

MAPA ESTILOMÉTRICO DE LA NARRATIVA DE EDUARDO MENDOZA: APROXIMACIÓN A UN ANÁLISIS ESTILISTICO COMPUTACIONAL DE TEXTOS LITERARIOS

**STYLOMETRIC MAP OF EDUARDO MENDOZA'S NOVELS:
AN APPROACH TO A COMPUTATIONAL STYLISTICAL ANALYSIS
OF LITERARY TEXTS**

MIGUEL ÁNGEL BOTO BRAVO
miguel.angel.boto.bravo@gmail.com

Fecha de recepción: 27-03-2017
Fecha de aceptación: 02-06-2017

RESUMEN

Gracias a la evolución de la lingüística computacional y su aplicación a diferentes disciplinas de las Humanidades Digitales, la estilometría ha recibido un notable impulso con la aparición de herramientas informáticas desarrolladas específicamente para el análisis estilométrico de corpus textuales, como *Stylo R*, la herramienta desarrollada por el Institute of Polish language, de la Polish Academy of Sciences, y el Institute of English Studies, de la Jagiellonian University. A través del paquete *Stylo R* analizaremos la obra narrativa de Eduardo Mendoza para comprobar, más allá de su funcionalidad de autoría, su eficacia para la clasificación tipológica y estilística de textos literarios, así como para presentar la correcta configuración de los valores de los elementos involucrados en el procesamiento de datos.

PALABRAS CLAVE: estilometría; paquete Stylo; Eduardo Mendoza; lingüística informática; estilística.

ABSTRACT

Thanks to the evolution of computational linguistics and its application in various disciplines of Digital Humanities, stylometry has received considerable momentum with the advent of informatics tools specifically developed for stylometry analysis of textual corpora, such as Stylo, a tool developed by the Institute of Polish Language at the Polish Academy of Sciences and the Institute of English Studies at the Jagiellonian University. Using Stylo package, the narrative work developed by Eduardo Mendoza will be analysed to prove, beyond his authoring functionality, its efficacy on classifying the typology and stylistics of literary texts, as well as to present the right configuration of the involved elements' values in data processing.

KEY WORDS: stylometry; Stylo package; Eduardo Mendoza; computer linguistics; stylistics.

1. LA ESTILOMETRÍA

La estilometría pertenece al área de investigación de las Humanidades Digitales que, a su vez, derivan de la aplicación de la lingüística computacional al campo específico de las diversas áreas de la lingüística aplicada, de la lingüística textual y de la filología. Los grandes avances en el área de la computación producidos en los últimos treinta años han beneficiado a todas las áreas del saber, incluidos los estudios humanísticos que han visto como se multiplicaban exponencialmente los recursos disponibles para el análisis lingüístico, filológico o literario, provocando con ello la aparición inclusiva de nuevos campos de estudio soportados por los avances de la ingeniería y de la programación informática.

Este ha sido el caso de la estilometría que, si bien nació como disciplina a finales del siglo XIX a partir de los trabajos de Wincenty Lutosławski (1890) y Thomas Mendenhall (1901), no ha podido desarrollarse por completo hasta la aparición de herramientas informáticas que facilitasen las tareas de procesamiento de datos masivos. A partir de la programación de herramientas informáticas específicas para el análisis textual, la estilometría se ha unido a la lingüística informática, junto con la lingüística de corpus, la lexicografía y sociolingüística computacionales, aunque dentro del subgrupo específico del análisis infoasistido de textos.

La estilometría, también llamada análisis computacional de estilo textual, es una técnica de análisis perteneciente a la lingüística informática cuya finalidad consiste en establecer la relación entre el estilo de escritura de los textos y los metadatos contenidos en tales textos (datación, género textual, sexo, o autoría, principalmente) a partir del análisis estadístico de corpus textuales (Kenny, 1981; Eder, 2016). Desde el punto de vista de los estudios literarios, la estilometría se refiere principalmente a una serie de técnicas computacionales y estadísticas de análisis de textos que se han denominado como «lectura lejana», «no lectura» o «macroanálisis» (Jockers, 2013). En lugar de la práctica tra-

dicional para el análisis literario de la «lectura cercana», la estilometría no parte de una sola lectura directa, sino que permite explorar grandes corpus textuales utilizando para ello técnicas computacionales y gráficas. Además, la estilometría ha demostrado que las palabras gramaticales (preposiciones, artículos, conjunciones, etc.) tienen un mayor peso estilístico a la hora de establecer la huella lingüística de una obra o autor que las palabras semánticas, dado que las primeras, independientemente del género textual o de la temática, se usan inconscientemente (Stamatatos, 2009). Del mismo modo sucede con la frecuencia de uso de *n-grams*.

De este modo, la estilometría amplia los horizontes de la investigación en humanidades mediante la ampliación de los recursos de investigación a grandes corpus textuales, con el fin de encontrar relaciones y patrones de similitud y diferencias invisibles a través del análisis manual.

Por otro lado, la ausencia de un volumen considerable de estudios estilométricos aplicados a las literaturas hispánicas (incluso de desarrollo de herramientas propias de análisis estilométrico) es reflejo de la carencia que de modo general presentan los estudios de lingüística computacional en el ámbito hispánico, a excepción de los proyectos de los grupos «RedHD» o del proyecto «Humanidades Digitales Hispánicas». La reciente incorporación de asignaturas específicas de la informática y su aplicación al tratamiento de textos en los nuevos currícula de los grados universitarios de lingüística, filología y estudios literarios y el esfuerzo investigador de los proyectos mencionados favorecerán en el futuro el crecimiento en las áreas de desarrollo de nuevos sistemas y su aplicación al análisis textual y literario.

