorship Attribution Program* (JGAAP), desarrollado por Patrick Juola en el Evaluation Variation in Language Laboratory (Juola, Sofko & Brennan 2006) y se ha puesto a prueba y ha demostrado su fiabilidad en el caso Robert Galbraith (Juola, 2013a, 2013b, 2015).

Robert Galbraith, un policía militar jubilado y con experiencia en la industria de seguridad, publicó en abril de 2013 la novela policiaca titulada *The Cuckoo's Calling* (*La llamada del Cuco*). Esta novela «was lavishly praised by critics» (*Sunday Times*, 14.7.13). Según contaba *The Sunday Times*, a uno de sus periodistas le pareció una novela demasiado buena como para ser una obra primeriza y que un autor con la formación que decía tener describiese con sumo detalle la ropa femenina, por lo que decidió solicitar los servicios de Juola para averiguar quién pudiera ser el autor real. La verdad es que tenían un soplo: que Robert Galbraith era un seudónimo de J. K. Rowling, con lo que pudo partir de una hipótesis que se trató de confirmar.

El procedimiento de Juola (2013b) fue seleccionar la única novela para adultos escrita por J. K. Rowling, *The Casual Vacancy*, y otras tres novelas policiacas escritas por mujeres: *The St. Zita*

---

[6] Su tesis doctoral trató de establecer si «Funeral Elegy [for] William Peter», firmada con las iniciales W. S. e impresa por Georg Eld para el librero londinense Thomas Thorp, las mismas personas que se hicieron cargo de la publicación de los sonetos de Shakespeare en 1609, había sido escrita por Shakespeare. No llegó a una conclusión definitiva, tan solo lo sugirió (Foster, 1989). Años más tarde se incluyó en tres ediciones estándar de Shakespeare, pero en 2002 Foster reconoció que se había equivocado (Niederkorn, 2002).

Society de Ruth Rendell, *The Private Patient* de P. D. James y *The Wire in the Blood* de Val McDermid para ver cuál era más similar a Galbraith y realizó cuatro análisis centrados en cuatro variables lingüísticas:

1. distribución de la longitud de las palabras
2. uso de las 100 palabras más comunes
3. distribución de 4-gram (grupos de cuatro letras juntas, pueden ser palabras, parte de una palabra o de dos palabras adyacentes)
4. distribución de bigramas (qué dos palabras aparecen juntas)

La conclusión a la que llegó es que de los cuatro autores, dos quedaban descartados y que «The only person consistently suggested by every analysis was Rowling, who showed up as the winner or the runner-up in each instance» (Juola, 2013b). Ante estos datos, el periodista de *The Sunday Times* preguntó a J. K. Rowling y esta confesó que era ella la autora.

Como puede verse, los métodos informáticos, ya sean *computer-assisted* (Foster) y sean *computer-driven* (Juola) pueden ser de gran ayuda para establecer la autoría de un texto anónimo o publicado anónimamente, pero la confirmación solo ha sido posible cuando los autores han confesado que efectivamente eran ellos los autores. Un caso interesante en el mercado editorial actual sería responder a la pregunta ¿Quién es Elena Ferrante?[7]

---

[7] Durante el proceso de maquetación de estas páginas saltó a los medios de comunicación que Elena Ferrante es la traductora Anita Raja. Ahora que ya hay una posible candidata sería el momento de proceder a los análisis estilométricos para establecer si Anita Raja es Elena Ferrante. El problema: no hay obras escritas por Anita Raja con las que comparar las novelas de Ferrante. Luego el misterio continúa.

3. Análisis de grupos

Como se ha señalado, hay varias técnicas informáticas que pueden ayudar a la hora de determinar la autoría de una obra y muchas de ellas están al alcance de cualquier investigador. En el resto de este trabajo voy a exponer un experimento para determinar la autoría de un corpus de novelas españolas contemporáneas por medio del análisis de grupos (también llamado de conglomerados o agrupamiento, en inglés *cluster analysis*).

El análisis de grupos es una técnica de análisis estadístico multivariante cuya finalidad es agrupar una serie de elementos en grupos de manera que se dé la máxima homogeneidad posible dentro de cada grupo y, a la vez, la mayor diferencia entre los diversos grupos. Por lo general, representan los resultados por medio de dendrogramas[8].

Las primeras aproximaciones a este tipo de análisis las realicé a la vista del libro de Jockers (2014), en el que, como un ejercicio para el aprendizaje del lenguaje de programación R para el análisis de textos literarios, propone determinar si un texto marcado como anónimo se agruparía de manera *natural* con algún autor de un grupo de textos de novelistas irlandeses (o de origen irlandés) de los siglos XVIII-XX. A la vista de que *funcionaba* con los textos ingleses, preparé un pequeño corpus de dieciocho novelas en español del siglo XIX (Eduardo Acevedo Díaz [3], Benito Pérez

[8] Un *dendrograma* es una representación arbórea de los datos. En gran medida es semejante a los *stemma* utilizados en crítica textual o a los *cladogramas* que se emplean en biología; solo que nos referimos a estos esquemas como dendrogramas cuando son el resultado de la aplicación de un algoritmo de agrupación jerárquica.

Galdós [8], Juan Valera [3], José María de Pereda [4])[9]. Uno de los ficheros se etiquetó como anónimo (anonimo.xml). Este corpus está constituido por 1 361 448 palabras token y 244 690 palabras tipo[10] y tan solo seleccionando 24 palabras tipo *–a, al, como, con, de, del, el, en, la, las, le, lo, los, más, me, no, para, por, que, se, su, un, una* e *y*–, aquellas cuya frecuencia relativa de aparición fuera >= .5, el análisis de agrupación por medio de la medida de la distancia euclidiana los reunió sin error alguno, como puede verse en el dendrograma correspondiente (fig. 1) y estableció que el texto marcado como anónimo era de Pérez Galdós, como efectivamente lo era. Se trataba del Episodio Nacional *Trafalgar*[11].

[9] Los textos se han extraído de las versiones ePub creadas por la Biblioteca Virtual Cervantes (http://www.cervantesvirtual.com) y distribuidas con el e-reader Papyre 6.1. Posteriormente se etiquetaron sucintamente de acuerdo con el sistema de marcado de la Text Encoding Initiative (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/) según el siguiente esquema: cada capítulo constituye un , el título y número de capítulo se marca con <head>, y el texto con <p> para cada párrafo, entendiendo por párrafo cada vez que hay un punto y aparte. No se ha tenido en cuenta ninguna otra característica gráfica. Evidentemente, esta colección requiere una revisión ulterior.

[10] Se entienden por *palabras token* todas y cada una de las palabras que constituyen un texto, con independencia de cuántas veces aparezca cada una de ellas. En cambio, las *palabras tipo* son cada una de las palabras diferentes que conforman el texto. Así, en el enunciado inicial del *Quijote* («En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor») hay 33 palabras-token, pero solo 27 palabras-tipo puesto que la preposición *de*, que aparece cuatro veces (tokens), solo se cuenta como un único tipo y lo mismo sucede con *en*, *un* y *no*, todas ellas aparecen dos veces, pero tan solo constituyen un tipo cada una de ellas. Por otra parte, a efectos de estos análisis, se entiende por *palabra* la secuencia de caracteres alfabéticos o numéricos entre dos espacios o delimitada por un signo de puntuación. Véase también la nota 16.

[11] Podría haber complicado el problema si hubiera incluido *Trafalgar* de José Luis Corral (2001), pero, como se verá, no habría cambiado el resultado.

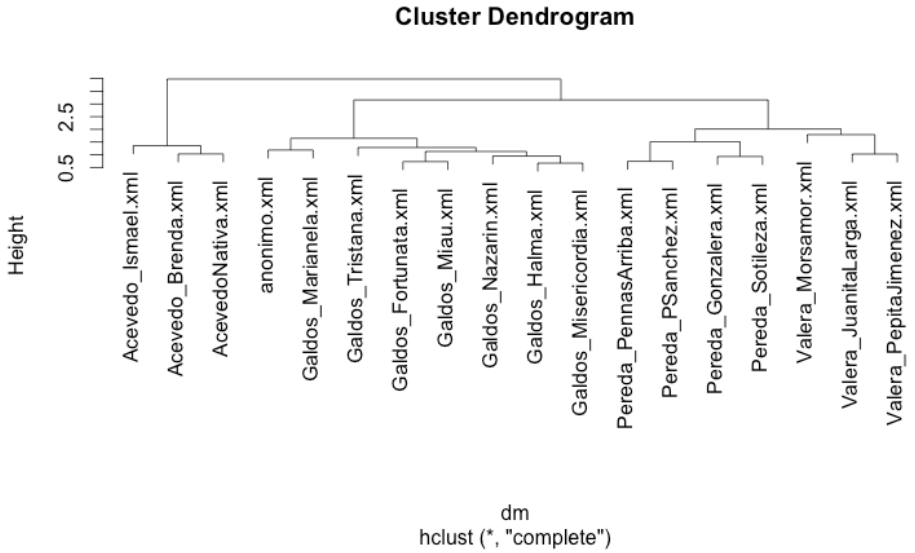

Figura 1. Agrupación de las 18 novelas, frecuencia >= .5

Ante este alentador resultado, se decidió incorporar cuatro novelas más. Para este nuevo experimento, el autor seleccionado fue Arturo Pérez-Reverte (*El Asedio*, *Cabo de Trafalgar*, *El Húsar* y *Hombres buenos*). Aunque está bastante alejado temporalmente de los anteriores (nacidos entre 1824 y 1851; fallecidos entre 1905 y 1921), podría encerrar un cierto problema y provocar un posible error de atribución ya que dos de ellas, una de Pérez Galdós y otra de Pérez-Reverte, tratan el mismo tema: la batalla naval de Trafalgar (21 de octubre de 1805).

La adición de las cuatro novelas de Pérez-Reverte ha aumentado el volumen de palabras-token a 1 874 532 y a 297 841 el de las palabras-tipo. De nuevo, tan solo 24 de ellas, las mismas que en el caso anterior, han permitido agrupar sin error alguno las 22 novelas (fig. 2). La igualdad de argumento en Pérez Galdós y Pérez-Reverte no ha supuesto ningún problema.

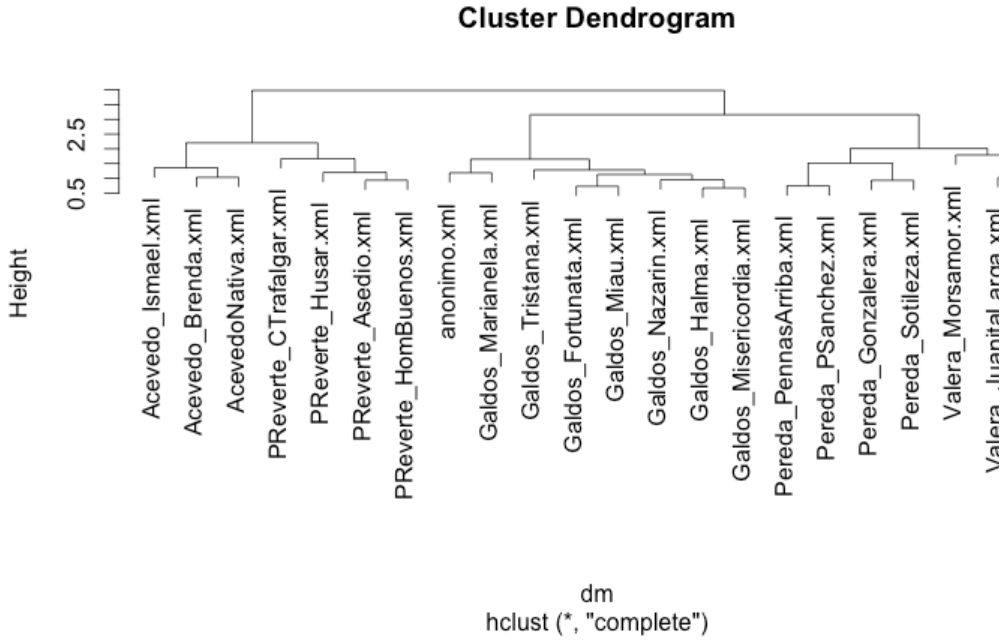

Figura 2. Agrupación de las 22 novelas, frecuencia >= .5

A la luz de los ensayos anteriores, propuestos como ejercicios para el aprendizaje de la programación en R, se procedió a diseñar un nuevo experimento con el paquete *stylo* (Eder, Rybicki & Kestemont, s. f.). Ya no se trata de unas pocas líneas de código para aprender a programar y analizar textos, sino de todo un paquete diseñado para el análisis estilométrico, en el mismo lenguaje que las pruebas que acabo de exponer.

El paquete *stylo* es una aplicación escrita en el lenguaje R desarrollado por Eder, Rybicki y Kestemnot (Eder, Rybicki & Kestemont, s. f.; 2013) para el análisis estilométrico, o estilística computacional, relativamente sencillo de usar puesto que dispone de una interfaz gráfica (GUI) que permite realizar varios tipos de análisis con solo unos clics, desde cargar los textos hasta obtener sofisticados gráficos. Tras esta interfaz se esconde un conjunto funcional de los principales métodos de clasificación que dominan

el ámbito de la estilometría. En este estudio solo se hace uso de uno de los posibles análisis que permite el paquete *stylo*: el análisis multivariado sin supervisión que ofrece como resultado un dendrograma en el que aparecen agrupados los textos, es decir, un análisis de grupos[12].

En esta ocasión el corpus se ha construido con novelas históricas y policiacas contemporáneas (véase en los apéndices I y II el elenco). Se han elegido estos dos temas porque presentan un rico panorama para comprobar problemas de autoría, puesto que hay obras escritas bajo seudónimo (dentro de las novelas históricas Peter Harris es el seudónimo de José Calvo Poyato; en el ámbito de la novela negra o policiaca contamos con Francisco González Ledesma quien, asimismo, escribió novelas de intriga policial como Enrique Moriel y novelas del oeste bajo el alias de Silver Kane)[13] y también hay novelas escritas a cuatro manos (Ángeles de Irisarri y Magdalena Lasala, y Ángeles de Irisarri y Martínez de Lezea). Eso es un terreno inmejorable para probar este software y alguna de las técnicas estilométricas que ofrece ya que sabemos las respuestas: quienes son los autores reales y sabemos de antemano cuál ha de ser el resultado[14].

La construcción del corpus ha sido relativamente sencilla. En la red hay numerosas copias en formato electrónico, por lo normal

---

[12] La versión de *stylo* empleada ha sido la 0.6.3 y se ha ejecutado con la versión 3.2.2 de R en un Apple iMac (21,5", con procesador Intel Core 2 Duo @ 3,06 GHz y 4 GB de memoria RAM) bajo OS X Yosemite (OS X 10.10.5).

[13] Francisco González Ledesma (1927-2015) ha utilizado otros seudónimos como Taylor Nummy, Silvia Valdemar o Rosa Alcázar, bajo este último ha escrito novelas románticas.

[14] Podríamos llevarnos una sorpresa: que alguna novela no haya sido escrita por quien dice que la ha escrito, sino que hubiera sido de un autor fantasma (*ghostwriter*). Pero no ha sido el caso.

en formato ePub, de casi cualquier título imaginable. Algunos están mejor construidos que otros. Puesto que se trata de ePub se procedió a su desempaquetado y a la extracción del texto de las novelas en texto plano. Se editaron para eliminar todo lo accesorio: paratextos, apéndices, notas al pie, aclaraciones, epígrafes, dedicatorias. Todos los textos presentan algunos problemas menores, pero no despreciables, como es el errático uso de los separadores de los millares (la norma actual indica que debe ser con un espacio, la tradicional con un punto[15]), puesto que el concepto palabra para un ordenador es la secuencia de caracteres entre dos espacios en blanco o entre dos caracteres no palabra (*non-word character*)[16], los separadores ya fuera el punto (tradicional) o el espacio en blanco podía convertir una cifra en dos o tres palabras distintas. Por otra parte, en el corpus histórico hay un elevado número de casos en los que aparecen nombres árabes con el artículo prefijado (577 casos diferentes con 1 812 ocurrencias) en los que el guion que separa el artículo de la palabra, se ha sustituido por un guion bajo (*underscore*); no hacerlo así habría aumentado escandalosamente el número de casos de la amalgama de la preposición *a* y el artículo *el* (53 282 casos en el corpus histórico) y habría podido producir algún sesgo en el análisis estadístico.

---

[15] Según las normas ortográfica de las RAE y de acuerdo con las normativas internacionales, el separador de millares consiste en «introducir un pequeño espacio en blanco, lo que se conoce en tipografía como espacio fino» (663), y «no deben utilizarse ni el punto ni la coma para separar los grupos de tres dígitos en la parte entera de un número» (664). Otro problema lo constituyen las fechas separadas por puntos (1.10.15) y los números de teléfono.

[16] Esto es algo sutil. Se consideran caracteres palabra cualquier carácter que sea una letra (con y sin diacríticos, de cualquier alfabeto), cualquier número (0 a 9) y el guion bajo (*underscore*), todo lo demás es *no palabra*.

Con un sencillo *script* de R se procesaron los 118[17] ficheros que constituyen ambos corpus y se han generado las listas de palabras que conforma cada una de las novelas con indicación del número de palabras-token, palabras tipo y la frecuencia relativa de los token[18] de cada una de las novelas. A partir de ellas se ha construido otras dos listas que contienen el total de palabras que constituye cada uno de los corpus analizados (histórica y policiaca) y cuyas estadísticas descriptivas resumidas se pueden ver en la tabla 1.

Histórica		**Policiaca**
69	textos	49
44 252 212	caracteres	22 732 809
7 659 249	tokens	4 010 728
122 038	tipos	91 789
195 805	párrafos	149 570
1.593342	Token-Tipo Ratio	2.288587
111 003.6	Media de tokens / novela	81 851.59
12 900.86	Media tipos / novela	10 645.39
2 837.754	Media párrafos / novela	3 052.449
641 336.4	Media caracteres / novela	463 934.9
39.11672	Media palabras/párrafo	26.81506
226.0014	Media caracteres / párrafo	151.9878
5.777618	Media caracteres / palabra	5.668001

Tabla 1. Estadísticas descriptivas resumidas

[17] En verdad son 122, pues al final se añadieron tres novelas firmadas por Silver Kane y una cuarta firmada por Enrique Moriel al elenco policiaco.

[18] No podemos facilitar acceso al corpus por evidentes razones legales, sin embargo, se puede acceder a todo el material procesado en <http://revistacaracteres.net/wp-content/uploads/2016/11/fradejas_datos.zip>. Véase al final de este artículo el apéndice III, titulado «Datos adjuntos» para una explicación detallada del contenido del fichero fradejas_datos.zip.

El análisis básico que se ha realizado ha sido un análisis de grupos (*cluster analysis*) y se ha ejecutado aplicándole varios métodos para calcular la distancia (Classic Delta, Argamon's Delta, Eder's Delta, Eder's Simple, Manhattan, Canberra, Euclidean y Coseno[19]) e inicialmente se han tenido en cuenta tan solo las 100 palabras más utilizadas (MFW *most frequent words*), aunque solo ofreceré los resultados obtenidos con Classic Delta.

4. Análisis del corpus de novelas históricas

El análisis de grupo de los 69 textos del corpus de novelas históricas, con la medida Delta clásica[20] y con tan solo las 100 palabras más frecuentes ha agrupado correctamente todos los autores (figura 3). Ha sido capaz de establecer que Harris y Calvo Poyato son el mismo autor. Los libros a cuatro manos de Irisarri y Lasala y de Irisarri y Martínez de Lezea los ha agrupado con los de Irisarri, pero formando un subgrupo diferente, aunque dentro de un grupo superior en el que se hallan Martínez de Lezea e Irisarri. Aquí lo *preocupante* es que la novela escrita por Irisarri y Lasala se encuentra muy alejada del grupo en el que se insertan los demás textos de Lasala.

[19] Tan solo enunciamos los métodos, quien esté interesado en la fórmulas matemáticas que subyacen véase Eder, Rybicki y Kestemont (2013: 15-17).
[20] Para una explicación sencilla de cómo se realiza este tipo de análisis veáse Calvo Tello (2016).

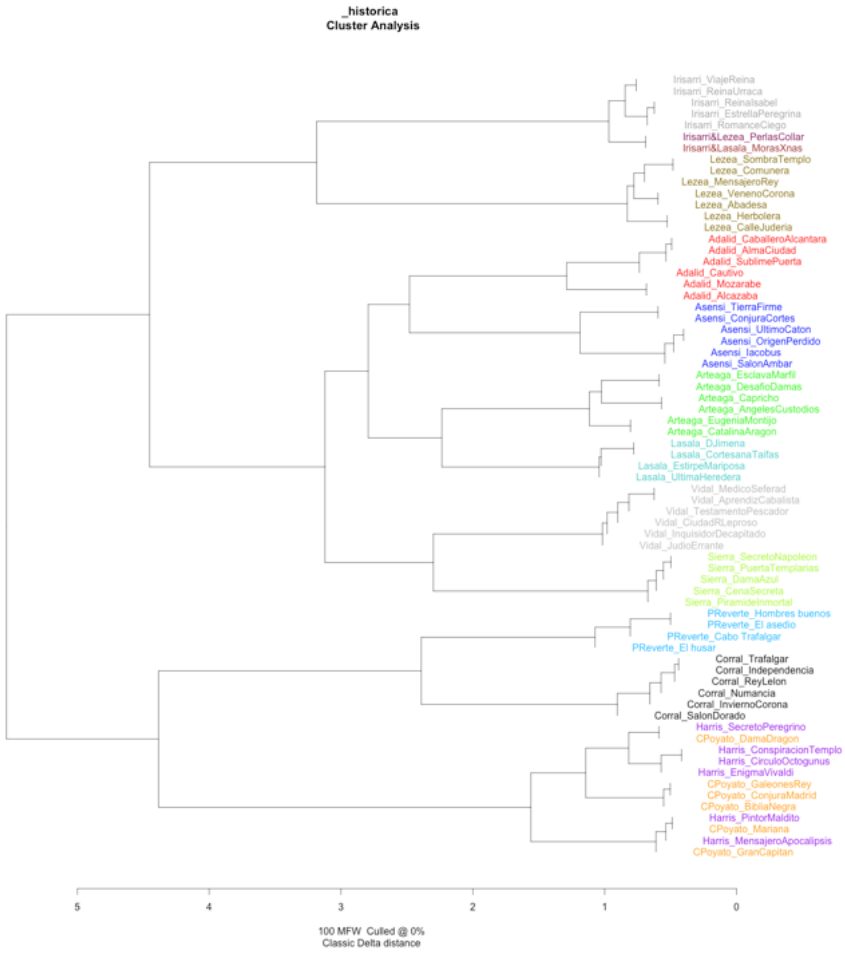

Figura 3. Dendrograma con 100 MFW y distancia Classic Delta con *stylo*

El mismo tipo de análisis, pero aumentando de las 100 a las 1000 palabras más frecuentes arroja el mismo resultado general: todos los libros se agrupan correctamente. Incluso el conjunto de Martínez de Lezea, Irisarri y las novelas escritas a cuatro manos. De nuevo Lasala está alejada del de Martínez de Lezea e Irisarri. Lo único que varía es la disposición y ordenamiento dentro del dendrograma (figura 4).