2. OBJETIVO, HERRAMIENTA Y METODOLOGÍA

Dado que la mayoría de estudios estilométricos publicados están centrados en la atribución de autoría o al análisis contrastivo de textos, el presente trabajo pretende mostrar el rendimiento de diferentes análisis estadísticos de MFW (siglas en inglés de «lista de las palabras más frecuentes») y de *n-grams*, o subsecuencia de «*n*» elementos de una secuencia dada, aplicados al conjunto de la obra de un solo autor. Asimismo, y debido a las posibilidades numéricas tanto de los modelos estadísticos seleccionados como de valores de MFW y *n-grams* aplicados en cada caso, mostraremos los diferentes resultados y sugeriremos los modelos estadísticos con mayor eficiencia para el análisis de textos literarios narrativos, así como los valores mínimos y máximos de configuración de MFW y *n-grams* que aportan resultados congruentes con análisis estilísticos tradicionales. De este modo, y desde el punto de vista estrictamente literario, podremos mostrar cómo el análisis estilométrico ayuda al especialista a definir cuestiones no sólo de autoría, sino también de datación, estilo o género literario y los vínculos estilísticos entre diferentes obras, variaciones formales y ascendencia temporal.

Para ello, el corpus seleccionado pertenece a la obra narrativa de Eduardo Mendoza, cuya selección justificamos por varios motivos: el considerable volumen de obras publicadas (quince novelas, un libro de relatos y un ensayo), la variedad gené-

rica entre ellas (obras paródicas, novelas pastiche, relatos breves, ensayos) y el notable lapso de tiempo transcurrido entre la publicación de la primera y la última de las obras (cuarenta y un años).

El programa informático seleccionado para el estudio que nos ocupa será *Stylo*¹. Se trata del software de análisis estilométrico de corpus textuales de mayor desarrollo y con mayores funcionalidades en la actualidad y, por este motivo, el elegido para el presente trabajo. La herramienta ofrece un conjunto de funcionalidades específicas para el análisis exploratorio de corpus textuales, así como de tareas de clasificación textual o procesos de atribución de autoría propias de la lingüística forense. Frente a otras herramientas, la ventaja principal de *Stylo* radica en la integración de procedimientos típicos de la estilometría (como por ejemplo, preprocesamientos) junto con procedimientos estrictamente estadísticos. Programado en lenguaje R, el código fuente y los binarios del paquete están disponibles gratuitamente en la Red de Archivos Comprehensive R².

Stylo proporciona a través de la interfaz de la consola de R cinco funcionalidades: classify() (para la clasificación de textos por afinidad estadística y su posterior aplicación a la autoría de textos), oppose() (para realizar un análisis contrastivo entre dos conjuntos de textos dados a fin de generar una lista de palabras clave usadas por el autor testado, o por un conjunto de autores, y otra lista de palabras voluntariamente evitadas por el primero, en comparación con otro conjunto de textos), rolling.delta() (para analizar trabajos colaborativos entre dos o más autores a fin de determinar la autoría de cada uno de los fragmentos) y, por último, stylo() (para evaluar y visualizar las similitudes estilísticas entre los textos procesados a partir de estadísticas multivariantes).

Dado que entre los objetivos del presente trabajo se encuentra la presentación de un mapa estilométrico del conjunto de la narrativa de Eduardo Mendoza para, posteriormente, extraer conclusiones relativas a vinculación estilística, genérica y temporal entre las obras analizadas, la funcionalidad que usaremos para tal fin será stylo().

Stylo () genera en primer lugar una lista de las palabras, también llamadas «tokens», más frecuentes (MFW, siglas en inglés de «most frequently words») para la totalidad del corpus cargado. Posteriormente, procesa las frecuencias de las MFWs de cada texto de forma individual para crear una matriz inicial de palabras ordenadas en filas por cada uno de los textos individuales del corpus, ordenados a su vez en columnas. Cada celda contiene la frecuencia de una sola palabra en un solo texto. A continuación, normaliza las frecuencias resultantes: selecciona las palabras de los rangos de frecuencia deseados para los análisis y realiza algunos procedimientos de procesamiento de la información, como la supresión automática de los pronombres personales o la selección de parámetros seleccionados por el usuario, para producir una lista de palabras final que será usada en los análisis definitivos. Después, compara los resultados de textos individuales, realizando, por ejemplo,

¹ Creado por el Institute of Polish language de la Polish Academy of Sciences, junto al Institute of English Studies de la Jagiellonian University. <https://sites.google.com/site/computationalstylistics/stylo>

² <https://cran.r-project.org/>

cálculos estadísticos de distancia y utilizando diversos procedimientos (análisis de conglomerados, escalamiento multidimensional o análisis de componentes principales). Por último, la función producirá representaciones gráficas de redes nódales a partir de las distancias entre textos y escribirá los candidatos de autoría (o similitud) resultantes en un archivo de registro (results.txt). Cuando se selecciona la opción de árbol de consenso para visualización del gráfico, el script produce además un análisis de clúster virtual de una variedad de parámetros que generan un diagrama final mostrando la proximidad de los textos entre los análisis de clúster subyacentes.