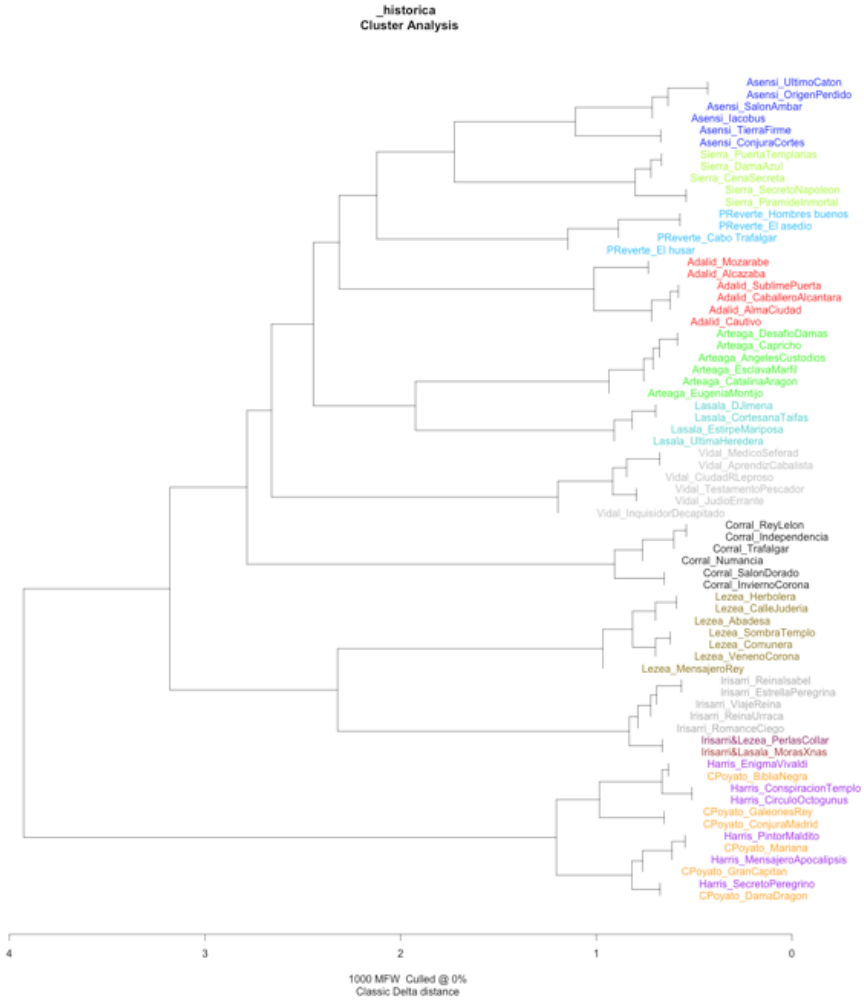

Figura 4. Dendrograma con 1000 MFW y distancia Classic Delta con *stylo*

Hay una pregunta que cabe hacerse: ¿cuál es el número mínimo y máximo de palabras necesario para que la clasificación sea efectiva? En los experimentos que hice con los novelistas españoles del siglo XIX bastó con 24 palabras[21]. En nuestro caso

[21] En aquel caso se utilizó la distancia Euclidiana. Utilizando esta medida los errores de agrupamiento son numerosos tanto con 25 como con 100 como con 1000 palabras. Según Eder, Rybick y Kestemont (2013: 15-16) la distancia Euclidiana, aunque es básica y la más natural, se debe evitar en los análisis estilométricos basados en la

con 25 palabras mínimas hay una cierta descolocación (figura 5). Básicamente todos los autores se agrupan correcta y consistentemente (Asensi, Pérez-Reverte, Vidal, Harris y Calvo Poyato), pero algunos se descolocan: Artega se entremezcla con Lasala y Vidal; dos obras de Sánchez Adalid se sitúan entre las de Corral; y una de Martínez de Lezea se acomoda con las de Pérez-Reverte.

frecuencia de palabras, salvo que estén normalizadas, que es el caso de la fórmula Delta introducida por Burrows (2002) o incluso la Delta Lineal de Argamon (2008) que no deja de ser una distancia Euclidiana aplicada la frecuencias de palabras normalizadas (*z-scored*). Pero no merece la pena introducirnos en las complejidades matemáticas que subyace en todo esto, como bien indica en varias ocasiones Juola (2006), pero puede verse una explicación muy sencilla en Calvo Tello (2016).

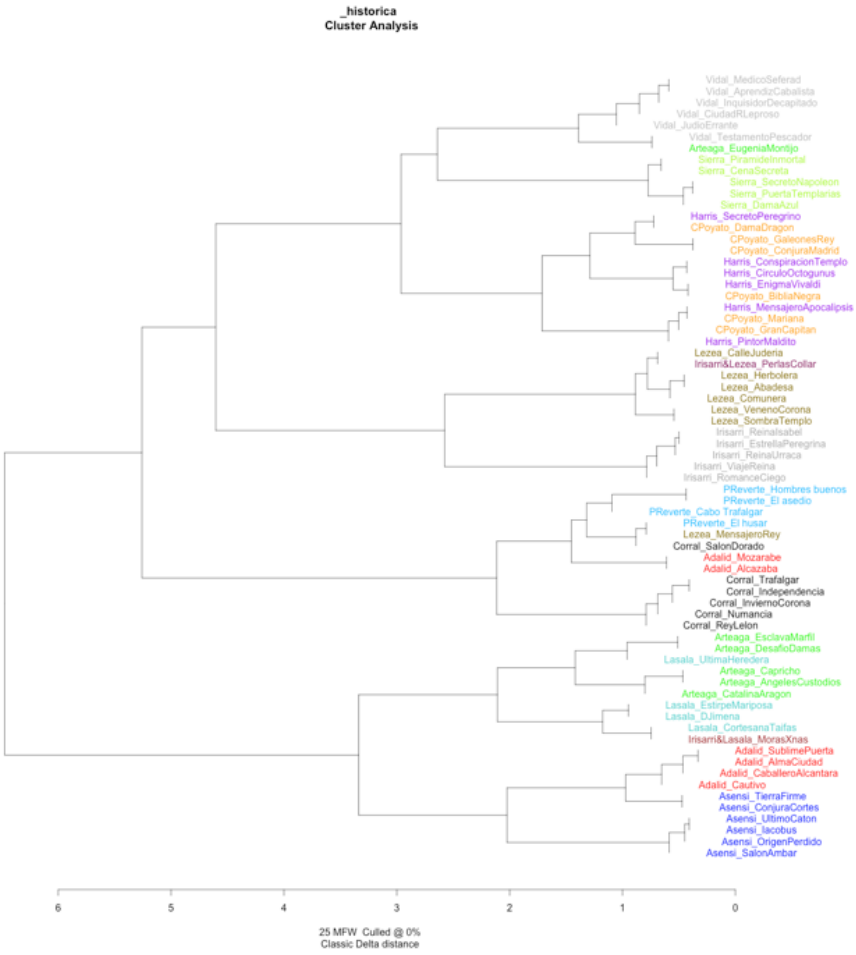

Figura 5. Dendrograma con 25 MFW y distancia Classic Delta con *stylo*

Lo más interesante en este caso es que la novela coescrita por Irisarri y Lasala se encuadra con Lasala y se aleja de las de Irisarri.

Con 50 palabras se obtiene una clasificación casi correcta (figura 6). La única nota curiosa es que *El húsar* de Pérez-Reverte se separa de las demás de este autor y se alinea con las del Corral, aunque se mantiene dentro del nodo en el que está todo Pérez-Reverte. La otra nota es que Irisarri y Lasala, como autoras independientes, se aproximan, mientras que Irisarri se aleja de Martínez de Lezea, que en todos los casos anteriores se

aproximaban; lo que, por otra parte, implica que la novela coescrita por Irisarri y Lasala se halla cómodamente asentada entre ellas.

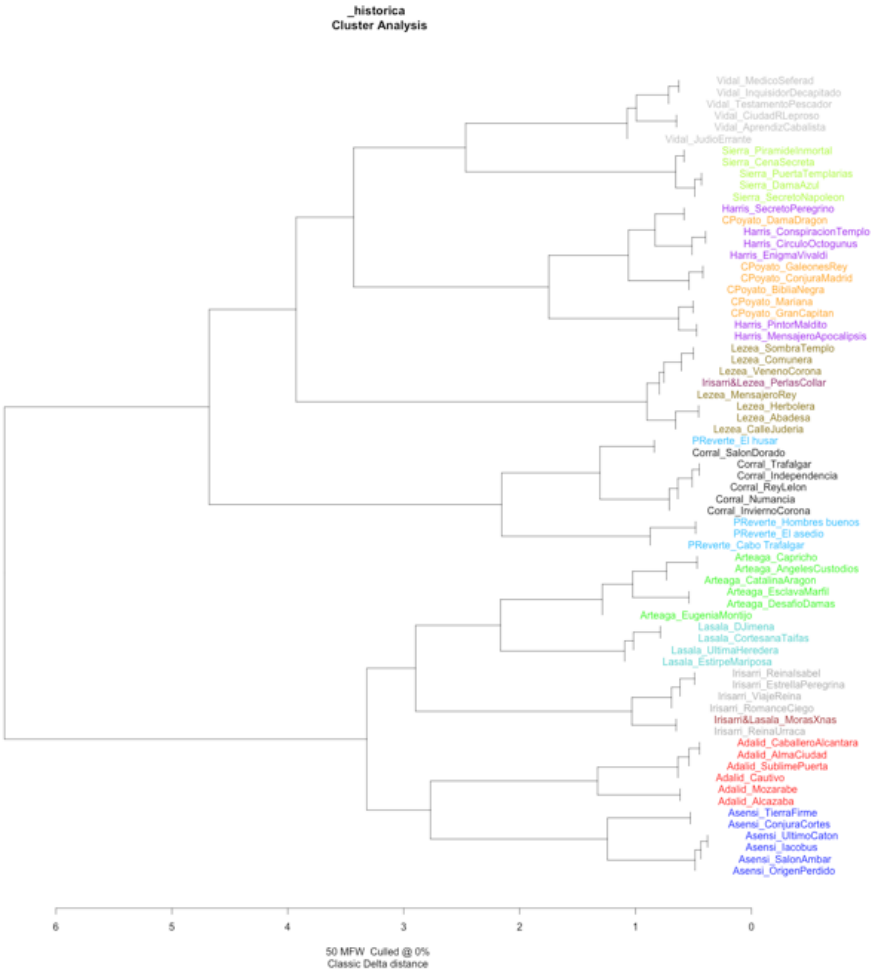

Figura 6. Dendrograma con 50 MFW y distancia Classic Delta con *stylo*

La conclusión parcial es que con un mínimo de 100 palabras la medida Delta clásica agrupa correctamente los textos. Al bajar de ese mínimo los problemas surgen y se van incrementando cuanto menor es el número de rasgos que se tienen en cuenta.

Otro rasgo para establecer estos agrupamientos es el de qué bigramas, es decir, qué dos palabras aparecen juntas con mayor frecuencia. De nuevo los análisis con Classic Delta agrupan correctamente los textos (figura 7). Lo más interesante es que las novelas de Irisarri, Martínez de Lezea y Lasala constituyen ahora un grupo homogéneo en el que se ha introducido Corral y que la novela coescrita por Irisarri y Lasala se ha agrupado en esta ocasión con su segunda autora (Lasala) y se ha alejado, relativamente, de la otra autora (Irisarri).

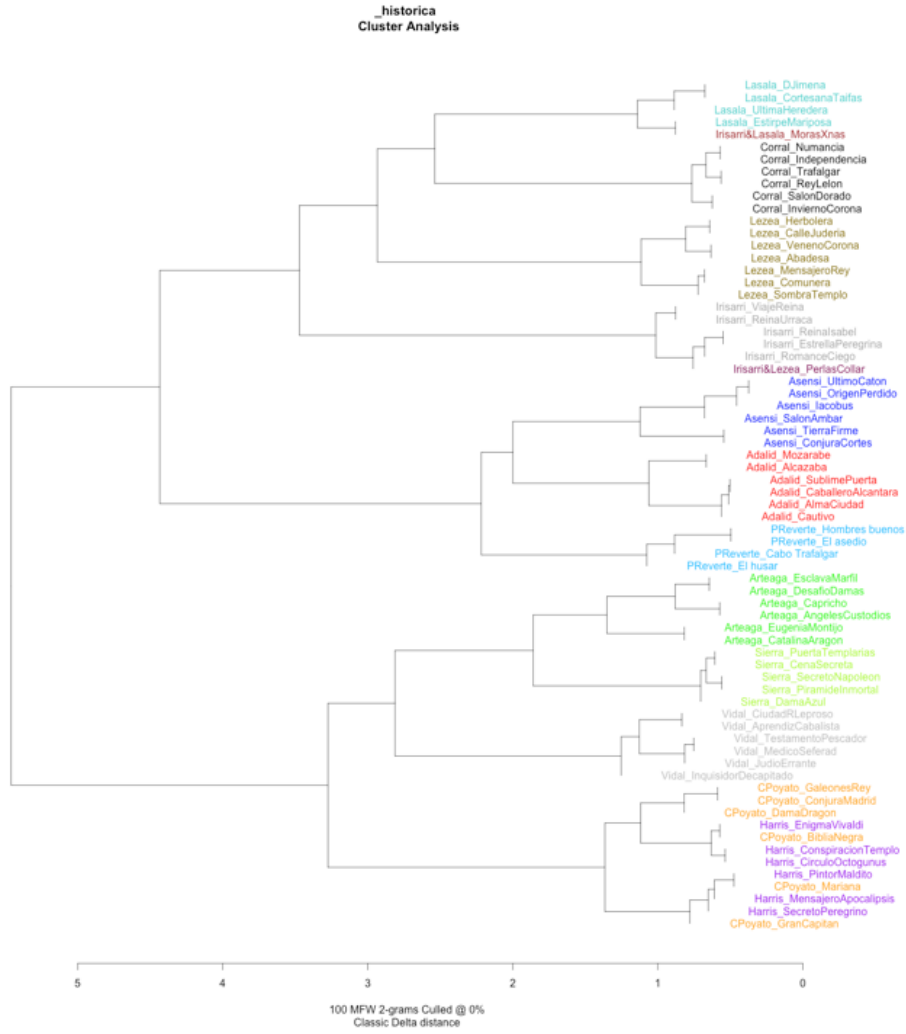

Figura 7. Dendrograma con los 100 bigramas más frecuentes y distancia Delta Classic con *stylo*

Ante el comportamiento un tanto oscilante de los textos coescritos por Ángeles de Irisarri y Martínez de Lezea –*Perlas para un collar*– por un lado y por Ángeles de Irisarri y Magdalena Lasala –*Moras y cristianas*– por otro; aunque en el caso de Irisarri y Martínez de Lezea suelen agruparse con ellas sin dificultad, es complicado que *Moras y cristianas* se aproxime a su otra coautora: Lasala, salvo en la última prueba con bigramas (figura 7), por lo

que se decidió *editar* los dos textos y someterlos a un nuevo análisis.

La edición de los textos ha consistido en dividirlos en otros dos que contuvieran lo que ha escrito cada una de las autoras. *Perlas para un collar* es una colección de treinta relatos sobre mujeres cristianas, judías y musulmanas y cada uno de ellos está firmado al final, mientras que en *Moras y cristianas* solo se indica que Irisarri se ha ocupado de las cristianas y Lasala de las moras. Así, pues, se han creado cuatro nuevos archivos Irisarri_Moras, Irisarri_Perlas, MLezea_Perlas y Lasala_Moras, se han retirado los *originales* del corpus y se han procesado con *stylo*. Se ha llevado a cabo un análisis de grupos con la medida Classic Delta y con 100 palabras más frecuentes (figura 8) y 1000 palabras más frecuentes (figura 9) y han ubicado cada nuevo texto entre las obras de cada una de las autoras y también ha ocurrido así cuando se ha hecho el análisis con los 100 bigramas más frecuentes (figura 10).

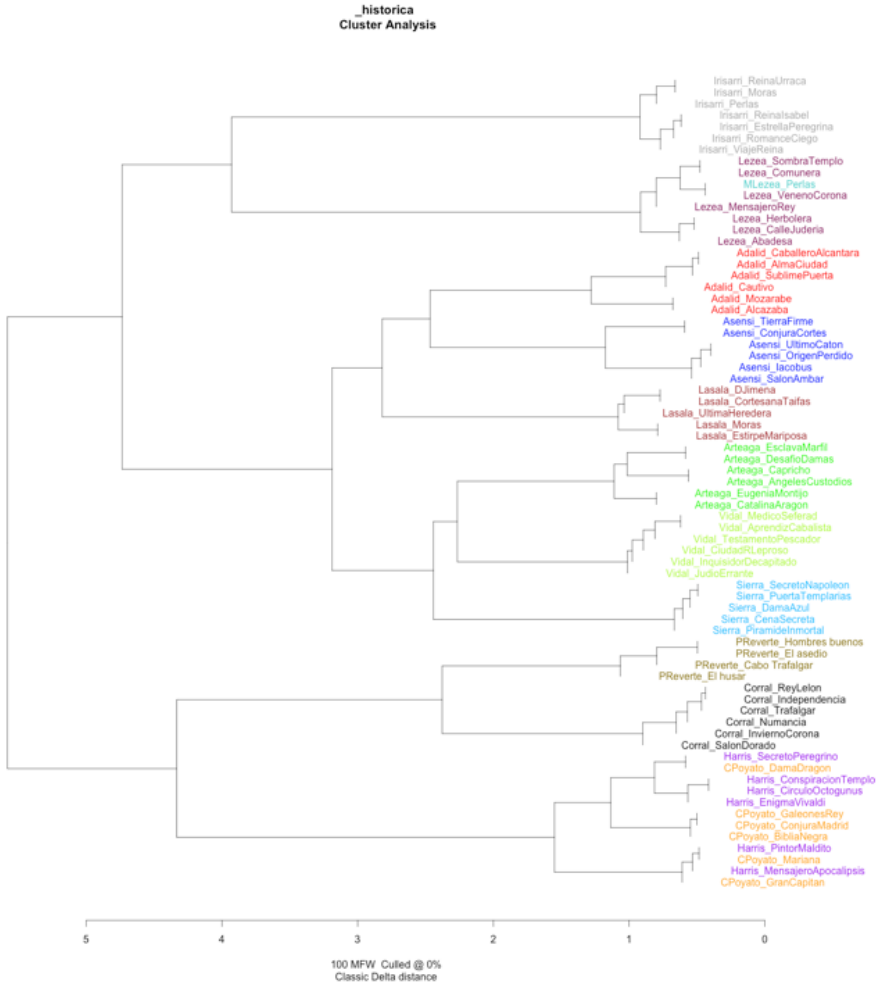

Figura 8. Dendrograma de los textos *editados*, medida Classic Delta y 100 MFW con *stylo*

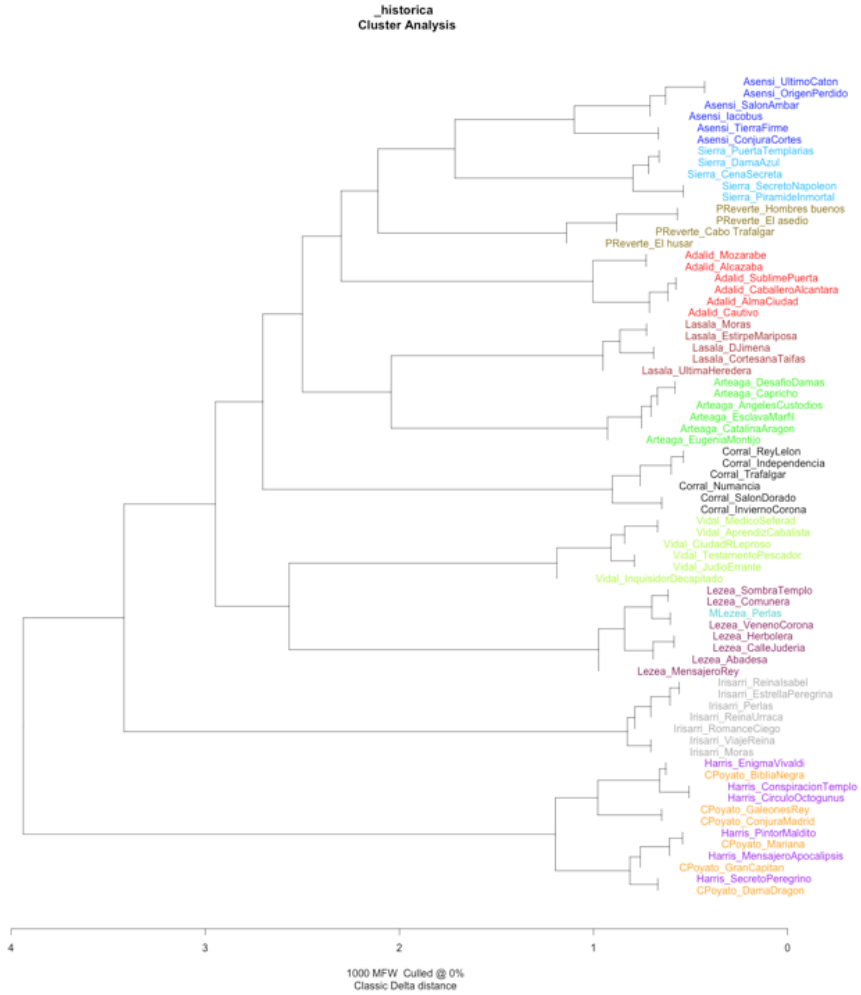

Figura 9. Dendrograma de los textos *editados*, medida Classic Delta y 1000 MFW con *stylo*