Para poder realizar el procesamiento del corpus, los textos han de estar formateados como archivos de texto plano (extensión .txt). La forma más sencilla de formatear los textos es convertir desde los formatos electrónicos publicados por las editoriales Seix Barral, Destino y Planeta de las novelas de Mendoza (extensión .epub o .mobi como extensiones más frecuentes de libros electrónicos). Por otra parte, previo al preprocesamiento de los datos, es necesario que el nombre de cada archivo esté compuesto por al menos dos términos separados por guión bajo. La primera referencia alude al nombre del autor, mientras que la segunda al título de la obra. Ordenamos los nombres de los archivos cronológicamente por fecha de publicación en nota al pie de página³.

3. EL ANÁLISIS ESTILOMÉTRICO DE LA NARRATIVA DE EDUARDO MENDOZA CON STYLO

Realizamos el procesamiento de datos y obtenemos los siguientes valores de tipo de palabras y tokens contenidos por cada uno de los textos del corpus (figura 1).

Los tipos de palabra son las formas léxicas únicas aparecidas en cada obra, mientras que los «tokens» son el número total de esas formas léxicas teniendo en cuenta el total de veces utilizadas. El ratio, por tanto, muestra el porcentaje de formas léxicas sobre el total de tokens. A menor porcentaje de ratio, mayor es la variedad léxica contenida en la obra, aunque no debemos obviar el hecho de que cuanto mayor sea la extensión de la obra, mayor la posibilidad de reducción del ratio.

³ savolta, para *La verdad sobre el caso Savolta* (1975); cripta, para *El misterio de la cripta embrujada* (1979); laberinto, para *El laberinto de las aceitunas* (1988); ciudad, para *La ciudad de los Prodigios* (1986); isla, para *La isla inaudita* (1989); gurb, para *Sin noticias de Gurb* (1991); diluvio, para *El año del diluvio* (1992); comedia, para *Una comedia ligera* (1996); new york, para *Nueva York* (1997); tocador, para *La aventura del tocador de señoras* (2001); horacio, para *El último trayecto de Horacio Dos* (2002); mauricio, para *Mauricio o las elecciones primarias* (2006); pomponio, para *El asombroso viaje de Pomponio Flato* (2008); duslav, para *El final de Duslav* (2009); malentendido, para *El Malentendido* (2009); ballena, para *La ballena* (2009); bolsa, para *El enredo de la bolsa y la vida* (2012); gatos, para *Riña de gatos* (2015).

Obras	Tipos de palabra	Tokens	Ratio
Ciudad	20553	165646	12%
Comedia	18757	153989	12%
Savolta	16108	115510	14%
Gatos	14971	116102	13%
Tocador	13519	92428	15%
Mauricio	13069	95689	14%
Isla	13024	86716	15%
Laberinto	12454	68222	18%
Bolsa	11452	66663	17%
Cripta	9492	45690	21%
Pomponio	8544	45550	19%
New York	8225	38516	21%
Horacio	7834	43634	18%
Diluvio	6999	35416	20%
Gurb	6437	27910	23%
Ballena	6024	27698	22%
Duslav	3280	11402	29%
Malentendido	2701	9919	27%
TOTAL	193443	1246700	16%

FIGURA I
Tabla de tipos de palabras, «tokens» y ratio por obra

Para la generación del primer gráfico, configuramos las opciones de MFW con valores mínimos y máximos de 100 MFW, las configuraciones de sangrado⁴ con 0 como valor máximo y mínimo. A continuación seleccionamos el modelo estadístico⁵ que deseamos aplicar al análisis y la distancia entre nodos⁶, en este caso Análisis de Conglomerados y Classic Delta respectivamente:

⁴ Los valores de sacrificio especifican el grado de eliminación de las palabras que no aparecen en todos los textos del corpus.

⁵ Trabajaremos con tres de los seis modelos disponibles: Cluster Analysis o Análisis de Conglomerados, Consensus Tree o Árbol de Consenso y PCA (cov.) o Análisis de Componentes Principales con matriz de covarianza.

⁶ La funcionalidad stylo () ofrece ocho distancias estadísticas, de las cuales utilizaremos para nuestro análisis solamente Classic Delta y Euclidean por su mayor rendimiento aplicado al análisis literario.