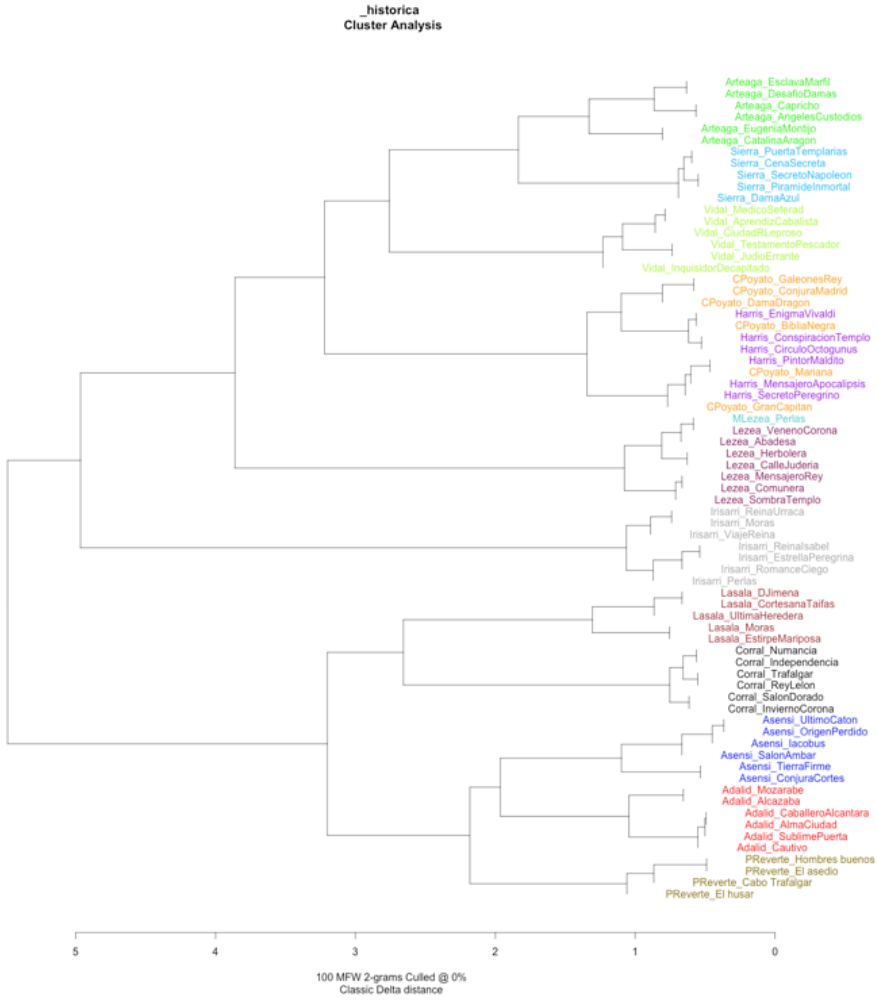

Figura 10. Dendrograma de los textos *editados*, constuido con los 100 bigramas más frecuentes y medida Classic Delta con *stylo*

5. Análisis del corpus de novelas policiacas

Con el corpus policial se han realizado las mismas pruebas que con las novelas históricas: análisis de grupos con las 100 y 1000 palabras más frecuentes y bigramas más frecuentes y los resultados

son análogos. Con 100 palabras y la medida Classic Delta se agrupan correctamente los autores (figura 11).

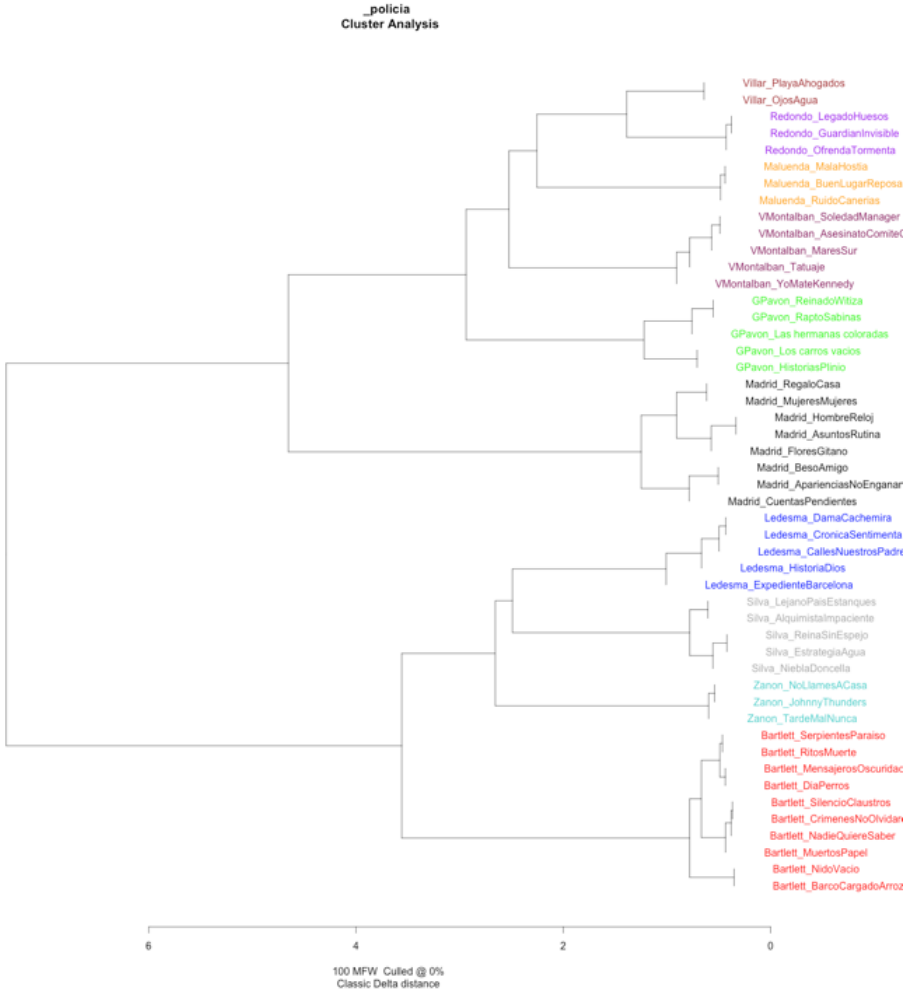

Figura 11. Dendrograma del corpus policial con 100 MFW y Classic Delta con *stylo*

Al aumentar a 1000 la agrupación sigue siendo correcta, lo único que varía, como de costumbre, es la ordenación interna (figura 12).

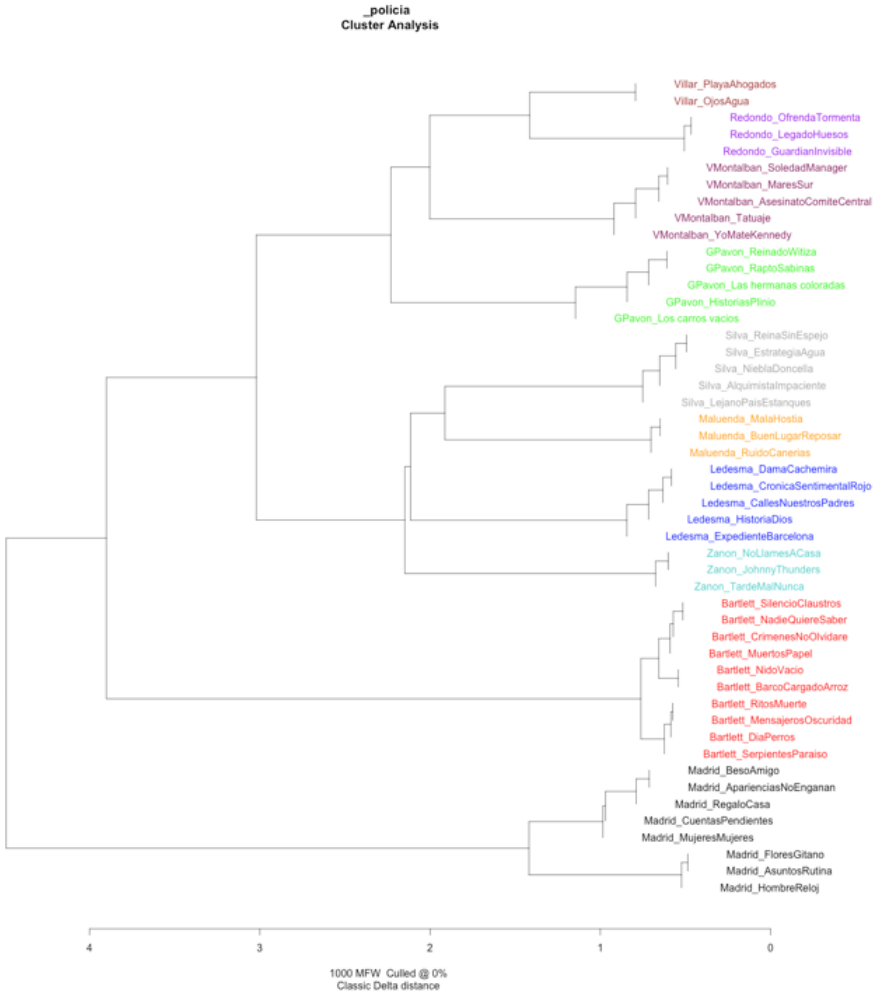

Figura 12. Dendrograma del corpus policial con 1000 MFW y distancia Classic Delta con *stylo*

Sin embargo, al reducir el número de rasgos a 25 no hay problemas. Las únicas diferencias se hallan en la ordenación final (figura 13).

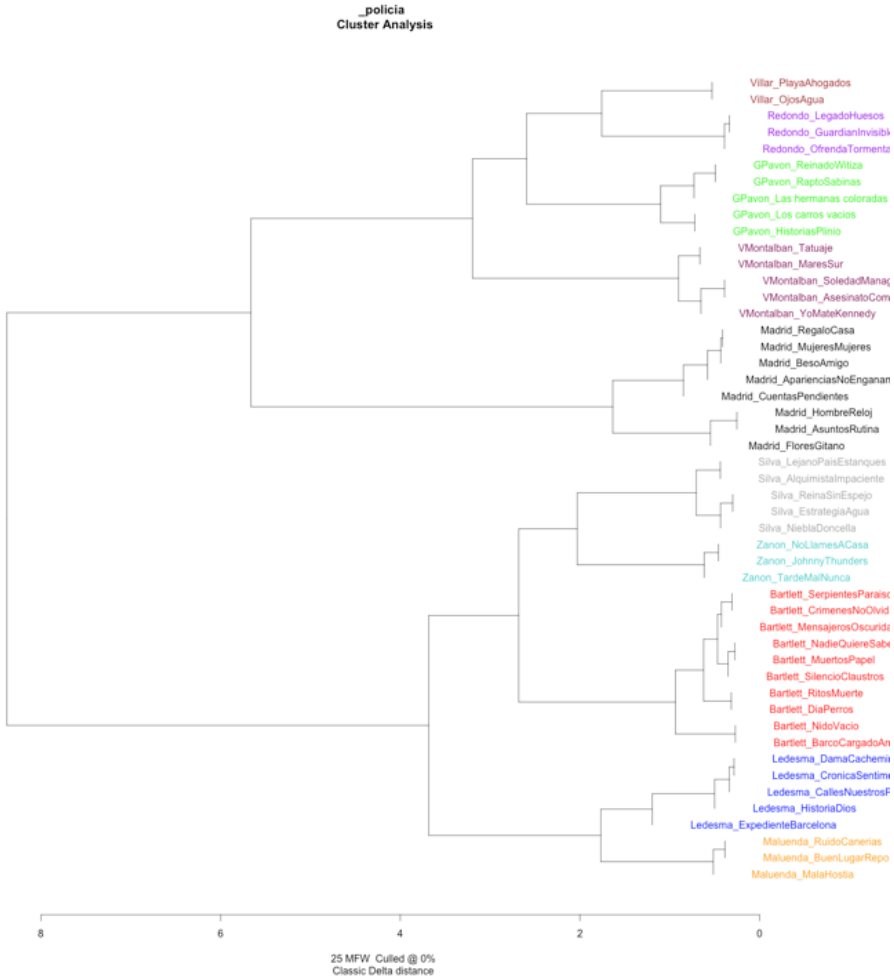

Figura 13. Dendrograma del corpus policial con 25 MFW y Classic Delta con *stylo*

Ahora vamos a introducir un elemento perturbador en el experimento. Francisco González Ledesma usó también el seudónimo de Silver Kane para publicar novelitas del oeste. Para esta parte hemos incorporado al corpus policiaco, aunque temáticamente estén muy alejadas, tres novelas de Silver Kane: *Póker de damas*, *Justiciero S. A.* y *Cadáver a subasta*. Los datos estadísticos básicos de estas tres novelitas se encuentran en la tabla 2.

	E. Moriel	**Silver Kane**	**Kane+Moriel**
textos	1	3	4
caracteres	589923	406566	996489
tokens	104356	71721	176077
tipos	11696	8389	15940
párrafos	3082	4235	7317
Token-Tipo Ratio	11.21	11.70	9.05
Media de tokens / novela	104356	23907	44019
Media tipos / novela	11696	4250	6112
Media párrafos / novela	3082	1411	1829
Media caracteres / novela	589923	135522	249122
Media palabras/párrafo	33.86	16.94	24.06
Media caracteres / párrafo	191.41	96.00	136.19
Media caracteres / palabra	5.65	5.67	5.66

Tabla 2. Estadísticas descriptivas resumidas de Silver Kane y E. Moriel

El análisis con tan solo 100 MFW (figura 14) sitúa perfectamente las novelas firmadas por Silver Kane entre las de González Ledesma.

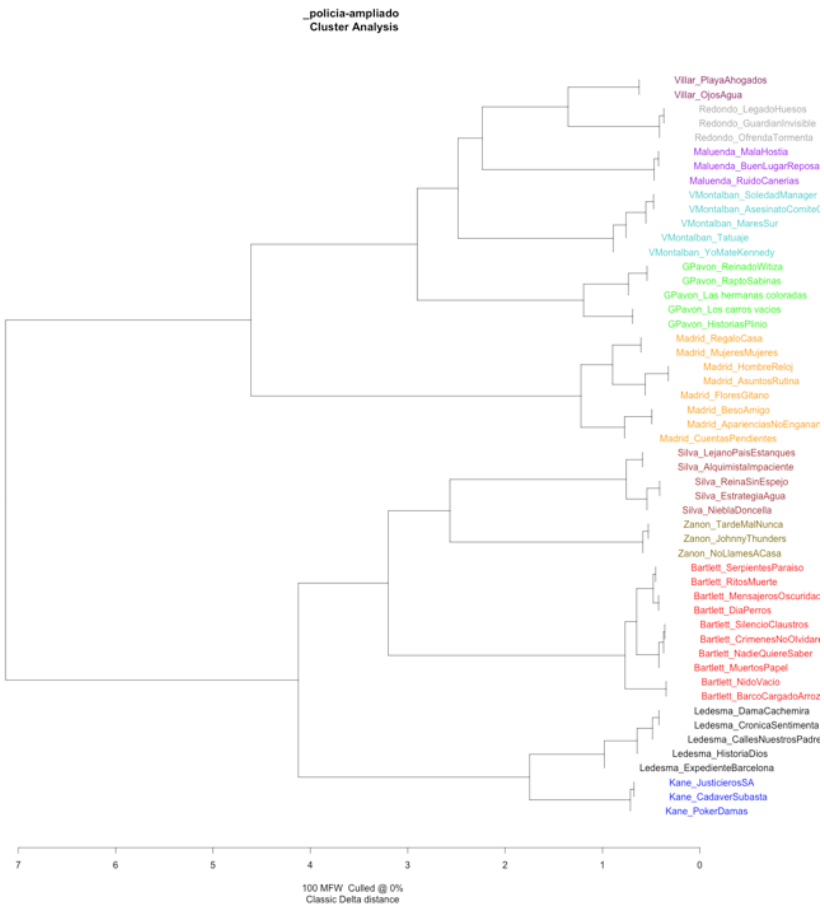

Figura 14. Dendrograma del corpus policial incorporado Silver Kane con 100 MFW y Classic Delta

Añadir la complicación de la otra firma que González Ledesma utilizó para novelas policiacas no mueve un ápice los resultados. Se sitúa claramente con Kane y González Ledesma, es más, ambos se sitúan a ambos lados de González Ledesma dentro del mismo nodo, tanto si solo tenemos en cuentas las 100 palabras más frecuentes (figura 15) como las 1000 (figura 16) e incluso si consideramos los bigramas (figura 17).

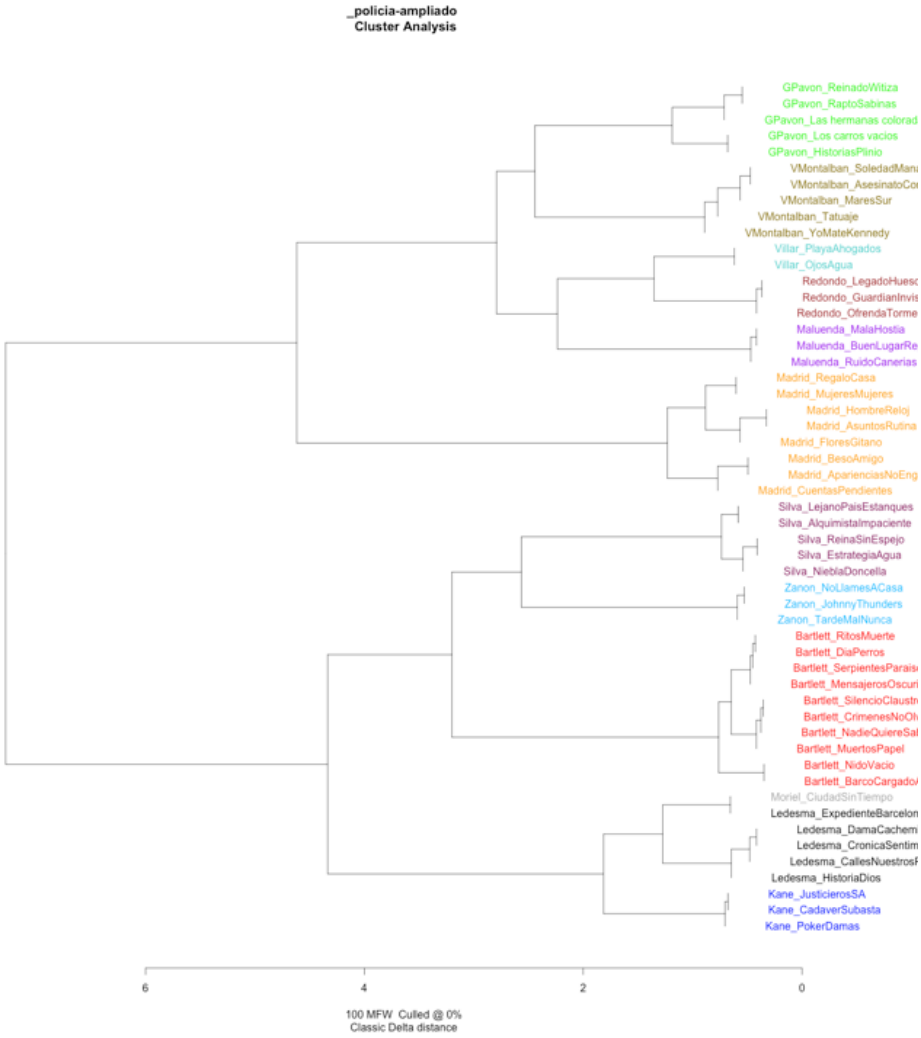

Figura 15. Dendrograma del corpus policial añadidos Kane y Moriel con 100 MFW Classic Delta

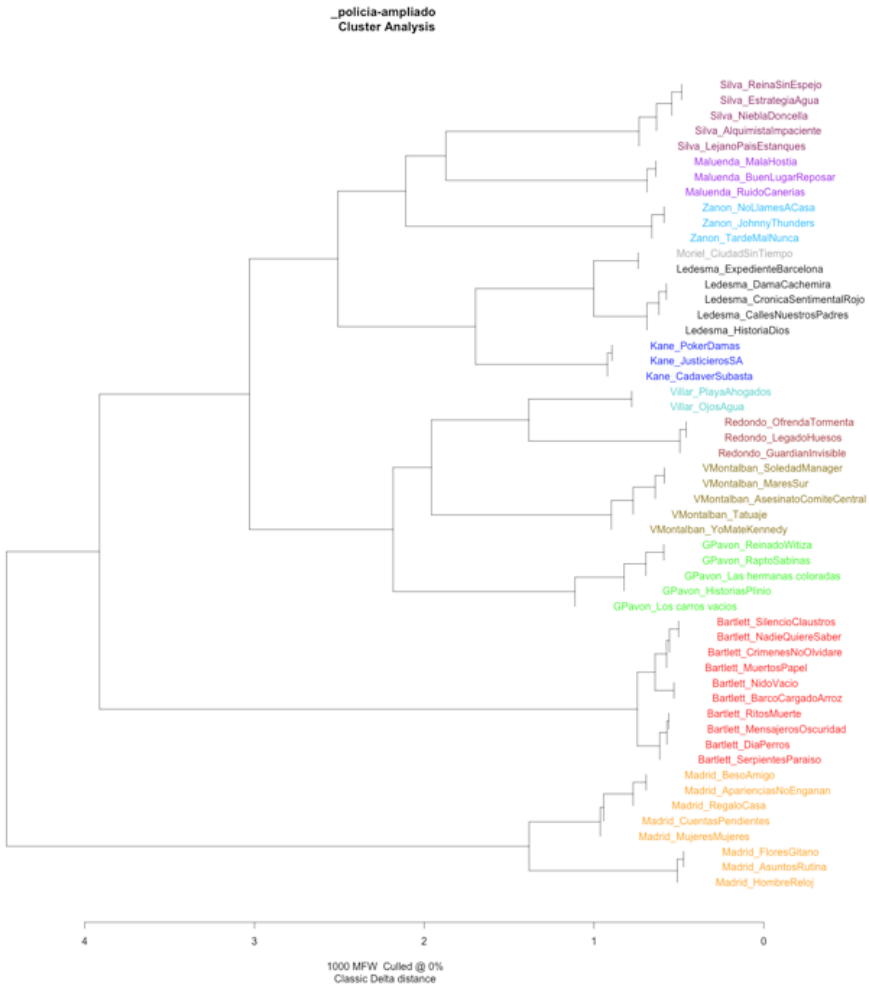

Figura 16. Dendrograma del corpus policial añadidos Kane y Moriel con las 1000 MFW y Classic Delta