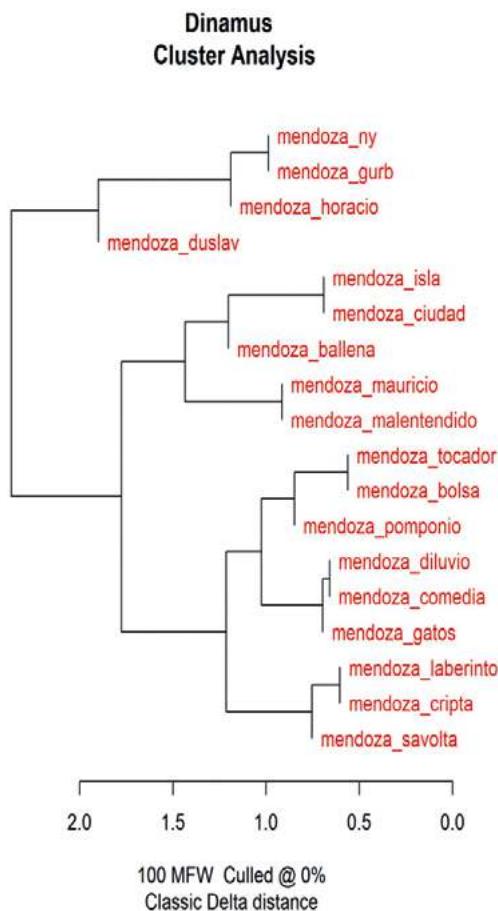


FIGURA 2

Dendograma de análisis de conglomerados con distancia Classic Delta con 100 MFW

Del dendograma resultante del primer análisis deducimos que *Stylo* ha ordenado las tres primeras publicaciones de Mendoza bajo el mismo nodo, aunque dos de ellas pertenecen a las parodias de las novelas de detectives protagonizadas por Ceferino. También, ha ordenado bajo el mismo nodo *Sin noticias de Gurb* y *El último trayecto de Horacio Dos*, dos parodias escritas bajo el formato de cuaderno de bitácora. Sorprendentemente, uno de los tres relatos que conforman *Tres vidas de Santos (El final de Duslav)* no aparece bajo el mismo nodo que los otros dos.

Debido a que la selección de MFW (100) puede haber resultado escasa para conseguir resultados más congruentes, ampliamos el número a 1.000:

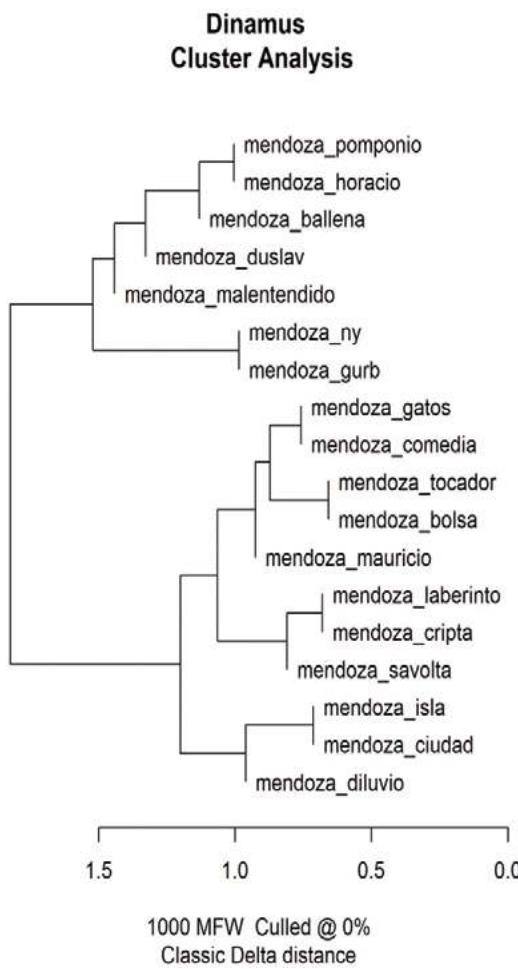


FIGURA 3

Dendograma de análisis de conglomerados con distancia Classic Delta con 1.000 MFW

Podemos observar que, por ejemplo, se ha obtenido un nuevo posicionamiento de *El año del Diluvio* respecto de *La ciudad de los Prodigios*, o que los tres relatos de *Tres vidas de santos* ahora sí aparecen vinculados bajo el mismo nodo. Asimismo, la distancia entre las cuatro novelas de Ceferino se ha reducido, aunque aún permanecen separadas en subgrupos diferentes.

Dado que al trabajar con corpus de novelas la extensión de estas permite ampliar significativamente los valores de MFW, realizaremos un nuevo análisis incrementando hasta 5.000 el valor de MFW:

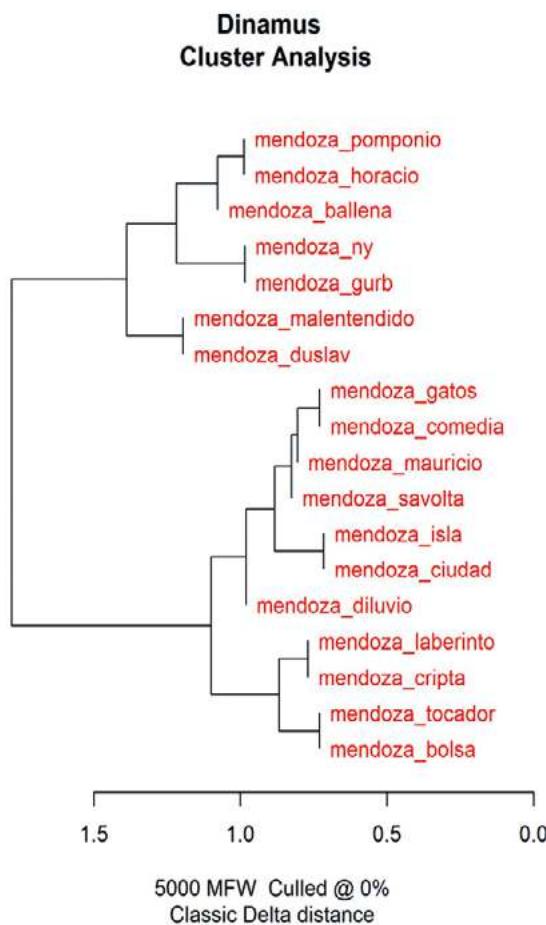


FIGURA 4

Dendograma de análisis de conglomerados con distancia Classic Delta con 5.000 MFW