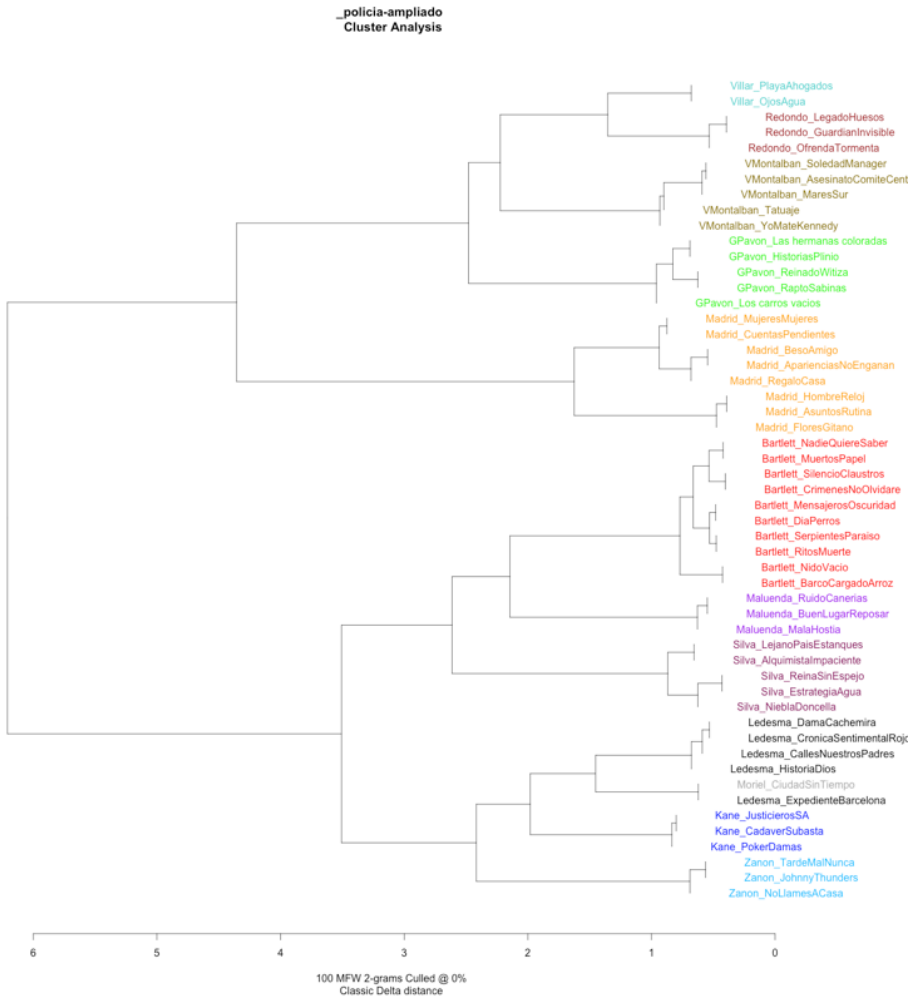

Figura 17. Dendrograma del corpus policial añadidos Kane y Moriel construido con los 100 bigramas más frecuentes y medida Classic Delta

6. Conclusión

La conclusión que se puede extraer a la luz de estos experimentos, elementales, de clasificación estilométrica, y la utilidad del paquete *stylo* es que estos métodos estadísticos para establecer la autoría de un texto funcionan sin error dentro de un

rango de rasgos de entre las 100 y las 1000 palabras más frecuentes (MFW) aplicando la medida Classic Delta y también con los 100 y 1000 bigramas (2-grams) más frecuentes. Luego, si ejecutamos el mismo tipo de análisis con los mismos rangos de rasgos, podríamos establecer quién se esconde tras el nombre de Alonso Fernández de Avellaneda y, a la luz del dendrograma obtenido (figura 18), no pueden ser ni el autor de la *Pícara Justina*, como propone Blasco (2005) ni el de la *Vida y trabajos de Jerónimo de Pasamonte*, que postula de Riquer (1988) y trata de confirmar informáticamente (¿?) Martín Jiménez (2007), puesto que ninguno de los dos se encuentra dentro de la rama en la que se ubica *el otro* Quijote[22]. Pero hay que hacer muchas más pruebas y el paquete *stylo* se ha mostrado como una interesante herramienta para los problemas de autoría que tiene las herramientas estadísticas pertinentes. Pero hay que tener en cuenta que quizá no sea ninguno de los autores mencionados, pues son muchos otros los que no se han considerado y, quizá, el autor sea otro del que no tenemos noticia alguna.

[22] Hay dos interesantes ensayos sobre el Avellaneda y su análisis estilométrico con *stylo* de mano de Rissler-Pipka (2016a), pero en uno de ellos –«Der falsche Quijote?» (Rissler-Pipka, 2016b)–, en el que ha aplicado la distancia del coseno (Consine Distance), los gráficos son ilegibles y no puede verse cómo se sitúa el Avellaneda.

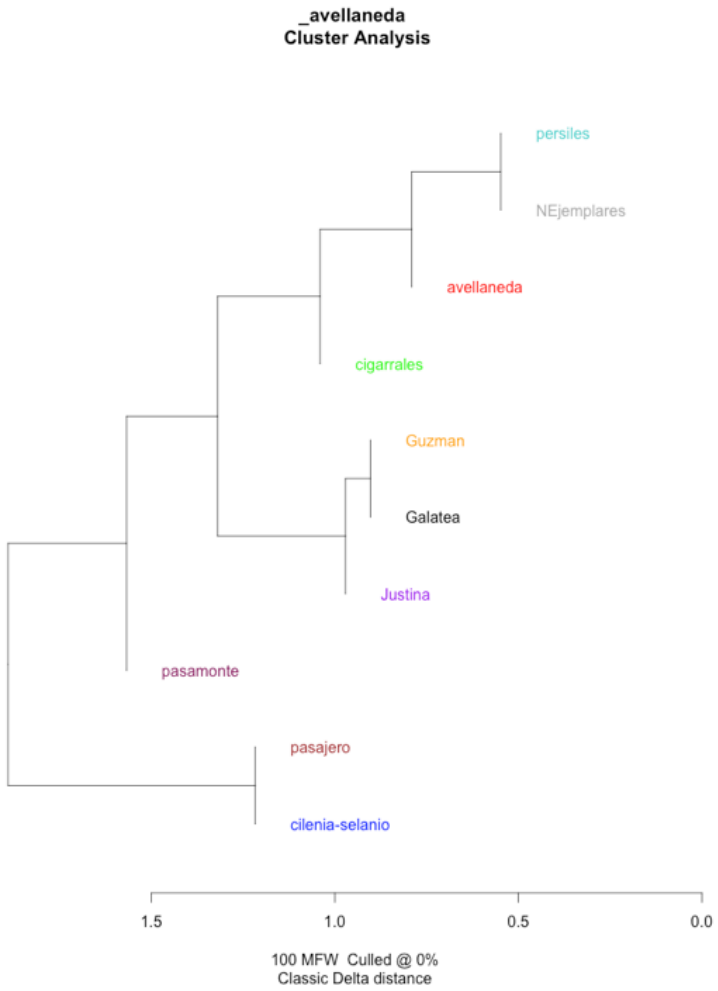

Figura 18. Dendrograma de algunos de los posibles responsables del «Avellaneda» realizado con *stylo* y las 100 MFW medidas con Classic Delta

7. Bibliografía

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

Blasco Pascual, Francisco Javier (2005). «La lengua de Avellaneda en el espejo de "La Pícara Justina"». *Boletín de la Real Academia Española* 85 (291): pp. 53-109.

Brinegar, C. S. (1963). «Mark Twain and the Quintus Curtius Snodgrass letters: A Statistical test of Authorship». *Journal of the American Statistical Association* 58: pp. 85-96.

Burrows, John (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. <http://revistacaracteres.net/revista/vol5n1mayo2016/entendiendo-delta/> (17/06/2016).

Craig, D. (2009). *Shakespeare, computers, and the mystery of authorship*. New York: Cambridge University Press.

Eder, Maciej, Jan Rybicki y Mike Kestemont (2013). *Stylo: a Package for Stylometric Analyses*. <https://sites.google.com/site/computationalstylistics/stylo/stylo_howto.pdf?attredirects=0&d=1> (17/06/2016).

Eder, Maciej, Jan Rybicki y Mike Kestemont (s. f.). «Stylometry with R: A Package for Computational Text Analysis». *The R Journal*: pp. 1-15. <https://journal.r-project.org/archive/accepted/eder-rybicki-kestemont.pdf> (17/06/2016).

Foster, Donald (1989). *Elegy by W.S.: a study in attribution*. Newark: University of Delaware Press.

Frías Delgado, Antonio (2009). «Distribución de frecuencias de la longitud de las palabras en español aspectos diacrónicos y de estilometría». Eds. Pascual Cantos Gómez y Aquilino Sánchez Pérez. *A survey of corpus-based research*.

<http://www.um.es/lacell/aelinco/contenido/pdf/51.pdf> (17/06/2016).

Gil-Albarellos Pérez-Pedrero, Susana (2010). «Algunas consideraciones teóricas sobre el fraude literario». Eds. Javier Blasco, Patricia Cepeda Marín y Cristina Ruiz Urbón. *Hos ego versiculos feci... Estudios de atribución y plagio*. Madrid: Iberoamericana - Vervuert. pp. 333-345.

Gil-Albarellos Pérez-Pedrero, Susana (2011). «"Que no hay tan diestra mentira/que no se venga a saber". Teorías de la falsificación literaria». Ed. Joaquín Álvarez Barrientos. *Imposturas literarias españolas*. Salamanca: Ediciones Universidad de Salamanca. pp. 17-32.

Holmes, David I. (1994). «Authorship attribution». *Computers and the Humanities* 28 (2): pp. 87-106.

Holmes, David I. (1999). «Stylometry». *Encyclopedia of Statistics*. Londres: Wiley.

Holmes, David I. (1998). «The Evolution of Stylometry in Humanities Scholarship». *Literary and Linguistic Computing* 13 (3): pp. 111-117.

Jockers, M.L. (2014). *Text Analysis with R for Students of Literature*. Cham: Springer.

Juola, Patrick (2006). «Authorship attribution». *Foundations and Trends in Information Retrieval* 1 (3): pp. 233-334.

Juola, Patrick (2013a): «How a Computer Program Helped Reveal J. K. Rowling as Author of A Cuckoo's Calling». *Scientific American* <http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/> (17/6/2016).

Juola, Patrick (2013b). «Rowling and "Galbraith": an authorial analysis». *Language Blog*

<http://languagelog.ldc.upenn.edu/nll/?p=5315> (17/6/2016).

Juola, Patrick (2015). «The Rowling case: A proposed standard analytic protocol for authorship questions». *Digital Scholarship in the Humanities* 30. <http://dsh.oxfordjournals.org/content/30/suppl_1/i100> (17/7/2016).

Juola, Patrick, John Sofko y Patrick Brennan (2006). «A prototype for authorship attribution studies». *Literary and Linguistic Computing* 21 (2): pp. 169-178.

Loper, Edward, Steven Bird y Ewan Klein (2009). *Natural language Processing with Python*. Sebastopol: O'Reilly.

López, Freddy (2011). «Donde se muestran algunos resultados de atribución de autor en torno a la obra cervantina». *Revista Colombiana de Estadística* 34 (1): pp. 15-37. <http://www.scielo.org.co/pdf/rce/v34n1/v34n1a02.pdf> (17/6/2016).

Madrigal, José Luis (2003). «De cómo y por qué *La tía fingida* es de Cervantes». *Artifara* 2. <http://www.cisi.unito.it/artifara/rivista2/testi/tiafingida.asp> (17.6.2016).

Madrigal, José Luis (2005). «El "Quijote" de Avellaneda, un crimen literario casi perfecto» *Voz y letra: Revista de literatura* 16 (1): pp. 247-294.

Martín Jiménez, Alfonso (2007). «Cotejo por medios informáticos de la "Vida" de Pasamonte y el "Quijote" de Avellaneda». *Etiópicas* 3: pp. 69-131. <http://www.uhu.es/revista.etiopicas/num/03/art_3_3.pdf> (17/6/2016).

Mendenhall, Thomas (1901). «A mechanical solution of a literary problem». *Popular Science Monthly* 60: pp. 97-105.

<https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_60/December_1901/A_Mechanical_Solution_of_a_Literary_Problem> (17/6/2016).

Merriam, Thomas (2013). «"Sir Thomas More"; sin estilometría». *Nueva revista de política, cultura y arte* 146: pp. 119-134.

Montoya Martínez, Jesús y Antonio Rubio Flores (1994). «De la comparación a la metáfora en Alfonso X. Cuestiones de estilometría en la prosa de la Partida Segunda». *Actas Primer Encuentro Interdisciplinar sobre Retórica, texto y Comunicación Cádiz 9, 10, 11 de diciembre de 1993*. Cádiz: Universidad Servicio de Publicaciones. pp. 156-162.

Mosteller, Frederick y David L. Wallace (1964). *Inference and disputed authorship: The Federalist*. Reading: Addison-Wesley.

Niederkorn, William S. (2002, 20 junio). «A Scholar Recants on His "Shakespeare" Discovery». *The New York Times*. <http://www.nytimes.com/2002/06/20/arts/a-scholar-recants-on-his-shakespeare-discovery.html> (17/6/2016).

Riquer, Martín de (1988). *Cervantes, Passamonte y Avellaneda*. Barcelona: Sirmio.

Rissler-Pipka, Nanette (2016a). «Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales». Ed. Hanno Ehrlicher. *El otro Don Quijote. La continuación de Fernández de Avellaneda y sus efectos*. Ausburgo: Institut für Spanien, Portugal- und Lateinamerikastudien (ISLA). pp. 27-51.

Rissler-Pipka, Nanette (2016b). «Der falsche Quijote? Autorschaftsattribution für spanische Prosa der frühen Neuzeit», *DHd 2016 Modellierung, Vernetzung, Visualisierung*, pp. 212-217. <http://dhd2016.de/boa.pdf> (17/6/2016).

Stamatatos, Efstathios (2009). «A survey of modern authorship attribution methods». *Journal of the American Society for Information Science and Technology*. <http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf> (17/6/2016).

Troya Déniz, Magnolia (2015). «*Quizá(s)* y *tal vez* en novelistas de España y América». *Philologica canariensia* 21. pp. 109-132. <http://ojsspdc.ulpgc.es/ojs/index.php/PhilCan/article/view/382> (17/6/2016).

8. Apéndice I

8.1. Lista de las novelas históricas analizadas

Autor	Título	Año de publicación
Jesús Sánchez Adalid	*Alcazaba*	2012
Jesús Sánchez Adalid	*El alma de la ciudad*	2007
Jesús Sánchez Adalid	*El caballero de Alcántara*	2008
Jesús Sánchez Adalid	*El Cautivo*	2005
Jesús Sánchez Adalid	*El mozárabe*	2001
Jesús Sánchez Adalid	*La sublime puerta*	2006
Almudena de Arteaga	*Ángeles custodios*	2010
Almudena de Arteaga	*Capricho*	2012
Almudena de Arteaga	*Catalina de Aragón, reina de Inglaterra*	2002
Almudena de Arteaga	*El desafío de las damas*	2006
Almudena de Arteaga	*La esclava de marfil*	2005
Almudena de Arteaga	*Eugenia de Montijo*	2000
Matilde Asensi	*La conjura de Cortes*	2012
Matilde Asensi	*Iacobus*	2000
Matilde Asensi	*El origen perdido*	2003

Matilde Asensi	*El salón de Ámbar*	1999
Matilde Asensi	*Tierra Firme*	2007
Matilde Asensi	*El último Catón*	2001
José Luis Corral	*¡Independencia!*	2005
José Luis Corral	*El invierno de la corona*	1999
José Luis Corral	*Numancia*	2003
José Luis Corral	*Rey Felón*	2009
José Luis Corral	*El salón dorado*	1996
José Luis Corral	*Trafalgar*	2001
José Calvo Poyato	*La Biblia negra*	2000
José Calvo Poyato	*Conjura en Madrid*	1999
José Calvo Poyato	*La dama del dragón*	2008
José Calvo Poyato	*Los galeones del rey*	2002
José Calvo Poyato	*El Gran Capitán*	2015
José Calvo Poyato	*Mariana, los hilos de la libertad*	2013
Peter Harris	*El círculo Octogunus*	2007
Peter Harris	*La conspiración del templo*	2006
Peter Harris	*El enigma de Vivaldi*	2005
Peter Harris	*El mensajero del Apocalipsis*	2012
Peter Harris	*El pintor maldito*	2013
Peter Harris	*El secreto de peregrino*	2010
Ángeles de Irisarri	*La estrella peregrina*	2010
Ángeles de Irisarri	*Isabel, la reina*	2001
Ángeles de Irisarri	*La reina Urraca*	2000
Ángeles de Irisarri	*Romance de ciego*	2005
Ángeles de Irisarri	*El viaje de la reina*	1991
Ángeles de Irisarri & Magdalena Lasala	*Moras y cristianas*	1998
Ángeles de Irisarri & Esperanza Martínez de Lezea	*Perlas para un collar*	2009
Magdalena Lasala	*La cortesana de Taifas*	2007
Magdalena Lasala	*Doña Jimena. La gran desconocida en la historia del Cid*	2006
Magdalena Lasala	*La estirpe de la mariposa*	1999
Magdalena Lasala	*La última heredera*	2015
Esperanza Martínez de Lezea	*La abadesa*	2002

Esperanza Martínez de Lezea	*La calle de la judería*	1998
Esperanza Martínez de Lezea	*La comunera. María Pacheco, una mujer rebelde*	2003
Esperanza Martínez de Lezea	*la herbolera*	2000
Esperanza Martínez de Lezea	*El mensajero del rey*	2002
Esperanza Martínez de Lezea	*A la sombra del templo*	2005
Esperanza Martínez de Lezea	*Veneno para la corona*	2011
Arturo Pérez-Reverte	*Cabo de Trafalgar*	2004
Arturo Pérez-Reverte	*El asedio*	2010
Arturo Pérez-Reverte	*El húsar*	1986
Arturo Pérez-Reverte	*Hombres buenos*	2015
Javier Sierra	*La cena secreta*	2004
Javier Sierra	*La dama azul*	2008
Javier Sierra	*La pirámide inmortal*	2014
Javier Sierra	*Las puertas templarias*	2000
Javier Sierra	*El secreto egipcio de Napoleón*	2002
César Vidal	*El aprendiz de cabalista*	2003
César Vidal	*La ciudad del rey leproso*	2009
César Vidal	*El inquisidor decapitado*	2014
César Vidal	*El judío errante*	2009
César Vidal	*El médico de Sefarad*	2004
César Vidal	*El testamento del pescador*	2004

9. Apéndice II

9.1. Lista de las novelas policiacas analizadas

Autor	Título	Año de publicación
Alicia Giménez Bartlett	*Un barco cargado de arroz*	2004
Alicia Giménez Bartlett	*Crímenes que no olvidare*	2015
Alicia Giménez Bartlett	*Día de perros*	1997
Alicia Giménez Bartlett	*Mensajeros de la oscuridad*	1999
Alicia Giménez Bartlett	*Muertos de papel*	2000
Alicia Giménez Bartlett	*Nadie quiere saber*	2013
Alicia Giménez Bartlett	*Nido vacío*	2007
Alicia Giménez Bartlett	*Ritos de muerte*	1996
Alicia Giménez Bartlett	*Serpientes en el paraíso*	2002
Alicia Giménez Bartlett	*El silencio de los claustros*	2009
Francisco García Pavón	*Historias de Plinio*	1968
Francisco García Pavón	*Las hermanas coloradas*	1970
Francisco García Pavón	*Los carros vacíos*	1965
Francisco García Pavón	*El rapto de las sabinas*	1969
Francisco García Pavón	*El reinado de Witiza*	1968
Francisco González Ledesma	*Las calles de nuestros padres*	1983
Francisco González Ledesma	*Crónica sentimental en rojo*	1984
Francisco González Ledesma	*La dama de Cachemira*	1986
Francisco González Ledesma	*Expediente Barcelona*	1983
Francisco González Ledesma	*Historia de Dios en una esquina*	1991
Juan Madrid	*Las apariencias no engañan*	1982
Juan Madrid	*Asuntos de rutina*	2010
Juan Madrid	*Un beso de amigo*	1980
Juan Madrid	*Cuentas pendientes*	1995
Juan Madrid	*Flores, el gitano*	2010

Juan Madrid	*El hombre del reloj*	2010
Juan Madrid	*Mujeres & Mujeres*	1995
Juan Madrid	*Regalo de la casa*	1986
Luis Gutiérrez Maluenda	*un buen lugar para reposar*	2012
Luis Gutiérrez Maluenda	*Mala hostia*	2011
Luis Gutiérrez Maluenda	*Ruído de cañerías*	2012
Dolores Redondo	*El guardian invisible*	2012
Dolores Redondo	*Legado en los huesos*	2013
Dolores Redondo	*Ofrenda a la tormenta*	2014
Lorenzo Silva	*El alquimista impaciente*	2000
Lorenzo Silva	*La estrategia del agua*	2010
Lorenzo Silva	*El lejano país de los estanques*	1998
Lorenzo Silva	*La niebla y la doncella*	2002
Lorenzo Silva	*La reina sin espejo*	2005
Domingo Villar	*Ojos de Agua*	2006
Domingo Villar	*la playa de los ahogados*	2009
Manuel Vázquez Montalbán	*Asesinato en el Comité Central*	1981
Manuel Vázquez Montalbán	*Los mares de Sur*	1979
Manuel Vázquez Montalbán	*La soledad del mánager*	1977
Manuel Vázquez Montalbán	*Tatuaje*	1974
Manuel Vázquez Montalbán	*Yo maté a Kennedy*	1971
Carlos Zanón	*Yo fui Johnny Thunders*	2014
Carlos Zanón	*No llames a casa*	2014
Carlos Zanón	*Tarde, mal y nunca*	2009

10. Apéndice III

10.1. Datos adjuntos

En el fichero de datos complementarios ("fradejas_datos.zip"), que se encuentra en <http://revistacaracteres.net/wp-content/uploads/2016/11/fradejas_datos.zip>, una vez descomprimido, en el primer nivel, se encuentran los ficheros "historica.txt" y "policiaca.txt" que contienen una tabla (separada con tabuladores) que recoge las informaciones básicas de cada uno de los textos considerados (autor, título, año de publicación, nombre abreviado –que es el que se utiliza en los dendrogramas–) y las estadísticas descriptivas básicas (número de párrafos, número de palabras-token, número de palabras-tipo, ratio token-tipo, número de caracteres, número de palabras por párrafo, número de caracteres por párrafo y número de caracteres por palabra). Los subdirectorios "historica-frecuencia-textos", "policiaca-frecuencia-textos" y "Kane+Moriel-frecuencias-textos" contienen un fichero por cada una de las novelas analizadas; en cada uno de ellos se halla el texto íntegro de la novela en forma de una tabla (separada con tabuladores) con todas las palabras tipo que constituyen el texto con sus frecuencias absolutas y relativas. Los subdirectorios "historica-tablas", "policial-tablas", "policial+ampliado-tablas" contienen dos tipos de ficheros: el de frecuencias y la lista de palabras (wordlist) que *stylo* ha manejado para realizar los análisis de grupos; pueden ser de las palabras consideradas individualmente o de los bigramas. En el subdirectorio "historica-tablas" algunos ficheros acaban en "4Manos", esos contienen las listas de palabras y las tablas de frecuencias de los textos del corpus histórico una vez *editadas* las novelas coescritas. Los ficheros de los subdirectorios acabados en

"–tablas" han sido generados por *stylo*. Todos los ficheros están en texto plano (UTF-8 y Unix LF) y pueden ser importados con sencillez en hojas de cálculo o reimportados en cualquier programa capaz de leer ficheros separados con tabuladores.

Este mismo texto en la web
http://revistacaracteres.net/revista/vol5n2noviembre2016/analisis-estilometrico