Los resultados obtenidos muestran que se han agrupado definitivamente las novelas de Ceferino bajo un mismo nodo, que separa a su vez las dos primeras publicaciones de las dos últimas de la saga (*La aventura del tocador de señoras* y *El enredo de la bolsa y la vida*). Del mismo modo, ha agrupado bajo un mismo nodo todas las obras consideradas por la crítica como «serias», esto es, las novelas pastiche, aplicando una subagrupación específica para *Riña de Gatos* y *Una comedia ligera*, por un lado, y aproximando las dos grandes obras de referencia de Mendoza, *La verdad sobre el caso Savolta* y *La ciudad de los Prodigios*. Por último, podemos observar que ha creado un nodo específico tanto para los cuentos como para las tres parodias más extremas de Mendoza, esto es, *Sin noticias de Gurb*, *El último trayecto de Horacio Dos* y *El asombroso viaje de Pomponio Flato*.

Los tres análisis realizados confirman que para obras de extensión considerable, como novelas, conviene ampliar el rango de MFW más allá de las 1.000 unidades⁷. Asimismo, sustituyendo la distancia Classic Delta por Euclidean, los resultados obtenidos muestran algunas similitudes con el análisis por Classic Delta, aunque la posición marginal que ocupa *El final de Duslav*, la división en nodos diferentes de las cuatro parodias de Ceferino y la posición bajo el mismo nodo de *La verdad sobre el caso Savolta* y *El asombroso viaje de Pomponio Flato* desaconsejan el uso de la distancia Euclidean, como podemos comprobar en el dendograma resultante:

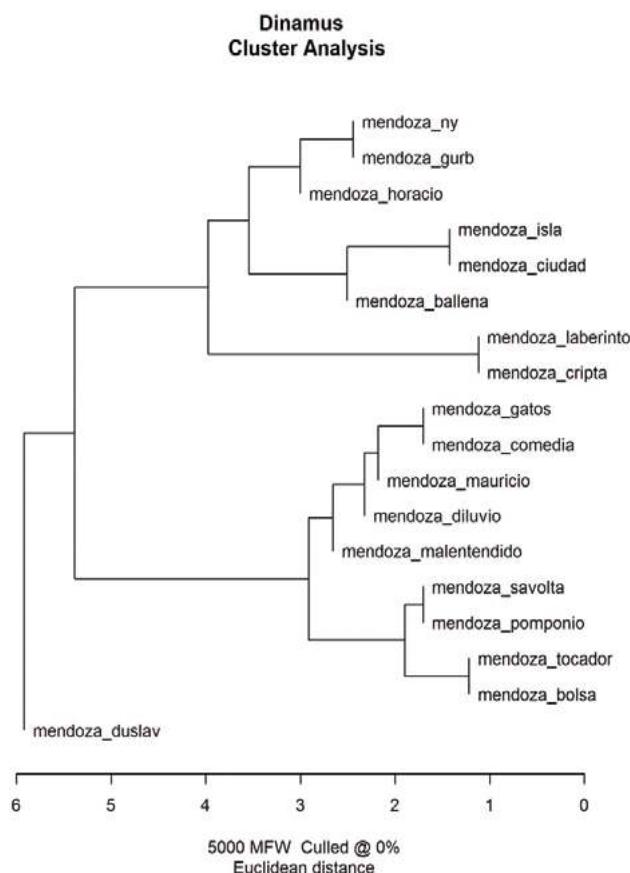


FIGURA 5

Dendrograma de análisis de conglomerados con distancia Euclidean con 5.000 MFW

⁷ Fradejas afirma respecto al número necesario de MFWs: «(...) 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» (Fradejas, 2016). En nuestro caso hemos podido comprobar que análisis por debajo de las 1.000 MFW no ordenan correctamente las obras.

A continuación, realizaremos una variación en el modelo Análisis de Conglomerados-Classic Delta, sustituyendo el procesamiento por palabras por el procesamiento por n-grams. Teniendo en cuenta que en español se producen con frecuencia agrupaciones de tres, cuatro o cinco «tokens» en el mismo orden debido a nuestra configuración de unidades fraseológicas, colocaciones y perifrasis verbales, incluiremos el número 3 como valor de gram:

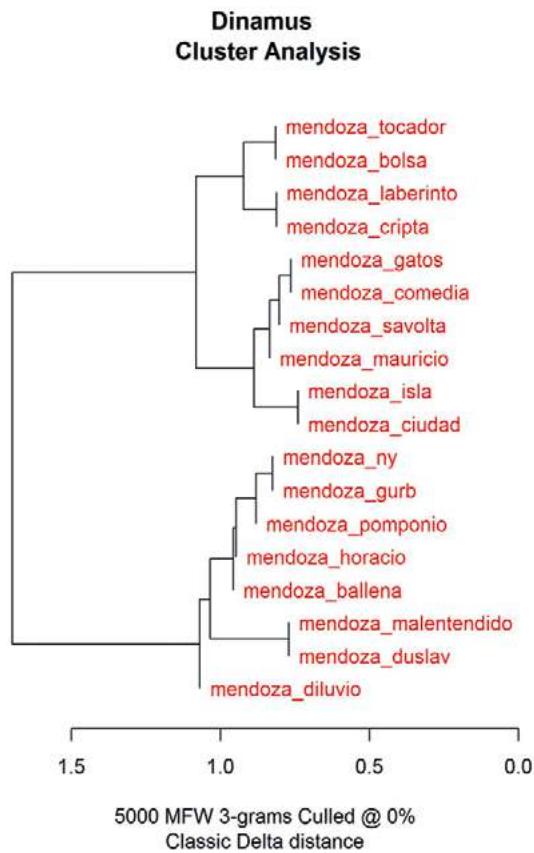


FIGURA 6
Dendrograma de análisis de conglomerados con distancia Classic Delta
con 5.000 MFW y 3 grams

Observamos en este caso que no se han producido variaciones en la distribución de las obras entre el análisis de 5.000 MFW y el de 3 grams. No obstante, ampliar el valor de grams a 5 desvirtúa los resultados, como podemos comprobar en el dendrograma resultante en el que el grupo de las cuatro novelas de Ceferino se separan en dos nodos diferentes:

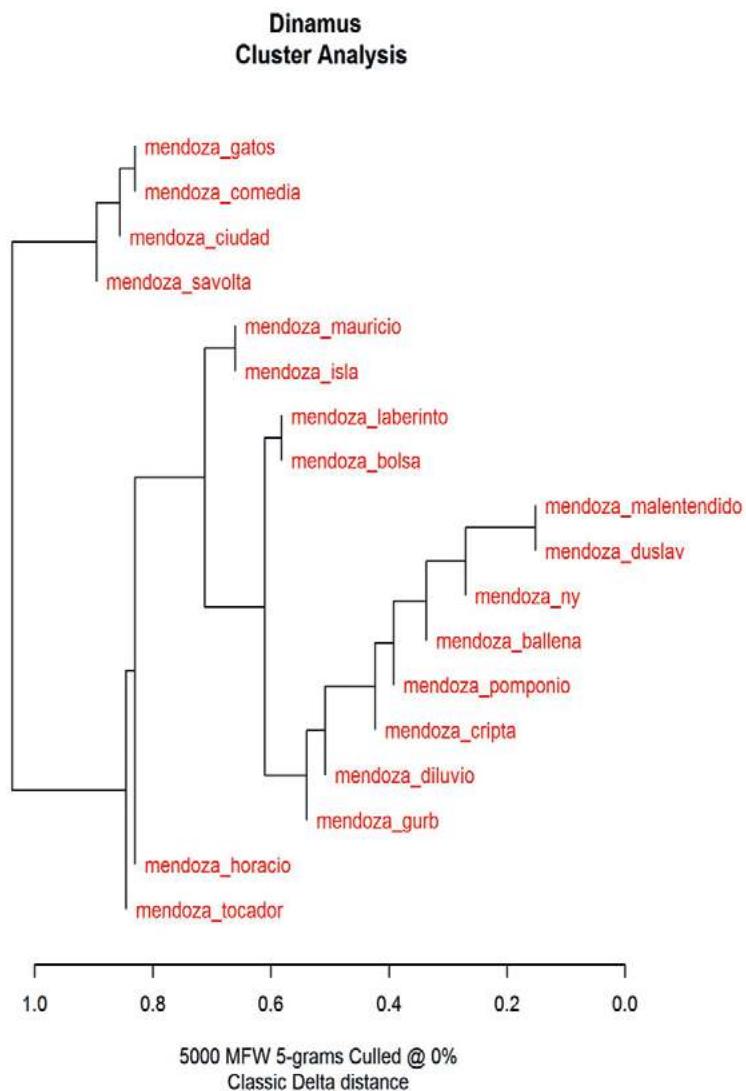


FIGURA 7

Dendograma de análisis de conglomerados con distancia Classic Delta con 5.000 MFW y 5 grams

Continuando con las pruebas combinatorias entre modelos estadísticos y distancias, aplicaremos el modelo PCA de Análisis de Componentes principales junto con la distancia Classic Delta. La única variación respecto a los dendogramas reside en el tipo de visualización obtenida, no observándose ningún tipo de variación entre los grupos:

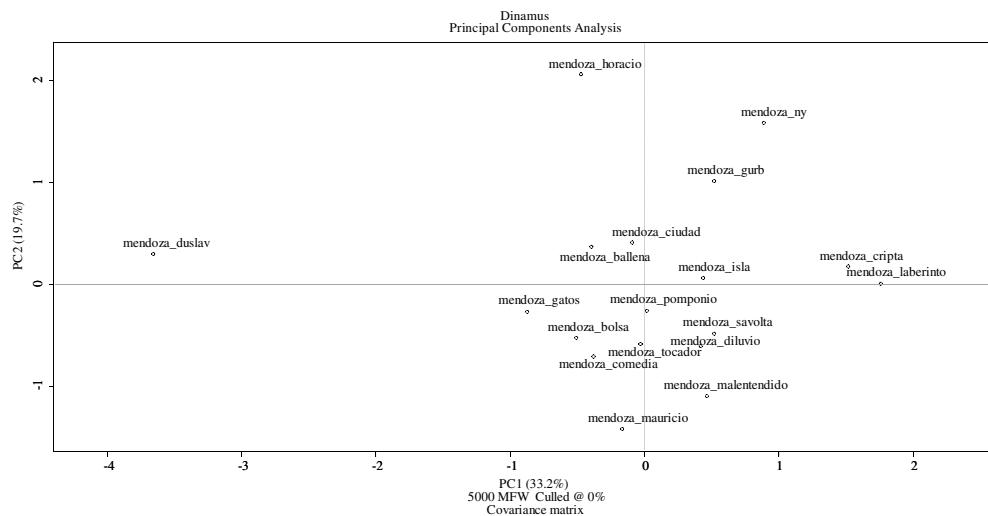


FIGURA 8

Dendograma de análisis de componentes principales con distancia Classic Delta con 5.000 MFW

Por último, mostraremos los resultados de la distancia Classic Delta bajo los parámetros que hemos demostrado más eficientes (MFW respecto de n-grams y valores superiores a 1.000 MFW) con el modelo estadístico de árbol de consenso:

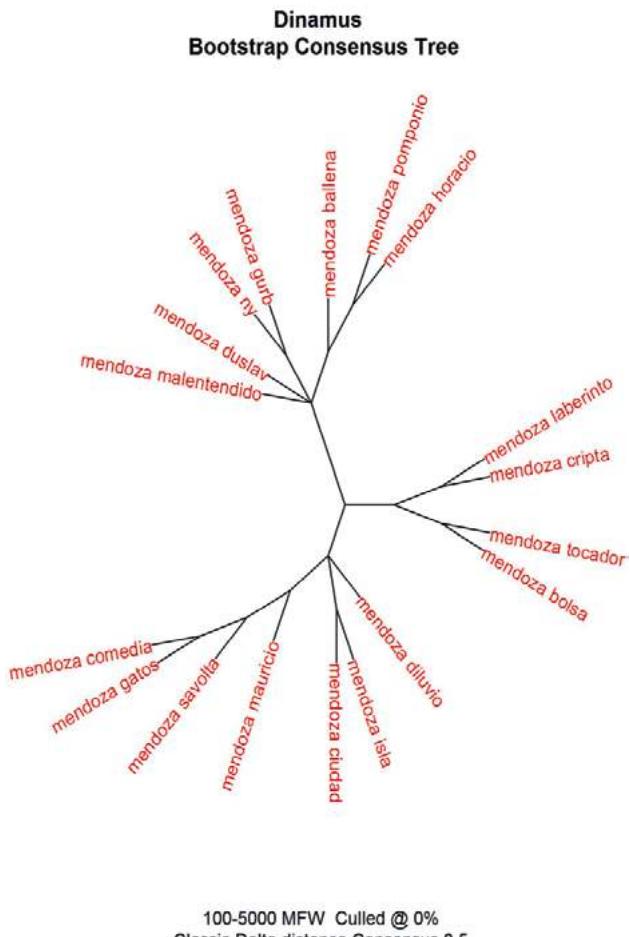


FIGURA 9
Árbol de consenso con distancia Classic Delta con 5.000 MFW

Se han mantenido las agrupaciones nodales aportadas por los dendogramas 4 y 6, pero obtenemos una mejora cualitativa respecto de su visualización. Cabe señalar en este sentido que la herramienta ofrece las tablas de distancia en formato .cvs para importación y procesamiento de los datos en programas específicos de configuración y visualización de gráficos y redes nodales, como Gephi (Bastian, Heymann y Jacomy, 2009), con cuyo mapa finalizamos el análisis.

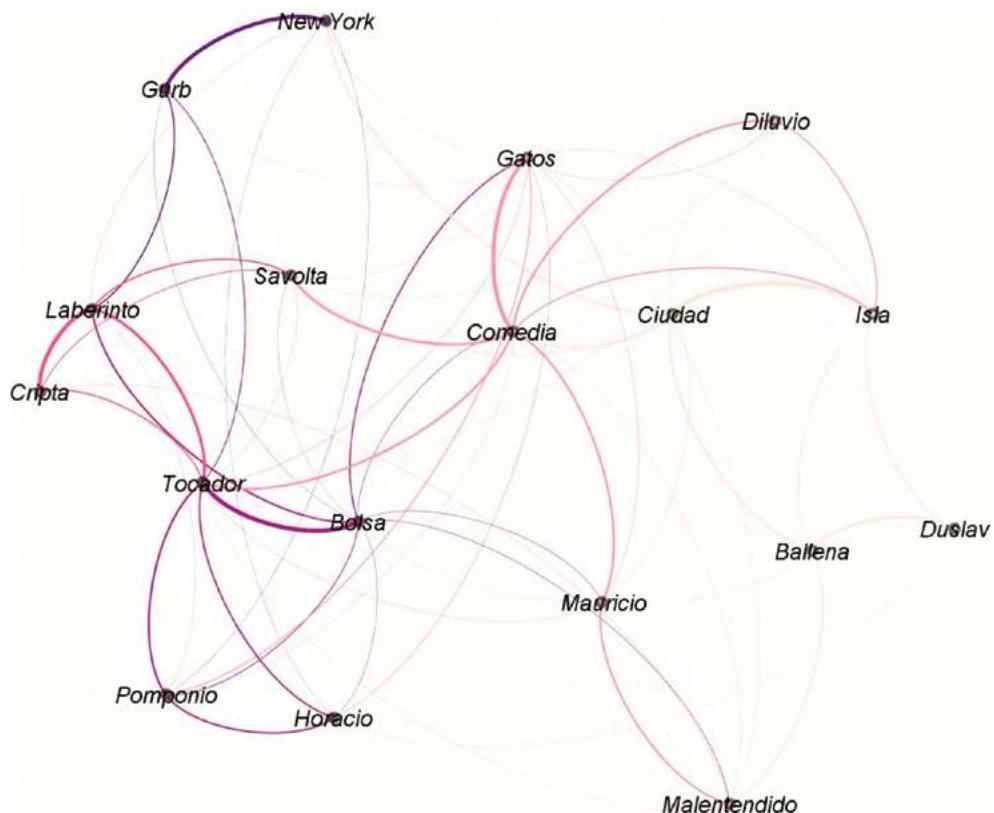


FIGURA 10
Mapa de redes nodales con distancia Classic Delta con 5.000 MFW

4. CONCLUSIONES

La herramienta *Stylo* se ha mostrado eficaz para la agrupación tipológica y estilística de un corpus de obras literarias de un mismo autor aplicando la configuración de análisis de 5.000 MFW y la distancia estadística Classic Delta.

El gráfico de árbol de consenso resultante demuestra la vinculación estilística de las novelas paródicas frente a las novelas pastiche (agrupadas bajo nodos opuestos), y de las novelas del detective Ceferino respecto del resto de parodias. También ha sido capaz de establecer diferencias de estilo cronológicas: las cuatro novelas de Ceferino, por ejemplo, las ha subdividido en dos grupos respondiendo a la publicación en lapsos de tiempo notablemente distantes.

No obstante, hemos comprobado que el conocimiento previo y profundo de la tipología genérica del corpus a analizar es un requisito necesario para la configuración de la he-

rramienta, ya que para poder delimitar correctamente los valores mínimos y máximos de «tokens», o la conveniencia o no de una determinada distancia estadística, deberemos contrastar cada gráfico resultante con el conocimiento previo que del corpus tenga el investigador, lo que vincula la estilometría basada en macroanálisis con el estudio estilístico tradicional basado en lectura cercana.

5. REFERENCIAS BIBLIOGRÁFICAS

- BASTIAN, MATHIEU, HEYMANN, SEBASTIEN y JACOMY, MATHIEU (2009). «Gephi: an open source software for exploring and manipulating networks». *International AAAI Conference on Weblogs and Social Media*. En línea: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.
- EDER, MACIEJ, KESTERMONT, MIKE and RYBICKI, JAN (2013). «Stylometry with R: a suite of tools». *Digital Humanities: Conference Abstracts*. University of Nebraska, Lincoln, NE, p. 487-89.
- EDER, MACIEJ, RYBICKI, JAN and KESTERMONT, MIKE (2016). «Stylometry with R: A Package for Computational Text Analysis». *R Journal*, Vol. 8, Issue 1, p. 107-121. En línea: <https://journal.r-project.org/archive/accepted/eder-rybicki-kestermont.pdf>.
- FRADEJAS, JOSÉ MANUEL (2016). «El análisis estilométrico aplicado a la literatura española: las novelas policíaca e histórica». *Caracteres. Estudios culturales y críticos de la esfera digital*. En línea: <http://revistacaracteres.net/revista/vol5n2noviembre2016/analisis-estilometrico/>
- JOCKERS, MATHEW (2013). «Macroanalysis: Digital Methods and Literary History». *Topics in the Digital Humanities*. Chicago, University of Illinois Press, p. 108.
- LUTOLAWSKI, WINCENTY (1898). «Principes de stylométrie», *Revue des Études Grecques*, volume 11, numéro 41, p. 61-81.
- KENNY, ANTHONY (1982). *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. Oxford, Pergamon Press, p. 12.
- MENDOZA, EDUARDO (1975). *La verdad sobre el caso Savolta*, Barcelona, Seix Barral.
 — (1979). *El misterio de la cripta embrujada*, Barcelona, Seix Barral.
 — (1982). *El laberinto de las aceitunas*, Barcelona, Seix Barral.
 — (1986). *Nueva york*, Barcelona, Destino.
 — (1986). *La ciudad de los prodigios*, Barcelona, Seix Barral.
 — (1989). *La isla inaudita*, Barcelona, Seix Barral.
 — (1991). *Sin noticias de Gurb*, Barcelona, Seix Barral.
 — (1992). *El año del diluvio*, Barcelona, Seix Barral.
 — (1996). *Una comedia ligera*, Barcelona, Seix Barral.
 — (2001). *La aventura del tocador de señoras*, Barcelona, Seix Barral.
 — (2002). *El último trayecto de Horacio Dos*, Barcelona, Seix Barral.
 — (2006). *Mauricio o las elecciones primarias*, Barcelona, Seix Barral.
 — (2008). *El asombroso viaje de Pomponio Flato*, Barcelona, Seix Barral.
 — (2009). *Tres vidas de santos*, Barcelona, Seix Barral.
 — (2012). *El enredo de la bolsa y la vida*, Barcelona, Seix Barral.
 — (2015). *Riña de gatos: Madrid 1936*, Barcelona, Planeta.
- MENDENHALL, THOMAS CORWIN (1901). «A Mechanical Solution of a Literary Problem». *The Popular Science Monthly*. LX (7), p. 97-105.
- STAMATATOS, EFSTATHIOS (2009). «A survey of modern authorship attribution methods». *Journal of the American Society for Information Science and Technology*, Volume 60, Issue 3, p. 538-556. En línea: <http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf>