Revista de Humanidades Digitales

Estilometría con fines geolingüísticos aplicada al corpus COSER

Stylometry for Geolinguistic Purposes Applied to the COSER Corpus

Pablo PEÑARRUBIA NAVARRO
Barcelona Supercomputing Center
pablop16n@gmail.com
https://orcid.org/0000-0002-1011-0110

RESUMEN

En el presente trabajo se propone la utilización de las técnicas estilométricas modernas para el estudio y clasificación de geolectos en español, realizando un análisis de Estilometría sobre el corpus COSER. Se utiliza el programa Stylo sobre una muestra de entrevistas transcritas para extraer parámetros fiables de categorización. Al final, se han aplicado los resultados concluyentes sobre todo el corpus, consiguiendo una clasificación no aleatoria que, incluso, es explicable desde algunos trabajos dialectales de referencia .

PALABRAS CLAVE

Estilometría, geolingüística, Stylo, clasificación dialectal, PLN.

ABSTRACT

In the present work, we propose the use of modern stylometric techniques for the study and classification of dialects in Spanish, by carrying out a stylometric analysis using the COSER corpus. We use Stylo on a sample of the transcribed interviews to extract reliable classification parameters. In the end, the outcomes have been applied to the entire corpus, achieving a non-random classification that is explainable considering some prestigious dialectal works.

KEYWORDS

Stylometry, Geolinguistics, Stylo, Dialect Classification, NLP.

RHD 6 (2021)

ISSN
2531-1786

Recibido 20/05/2021 – Aceptado 26/07/2021

1. INTRODUCCIÓN

Los métodos computacionales han revolucionado los estudios relacionados con el lenguaje, puesto que permiten procesar grandes cantidades de datos. Todas las ramas de investigación lingüística han querido explotar las posibilidades que ofrece el procesamiento automático de datos y el cálculo estadístico de los mismos, especialmente después de los enormes avances en el campo del Procesamiento del Lenguaje Natural y de la Lingüística Computacional. Podemos encontrar ejemplos actuales en áreas como la Lingüística Forense (Crespo Miguel, 2017; Queralt Estévez, 2020), la Glotocronología (Muñoz Acebes, 2019), la Geolingüística (Julià Luna, 2020) o los estudios literarios (Calvo Tello, 2019; Hernández Lorenzo, 2019). Por tanto, también evidenciamos que el interés por la investigación asistida por ordenador no solo no ha perdido vigencia, sino que parece estar estableciéndose como práctica básica en las distintas ciencias interesadas por el lenguaje humano.

Una de las técnicas computacionales más utilizadas en los estudios lingüísticos y literarios es la Estilometría. Se trata de un método de clasificación lingüística que goza de mucha popularidad en los últimos años, aunque sus orígenes se remontan a mediados de los sesenta. Se fundamenta en la idea de que el autor de un texto imprime siempre en sus creaciones una huella estilística o autorial, un estilo propio que puede ser rastreado por medio de métodos cuantitativos (Fradejas Rueda, 2016, pp. 199-205). A pesar de que esta técnica es utilizada principalmente para la atribución de autoría de obras literarias, es empleada también por la Lingüística Forense. En este sentido, permite, por ejemplo, determinar la autoría de una prueba judicial, detectar plagios o establecer perfiles lingüísticos (Eder et al., 2015; Dunn et al., 2015). Además, ya que el estilo es en principio inconsciente, resulta muy difícil de manipular deliberadamente para la mayoría de hablantes o escritores. Por otro lado, la Estilometría también puede ser utilizada para el estudio de la variación idiolectal o intrapersonal a lo largo del tiempo, con la finalidad de detectar cambios sustanciales en el estilo de un mismo autor: es la llamada Cronoestilometría (Stamou, 2008).

Por lo general, los estudios estilométricos utilizan los propios textos para extraer los datos que les permitirán categorizar las muestras por similitud estadística. Se suele medir parámetros tales como la longitud de caracteres por palabra[1], la cantidad de palabras por frase o las palabras más frecuentes, así como sus distintas combinaciones (n-gramas), las cuales permiten, por ejemplo, detectar colocaciones y locuciones.

El interés por la Estilometría ha llegado hasta tal punto que incluso se han creado programas que automatizan todo el proceso, tanto el procesamiento textual, como el cálculo y representación de datos. Dentro de estos programas podemos destacar Stylo (Eder et al., 2016) o JGAAP (Juola, 2009).

En resumen, las técnicas estilométricas tienen por finalidad medir la variación idiolectal de un grupo de textos, con el objetivo de clasificarlos en razón del estilo autorial. Sin embargo, la va-

[1] En Estilometría se entiende la palabra en sentido gráfico, como conjunto de caracteres separados por espacios, ya que es la manera más sencilla de computar los textos.

riación idiolectal no es la única que existe en el lenguaje humano. En efecto, los hablantes de todas las lenguas naturales muestran variaciones entre individuos (idiolectales), personas de distinto género, condición social o edad (sociolectales) y, lo que es más importante para este trabajo, también muestran variación geográfica (geolectal/dialectal) (Moreno Fernández, 1998-2009). La competencia lingüística de los usuarios de una lengua difiere a razón de la distancia que los separe y de los elementos geográficos que medien entre ellos (Hernández Campoy, 2008, pp. 521-525).

Llegados a este punto, podríamos preguntarnos si todos los tipos de variación lingüística son compatibles con las técnicas estilométricas. Teóricamente, la respuesta debería ser afirmativa, puesto que los sociolectos y los geolectos son grupos abstractos, formados realmente por la suma de idiolectos, al igual que una población está lógicamente compuesta por individuos. De esta manera, si los idiolectos son clasificables a través del uso de técnicas estilométricas, no hay ninguna razón para pensar que los textos de hablantes sociolectal o geolectalmente diferentes no puedan ser clasificables de esta manera. En cualquier caso, esto no quiere decir que podamos exigir a la Estilometría que consiga altas tasas de éxito con, por ejemplo, hablantes de distinta clase social o edad sin aislar previamente o, al menos, controlar el resto de variables que podrían afectar al análisis.

Para los estudios de variación geográfica, dentro del marco de la Geolingüística[2] o Dialectología moderna, la clasificación automática de datos por medios computacionales no es algo nuevo. Tampoco les son ajenos los cálculos con distancias geométricas, muy usados en Estilometría, o los análisis de grupos (Aurrekoetxea Olabarri, 2019). En este sentido, desde los trabajos de Goebl (1981) hasta ahora, se ha desarrollado la llamada Dialectometría. Este método utiliza grandes volúmenes de muestras lingüísticas, obtenidas, principalmente, de atlas para generar clasificaciones de hablantes, a partir de frecuencias de uso (Moreno Fernández, 2003, pp. 10-12).

Por tanto, podemos ver que la Estilometría y la Dialectometría son muy similares. Ambas pretenden convertir datos lingüísticos en datos numéricos y, con ellos, establecer relaciones de semejanza. Además, debemos tener en cuenta que el geolecto y el idiolecto parten de una misma premisa, esto es, existen rasgos distintivos entre los hablantes/escritores que permiten clasificarlos. La principal diferencia entre ambas técnicas la encontramos en el origen y el tipo de los datos. Mientras la Dialectometría parte de los rasgos preseleccionados y extraídos por lingüistas, la Estilometría utiliza directamente los textos de interés, sin necesidad de que estos tengan notaciones manuales, etiquetados o algún trato particular.

Lo que se propone en este trabajo es el uso de la Estilometría para el estudio de la variación geográfica. Aunque estos métodos han sido ya utilizados con este tipo de variación en la perfilación forense (Crespo Miguel, 2017), creemos que no han sido suficientemente explotados para

[2] Ya que el propósito de este trabajo tiene que ver con la variación lingüística motivada por la distancia, no trataremos otros tipos, como la variación social o incluso contextual (diafásica). Tampoco podemos atender a los accidentes geográficos, aunque pueden influir notablemente en las lenguas, puesto que no disponemos para este trabajo de medios para hacerlo.

la creación de mapas geolectales extensos. No se quiere con ello sustituir a la Dialectometría, puesto que, como hemos dicho, cada una parte de datos distintos. Lo que se propone es, más bien, el uso de la Estilometría como método complementario, destinado a clasificar textos dialectales que no cuenten con etiquetados de rasgos lingüísticos.

2. METODOLOGÍA

Para llevar a cabo un primer acercamiento a nuestra propuesta, utilizaremos la librería Stylo[3] para el lenguaje de programación R sobre el Corpus Oral y Sonoro del Español Rural (COSER) (Fernández-Ordóñez Hernández, 2005). De este modo se quiere, por un lado, comprobar si los métodos estilométricos son sensibles a la variación léxica geolectal del español de España y, por otro, saber si estas herramientas son capaces de hacer agrupaciones no aleatorias de hablantes.

2.1. Primer análisis

En primer lugar, se aplicarán algunas de las funcionalidades de Stylo sobre tres grupos de muestras extraídos del COSER. Cada grupo contiene las grabaciones de nueve provincias distintas y han sido escogidas de manera aleatoria. Tan solo se han aplicado dos condiciones: que no fueran limítrofes entre sí y que, además, pertenecieran a comunidades autónomas distintas[4]. Lo que le pedimos al programa es que agrupe las grabaciones por provincia correctamente. De este modo, extraeremos los mejores parámetros de clasificación para aplicarlos posteriormente en un segundo análisis con todo el corpus.

Figura 1. Detalle del mapa de archivos del COSER. Fuente: http://www.corpusrural.es/archivos.php.

[3] Versión 0.7.4.

[4] En los casos de Canarias, Baleares, Galicia y Cataluña se ha tomado toda la comunidad como una sola provincia debido al pequeño número de entrevistas.

El sentido de exigir provincias alejadas entre sí reside en que, como tan solo podemos identificar las entrevistas por el etiquetado de provincia que ofrece el COSER, no solo no podemos utilizar la distancia como unidad de medida, sino que dependemos de las delimitaciones políticas. Así, con este condicionante, si atendemos a la figura 1, veremos cómo algunas localidades registradas se encuentran muy cercanas, a pesar de pertenecer a provincias distintas. Si no aplicáramos la imposición de *no limítrofes* y se creara una agrupación entre las seis grabaciones que se ven al centro de la imagen, podríamos considerarlo un fallo en el cálculo estilométrico. No obstante, es muy probable que exista una relación lingüística entre estas seis localidades, puesto que se encuentran muy próximas entre sí y la única barrera que existe *a priori*, la delimitación política, poco o nada tiene que ver con los límites geolectales.

Además, hay que tener en cuenta el continuum lingüístico. Las lenguas no difieren de manera discreta, sino continua. Las hablas que se distinguen entre sí por motivos geográficos se distribuyen a modo de transiciones, donde es muy difícil establecer límites (Hernández Campoy, 1993). Si elegimos provincias fronterizas en este primer análisis, el programa se encontraría con el problema de categorizar en un grupo concreto localidades que podrían estar a caballo entre otros grupos más grandes. En ese caso, la decisión del programa variaría arbitrariamente

Por otro lado, la condición *distinta comunidad autónoma* nos permite crear un área de muestras lo suficientemente amplia como para abarcar todo el país, consiguiendo la mayor representatividad posible.

En cuanto a las pruebas, ya que existen muchos métodos y medidas, se intentarán explorar las diferentes opciones hasta encontrar las que ofrezcan los mejores resultados de categorización. Creemos importante realizar muchas pruebas con distintos parámetros y no usar simplemente los más populares, ya que no sabemos cómo reaccionarán los cálculos a una tipología textual tan diferente a la usada habitualmente en otros trabajos estilométricos. Para encontrar estos parámetros idóneos se realizará, en primer lugar, una exploración manual, utilizando la función stylo() del paquete Stylo para R. Con ello se pretende corroborar que las clasificaciones estilométricas de estos textos dialectales no son aleatorias y, además, se quiere conocer a grandes rasgos el comportamiento de los datos al cambiar los parámetros. Una vez hecho esto realizaremos una validación cruzada por medio de un script en R, utilizando la función crossv() de Stylo, con la finalidad de obtener los mejores parámetros de clasificación de una manera objetiva y contrastada.

2.2. Segundo análisis

Una vez hecho esto, utilizaremos los ajustes que hayan demostrado ser los más adecuados en las pruebas anteriores, pero esta vez se aplicarán sobre casi todas las grabaciones disponibles en el COSER, en total 204. Se excluyen aquellas entrevistas inferiores a las 3.000 palabras, teniendo en cuenta la extensión media de los textos (10.183 palabras) y los resultados de otros investigadores (Eder, 2015; Hernández Lorenzo, 2019).

Se realizará una clasificación de análisis de grupos utilizando la función stylo() y se intentará traducir la información del dendrograma resultante a un mapa organizado por colores, para

que el lector pueda comprender mejor los resultados. Buscamos con este paso, observar el comportamiento de los métodos estilométricos con datos más complejos. No vemos conveniente realizar pruebas de parámetros en este análisis, puesto que, al existir provincias limítrofes, existen también los problemas que hemos expuesto en 2.1 y, por tanto, no podemos establecer objetivamente fallos o aciertos. En este sentido, y como hemos argumentado antes, deberíamos encontrarnos con el problema del continuum lingüístico, lo cual provocaría que textos situados en áreas de transición cambiaran de grupo según los parámetros asignados a Stylo. Esto podría, incluso, utilizarse de manera malintencionada, para respaldar teorías dialectales a cerca de la extensión de cierta variedad lingüística, convirtiendo el método estilométrico en una técnica totalmente subjetiva y a conveniencia del investigador (Cuéllar González, 2018, p. 304).

2.3. Elección del corpus

Se ha elegido el corpus COSER entre los muchos que existen por varias razones. La principal ventaja reside en su equilibrio y extensión. COSER cuenta con transcripciones de entrevistas repartidas por toda la geografía española. Además, suele haber al menos 4 o 5 grabaciones por cada provincia. Otra ventaja del mismo es que la metodología que se sigue en la recolección de datos es muy homogénea: entrevistas informales, levemente dirigidas, donde se tratan las mismas temáticas. Todo esto nos permite, por un lado, disminuir al mínimo la variable diafásica o contextual y, por otro lado, reducir la asociación indeseada entre textos que traten casualmente el mismo tema, pero que realmente no contengan rasgos geolectales comunes.

El hecho de ser un corpus exclusivamente rural también nos ofrece una ventaja importante. Además de la variación diafásica y diatópica de las lenguas, hay que tener en cuenta la diastrática o social. Gracias a que este corpus solo registra en pequeñas poblaciones, se nos brinda la oportunidad de salvar este obstáculo, porque la variación de tipo diastrático apenas es relevante en poblaciones menores donde las redes sociales son mucho menos complejas. Igualmente podemos prestar menos atención al factor edad, puesto que los entrevistados son por lo general personas de 65 años en adelante.

En cuanto a la transcripción del material, se nos dice en su sitio web que se ha optado por una transcripción ortográfica con poca atención a variaciones de pronunciación dialectales, puesto que su objetivo se centra en la morfosintaxis. Esto nos permite que las clasificaciones se centren en el léxico. De otra manera, si atendieran en exceso a aspectos fonéticos, no podríamos saber nunca si las asociaciones estadísticas se están realizando por criterios léxicos o fonéticos.

Por último, no hay que olvidar la duración de las entrevistas. Los autores de COSER han pretendido que las grabaciones tuvieran una extensión similar[5] y, además, han procurado ceder al máximo el turno de palabra a la persona entrevistada. Por todo ello, pensamos que este corpus es uno de los mejores que podemos utilizar para poner a prueba la Estilometría con fines geolingüísticos.

[5] En la metodología del COSER se indica que la media es de 1 hora y 15 minutos.

2.4. Extracción de datos

La limpieza de los textos[6] se ha realizado utilizando un programa que hemos creado utilizando el lenguaje Python. En él hemos procurado extraer las intervenciones de cualquier hablante local, por tanto, si en una grabación hay tres interlocutores, sin tener en cuenta los entrevistadores, se han extraído indistintamente a los tres. Esto es debido a que, como dijimos, nuestro interés se centra en el geolecto, por lo que nos es indiferente que se mezclen varios idiolectos en una misma muestra. Además, hay que tener en cuenta que en una extracción automatizada es imposible predecir qué interlocutor es el que más nos interesa de entre los distintos intervinientes.

Tanto este código fuente, como los que usaremos después se encuentran en este enlace de GitHub[7] para que puedan ser revisados, reutilizados y mejorados por el lector. Su comprensión requiere de ciertos conocimientos de programación, por ello hemos comentado muchas de las líneas del código, con el fin de que resulte más sencillo, incluso para aquellos menos acostumbrados a este tipo de programas.

2.5. La utilización de Stylo

Stylo es probablemente la herramienta más utilizada en la Estilometría actual. Permite muchos tipos de métodos estadísticos[8] para establecer relaciones entre palabras, caracteres, o las posibles combinaciones de ellos. Stylo permite automatizar el proceso de la tokenización, es decir, la extracción de palabras que se debe realizar antes de empezar con la cuantificación. También resulta de gran utilidad la manera en que ofrece los resultados, a modo de gráficas o mediante listas de datos.

En cuanto a cómo se determina la semejanza entre textos, lo que se utiliza es una serie de marcadores, los z-scores, aplicados sobre las palabras más frecuentes. Estos establecen puntuaciones a partir de medidas estadísticas relativas extraídas de los textos. Se consigue con ello medir la extrañeza de que una palabra aparezca en un texto y en otros no. Al final del proceso, se crea una clasificación, partiendo de los resultados de los z-scores, que se puede representar a modo de dendrograma jerárquico binario. En él cada división a partir del nodo central significa mayor heterogeneidad entre los siguientes grupos creados (Calvo Tello, 2016).

En nuestro caso utilizaremos el método multivariante sin supervisión de análisis de grupos. También se utilizará el análisis de árboles de consenso, el cual se basa en aunar muchos análisis de grupos con parámetros distintos en una sola representación. Y, por último, dentro de las opciones que existen, utilizaremos la validación cruzada, un método que permite establecer índices numéricos de acierto en una clasificación. Este método se suele utilizar en grupos de textos cuyo autor es conocido para validar los parámetros antes de utilizarlos sobre un corpus con autores desconoci-

[6] Para este análisis se trabajó con la versión en formato texto, puesto que aún no se había publicado la versión en XML cuando se analizaron los datos.
[7] Accesible desde: https://github.com/pablop16n/estilometria-dialectal.
[8] Para más información sobre el funcionamiento de estos métodos se puede consultar Calvo Tello (2016) y Evert et al. (2017).

dos. En nuestro caso podemos aprovechar este primer paso, puesto que nuestro autor, la provincia, es conocido en todos los casos.

RESULTADOS

3.1. Primer análisis

3.1.1. Fase manual: análisis de grupos y árbol de consenso

Después de muchos ensayos, explorando parámetros distintos en la GUI de stylo(), obtuvimos asociaciones muy alejadas de la aleatoriedad a partir de las 2.000 palabras más frecuentes (MFW) en la mayor parte de las distancias Delta. No obstante, los mejores resultados de clasificación se consiguieron con Cosine Delta[9] (Jannidis et al., 2015): 5.000 palabras más frecuentes, monogramas[10] y fuerza de consenso de 0,5. Mostramos el análisis de grupos que obtuvo mejores resultados en cada uno de los tres grupos, figura 2, figura 3 y figura 4:

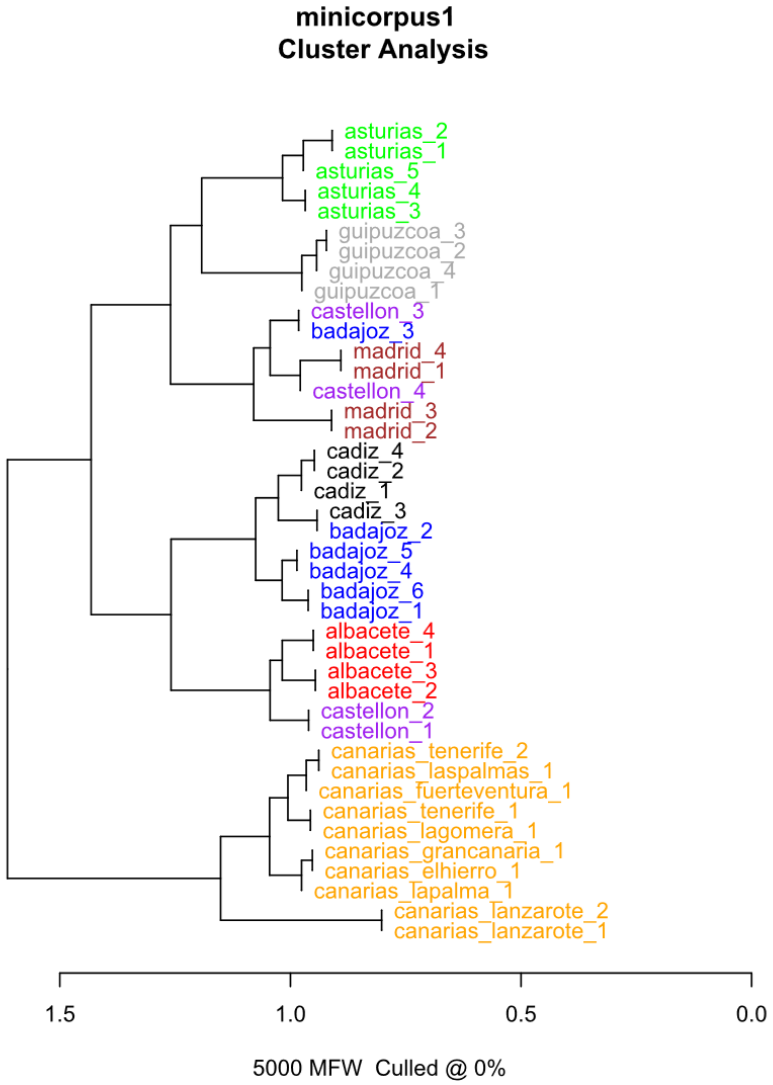


Figura 2. Análisis de grupos con Stylo, grupo 1. Fuente: elaboración propia.

[9] Otros autores también coinciden en el uso de esta medida, aunque aplicada al inglés (Evert et al., 2017).
[10] Crespo Miguel (2017) igualmente concluye que los monogramas de palabras funcionan mejor que otros parámetros con textos similares a los que usamos aquí.

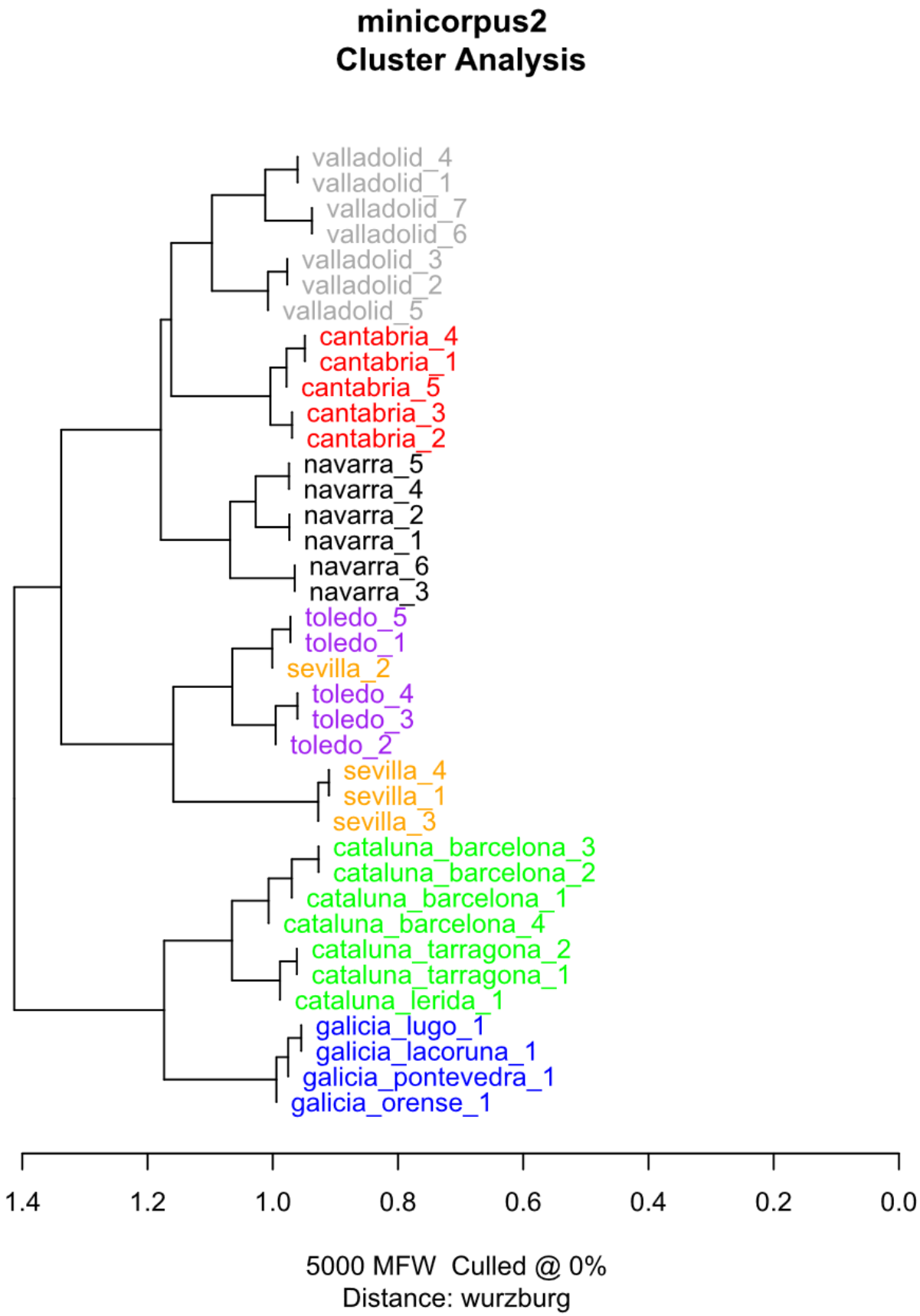


Figura 3. Análisis de grupos con Stylo, grupo 2. Fuente: elaboración propia.

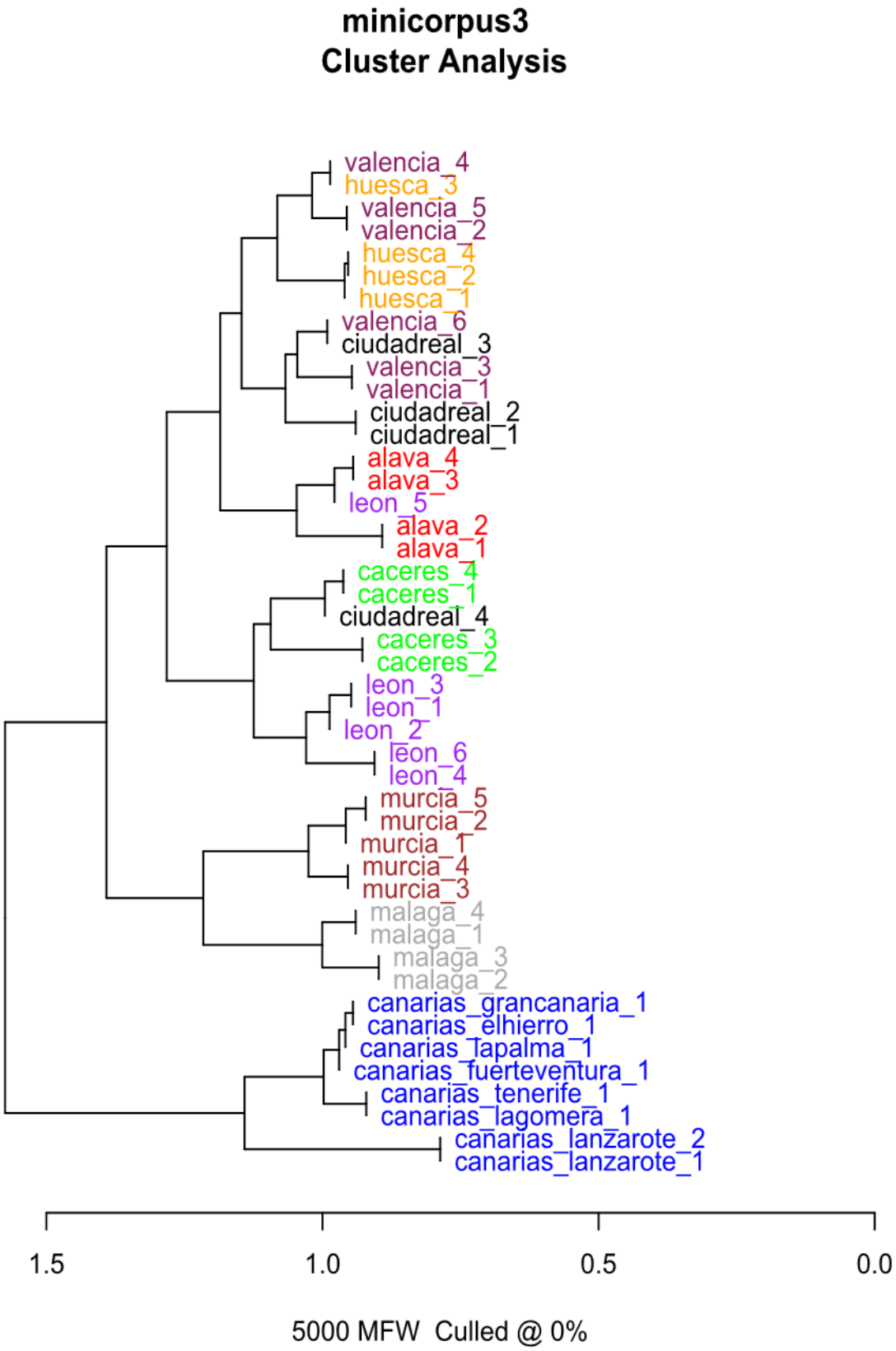


Figura 4. Análisis de grupos con Stylo, grupo 3. Fuente: elaboración propia.

Como se puede observar, se generan asociaciones poco aleatorias entre las diferentes provincias. Es destacable la Figura 3, donde solo se ha obtenido un fallo en la clasificación.

En cuanto a los árboles de consenso, los resultados fueron, por lo general, peores que los obtenidos con análisis de grupos de 5.000 palabras.

3.1.2. Fase automatizada: validación cruzada

Una vez explorados los textos manualmente podemos calibrar adecuadamente el estudio automatizado de validación cruzada. Para realizarlo hemos empleado una versión modificada del script de R presentado en Cuéllar González (2018). La propuesta de este autor permite lanzar una gran cantidad de validaciones cruzadas en modo *leaveoneout* con parámetros distintos, de tal manera que se llega a abarcar la práctica totalidad de posibilidades. Los datos resultantes se guardan en un archivo de texto, de tal modo que se pueden comparar y extraer conclusiones.

En nuestro caso hemos modificado este programa para que además de ir probando distintos Culling (fuerza de consenso[11]) y MFW (palabras más frecuentes), abarque automáticamente todas las distancias Delta disponibles[12]. Además, gracias a la exploración en la fase manual, determinamos que debíamos aumentar el máximo de MFW, que estaba ajustado en 1.000 palabras, hasta las 5.000. Por otro lado, también creamos un pequeño programa de Python para ordenar de mayor a menor tasa de éxito los resultados.

Mostramos a continuación, en las tablas 1 y 2, los resultados de los tres mejores datos de acierto por cada grupo en 1-gramas:

Corpus	Muestra 1	Muestra 2	Muestra 3
Acierto - Puesto 1	90%	97%	96%
Parámetros	Wurzburg Culling: 0 MFW: 2900-3000-4000-4100-4200-4300-4400-4500-4600-4700-4800-4900-5000	Wurzburg Culling: 0 MFW: 2000-2100-2200-2300-2400-2500-2600-2700-2800-2900-3000-3100-3200-3500-3600-3700-3800-3900-4100-4200-4600-4700-4800-4900-5000 Wurzburg Culling: 10 MFW: 2000-2100-2200-2300-2500-2600-2700-2800-2900-3000-3100	Wurzburg Culling: 0 MFW: 4900

[11] El Culling sirve para vetar los tokens que no comparten entre sí los textos, se mide en una escala porcentual del 0 al 100, dependiendo del porcentaje de textos que queremos que compartan como mínimo un determinado token.

[12] Se ha excluido la opción de *minmax* porque generaba muchos problemas durante la ejecución. En cualquier caso, todas las pruebas que se pudieron hacer con esta distancia Delta no mostraron una tasa de éxito mejor que el resto .

Acierto - Puesto 2	88%	95%	93%
Parámetros	Wurzburg Culling:0 MFW: 2500-2600-3100-3500-3600-3700-3800-3900	Wurzburg Culling: 0 MFW: 1900-3300-3400-4000	Wurburg Culling:0 MFW: 2900-3000-3100-3200-3300-3400-3600-3700-3800-4200-4300-4400-4700-4800
		Wurzburg Culling: 10 MFW: 1400-1800-1900-2400	
		Wurzburg Culling: 20 MFW: 1500	

Tabla 1. Mejores 3 (1-2) resultados obtenidos en la validación cruzada con crossv() de Stylo (monogramas). Fuente: elaboración propia.

Corpus	Muestra 1	Muestra 2	Muestra 3
Acierto Puesto 3	85%	92%	91%
Parámetros	Wurzburg Culling: 0 MFW: 1500-1600-1700-1800-1900-2000-2100-2200-2700-3200	Wurzburg Culling: 0 MFW: 1300-1400-1500-1600-1700-1800	Wurzburg Culling: 0 MFW: 2200-2300-2500-2800-3500-3900-400-4100-4500-4600-5000
	Wurzburg Culling: 10 MFW: 1500-1700-1800-1900-2100-2300-2400	Wurzburg Culling: 10 MFW: 1300-1500-1600-1700	
		Wurzburg Culling: 20 MFW: 1400	Wurzburg Culling: 10 MFW: 2900-300-3200-3600
		Delta simple Culling: 0 MFW: 1600-1800-1900-2000-2100-2200-2300-2400-2600-2700-2800-3200-33003400	

Tabla 2. Mejores 3 (3) resultados obtenidos en la validación cruzada con crossv() de Stylo (monogramas). Fuente: elaboración propia.

De estos datos se puede obtener una conclusión clara: Cosine Delta, denominado aquí W*urzburg*, es la distancia que mejor funciona con estos textos. En cuanto a los rangos de palabras más frecuentes, se puede observar que la conclusión que extrajimos antes a cerca de la necesidad de aplicar MFW altos era cierta. En efecto, y por lo general, cuanto más nos acercamos a las 5.000 palabras, mayor tasa de acierto conseguimos. Por lo que respecta al Culling, este parece ser poco útil a la hora de obtener aciertos.

A los lectores interesados por la Estilometría les podrán resultar extraños los resultados. Los rangos de MFW más utilizados suelen ser más bajos y, además, el Culling suele resultar útil para evitar asociaciones por motivos casuales, como que dos novelas compartan nombres en algunos de sus protagonistas. Esto es así porque en Estilometría lo que interesa medir, y lo que suele diferenciar dos estilos autoriales, son las palabras gramaticales, muy frecuentes y compartidas por cualquier tipo de escrito o discurso. Sin embargo, parece que al trabajar con el estilo geolectal, y no con el idiolectal, ocurre prácticamente al contrario: cuantas más palabras no gramaticales o poco frecuentes mejor categorización. Esto es así hasta tal punto que si eliminamos las 100 palabras más frecuentes de cada entrevista[13], obtenemos muy poca diferencia en el índice de acierto (Tabla 3), por lo que entendemos que estos *tokens* frecuentes apenas tienen peso en la clasificación.

Corpus	Muestra 1	Muestra 2	Muestra 3
Acierto sin 100MFW	95%	97%	93%
Acierto con 100MFW	90%	97%	96%[14]

Tabla 3. Mejor resultado obtenido en la validación cruzada con crossv() de Stylo (100 primeras MFW eliminadas). Fuente: elaboración propia.

En nuestra opinión, esto puede estar relacionado con el hecho de que en el léxico la clase cerrada, concepto muy relacionado con el de palabra gramatical, sufre menos variación que la clase abierta. Por citar solo un ejemplo, en español el nombre de muchos objetos cotidianos cambia enormemente de región en región (fogón, fuego, hornalla, estufa, etc.), pero la variación en el uso de preposiciones o determinantes es prácticamente inexistente. Por tanto, los cambios en las frecuencias de uso de estas palabras gramaticales estarían más ligados al estilo autorial o idiolectal que al geolectal. En cuanto al resto de n-gramas, los bigramas y trigramas se han mostrado ligeramente peores que los monogramas. Mostramos en la Tabla 4 solo los resultados máximos:

Corpus	Muestra 1	Muestra 2	Muestra 3
Acierto bigramas	88%	97%	80%
Acierto trigramas	88%	84%	71%

Tabla 4. Mejor resultado obtenido en la validación cruzada con crossv() de Stylo (bigramas y trigramas). Fuente: elaboración propia.

[13] Las 100 primeras palabras más frecuentes se han eliminado con un script de Python y luego se ha ejecutado la validación cruzada.

[14] Hay que recordar aquí que solo un parámetro obtuvo este resultado, el resto no llegó más allá del 93%.

3.2. Segundo análisis

Para el corpus completo del COSER se utilizaron los parámetros que hemos considerado más eficientes, a partir de los datos que se han mostrado en el apartado anterior: Cosine Delta, 5.000 MFW, 1-gramas y fuerza de consenso de 0. En este caso nuestra intención no es que Stylo agrupe las provincias correctamente, puesto que los límites políticos no se relacionan directamente con las transiciones lingüísticas. Tampoco queremos forzar un análisis que apoye cierta teoría con respecto a la distribución de isoglosas o áreas geolectales, sino realizar un estudio lo más objetivo y desinteresado posible. En consecuencia, consideraremos como exitosos los resultados que muestren coherencia espacial, es decir, que exhiban una relación clara entre las clasificaciones estilométricas y las distancias que separan las grabaciones, así como otras consideraciones puramente lingüísticas.

Debido a la gran cantidad de textos, hemos querido convertir el dendrograma ofrecido por Stylo (Anexo I) en un mapa, ya que nos parece más adaptado a nuestras necesidades y seguramente resulte más amigable para el lector. De este modo, se ha creado un mapa con la ayuda de Maphub[15], para el que hemos utilizado varias gamas cromáticas que intentan representar las relaciones jerárquicas de los nodos del dendrograma. Así, cada entrevista ha sido representada en el mapa con un color a razón del nodo al que pertenece.

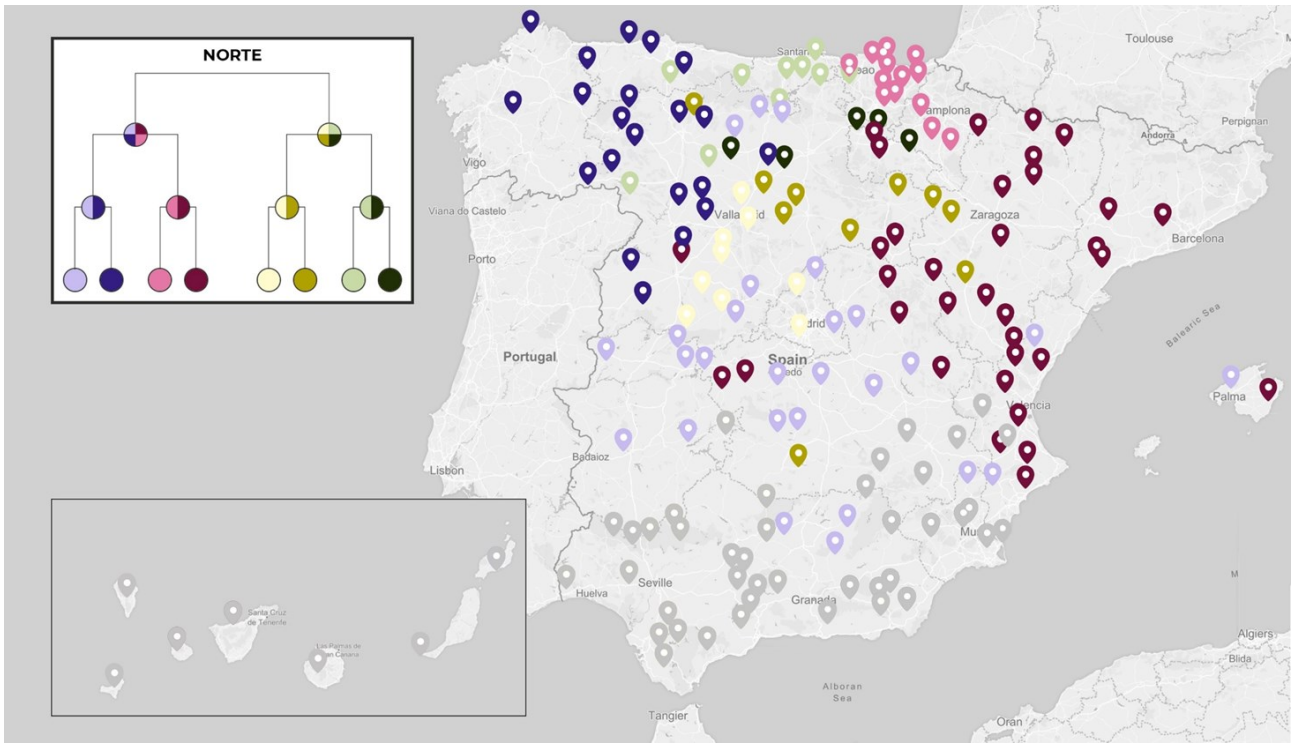


Figura 5. Primera división jerárquica ofrecida por el análisis de grupo con Stylo, grupo norte. Fuente: OpenStreetMap contributors.

[15] Accesible desde: https://maphub.net.

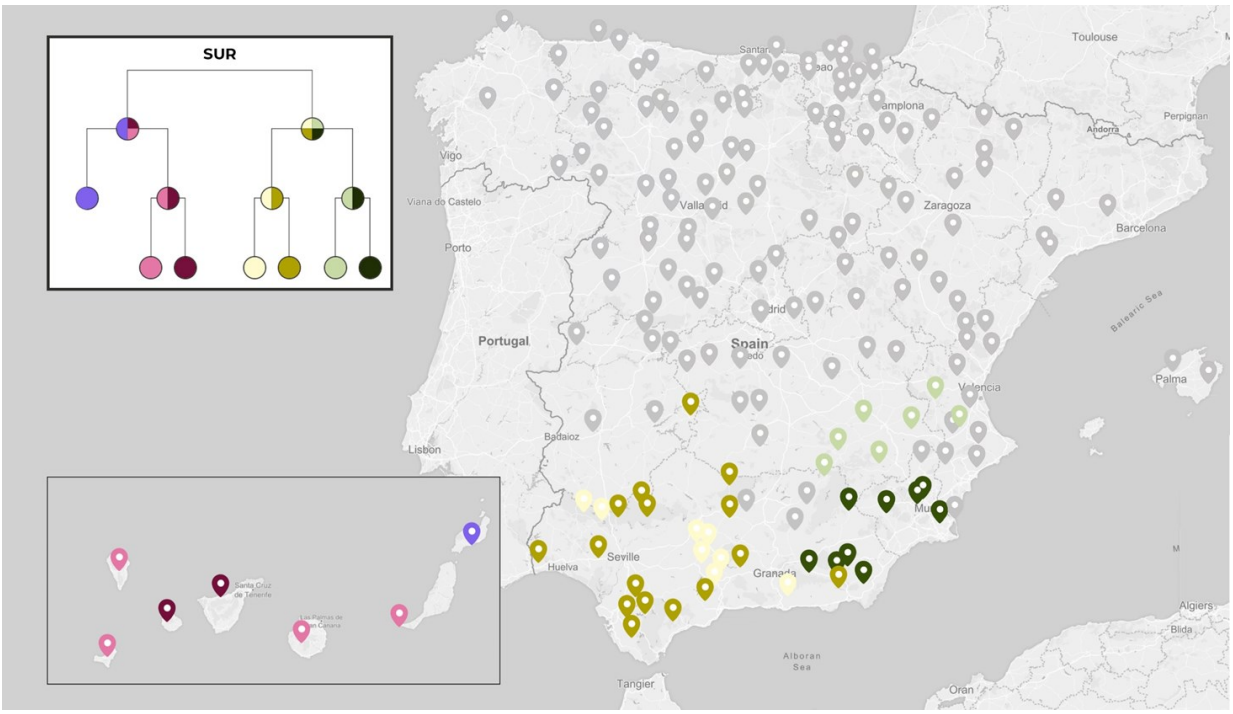


Figura 6. Primera división jerárquica ofrecida por el análisis de grupo con Stylo, grupo sur. Fuente: OpenStreetMap contributors.

En cuanto a los resultados, en primer lugar, Cosine Delta ha dividido el total en dos grupos de una manera muy parecida a la clásica diferenciación entre español meridional y septentrional de la Dialectología tradicional. Además, estas dos áreas son muy homogéneas, ya que las pocas interferencias que existen se dan precisamente en una zona intermedia de transición. Podemos ver el grupo norte en la figura 5 y grupo sur en la figura 6. Por lo general, las áreas que se proyectan están asociadas a espacios geográficos concretos de una manera incompatible con la arbitrariedad.

3.2.1. Norte

En cuanto al primer nivel jerárquico del mapa norte, el grupo verdoso norte se sitúa en el centro norte y el grupo violáceo norte ocupa el resto (figura 7).

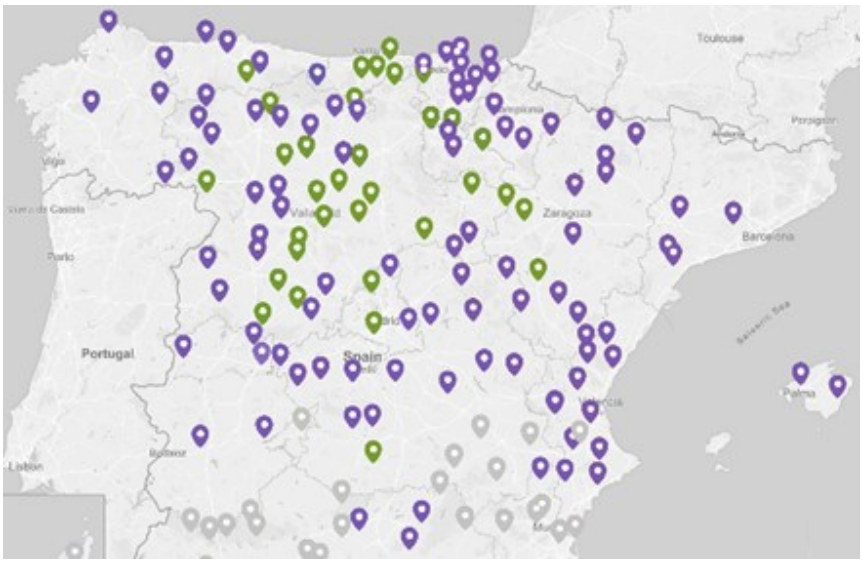
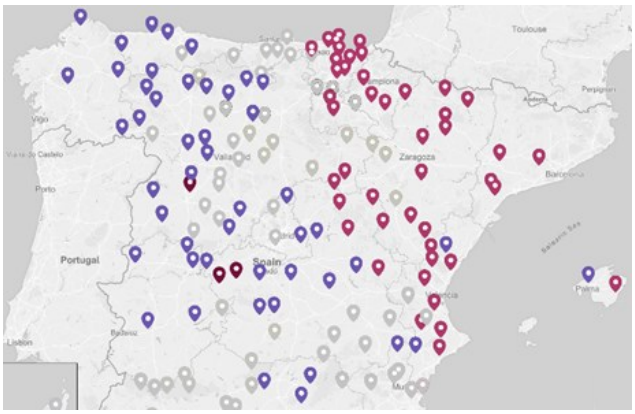


Figura 7. División del grupo norte. Fuente: OpenStreetMap contributors.

El grupo violáceo norte se parte en dos a este y oeste (figura 8). A su vez, estas dos columnas presentan, cada una, una tajante división norte-sur (figura 9).



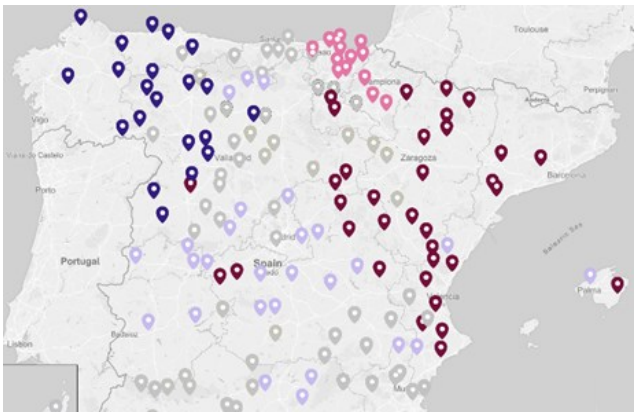


Figura 8. División del grupo violáceo norte. Figura 9. Las cuatro divisiones terminales del grupo violáceo norte. Fuente: OpenStreetMap contributors.

Hay que destacar que algunas de estas áreas no solo tienen coherencia espacial, sino que además tienen cierto sentido desde el punto de vista dialectal:

	Este grupo (Figura 9), situado al noroeste, parece coincidir con las áreas gallego-asturiano-leonesas, claramente relacionadas histórica y lingüísticamente (García de Diego, 1959, pp. 54-190).
	Esta (Figura 9), que abarca País Vasco y Navarra, probablemente sea la más homogénea de todas las agrupaciones y debe tener relación con el bilingüismo euskera-castellano. Además, la clasificación secundaria de esta junto con el grupo color vino, al este, es lógica desde el punto de vista de la Dialectología, puesto que se suele relacionar Navarra y La Rioja con las variedades aragonesas (González, 1996/2016, pp. 305-316).
	Parecen incluirse en este bloque (Figura 9), que se aglomera principalmente al este, las regiones aragonesas y también aquellas influidas por el catalán/valenciano/balear, algo que no es de extrañar debido a la interconexión entre estas dos zonas lingüísticas (Martín y Fort, 1996/2016, pp. 293-304).

Tabla 5. Áreas destacadas del grupo violáceo norte (Figura 9). Fuente: elaboración propia.

3.2.2 Sur

La primera clasificación del mapa sur ha separado Canarias[16] de la Península. Esta, a su vez, se ha dividido en este y oeste (figura 10), algo que coincide con muchos de los datos andaluces en los atlas hispánicos (Alvar, 1996-2016, pp. 245-256; Zamora, 1960-1985, pp. 288-331).

[16] No vemos pertinente comentar la clasificación interna de Canarias, debido a que los datos no son suficientes para extraer conclusiones

Figura 10. División del grupo verdoso sur. Figura 11. Las cuatro divisiones terminales del grupo verdoso sur. Fuente: OpenStreetMap contributors.

Se pueden destacar algunas agrupaciones hechas por Stylo en este grupo sur:

	Dentro de la agrupación oriental (Figura 11) se ha establecido relación con el este andaluz, Murcia, Albacete y, en menor medida, con Valencia. Esta unión para algunos autores no sería casual, pues como afirma García de Diego (1959), "El andaluz […] no ha de estudiarse solamente en las fronteras del reino de Andalucía, sino también dentro de las provincias de Murcia y Albacete" (p.350).

Tabla 6. Áreas destacadas del grupo verdoso sur (Figura 11). Fuente: elaboración propia.

4. CONCLUSIONES

Con este estudio se ha querido explorar otras funcionalidades para los métodos estilométricos modernos, que hasta hoy han sido principalmente utilizados para clasificar textos a partir del estilo personal del autor. Se partió de la hipótesis de que los geolectos no son más que la suma de estilos o idiolectos distantes geográficamente. Así, se propuso estudiar la variación geolectal utilizando las técnicas habitualmente empleadas para la investigación estilística.

Para validar la propuesta, se ha puesto en marcha un experimento estilométrico con Stylo aplicado al COSER, un corpus que cuenta con entrevistas dialectales grabadas por toda España. En primer lugar, se ha procedido a la limpieza de las transcripciones. Después, ha comenzado el experimento en sí, que consta de dos fases. En la primera, se han clasificado grupos de entrevistas pertenecientes a nueve provincias distantes entre sí, donde se ha conseguido un índice de acierto superior al 90% con Cosine Delta, Culling 0 y 5.000 MFW. En este punto, hemos podido comprobar que los textos dialectales de COSER son mejor clasificables con altas tasas de MFW; se ha visto, además, que las palabras más frecuentes apenas determinan el estilo geolectal de los hablantes de este corpus y, por último, se ha llegado a la conclusión de que el Culling es incluso contraproducente. Por otro lado, en una segunda fase, se han utilizado los resultados del paso anterior para aplicar un análisis de grupos sobre todas las grabaciones del corpus. De esta manera, se han conseguido clasificaciones muy alejadas de la aleatoriedad, dando lugar a áreas lingüísticas homogéneas y, en muchos casos, similares a las propuestas por otros trabajos dialectales anteriores.

Llegados a este punto, parece evidente que la Estilometría y, particularmente, Stylo y las distancias Delta, son capaces de crear clasificaciones geográficas coherentes, incluso utilizando

textos de localidades próximas. Esto abre algunos interrogantes que pueden motivar trabajos futuros. Por ejemplo, queda por saber si los resultados mejorarían notoriamente al añadir grabaciones de más localidades intermedias o si, por el contrario, el continuum lingüístico comenzaría a empeorar las clasificaciones. También sería interesante estudiar qué ocurre con textos mucho más extensos, o con otros más breves, que los que hemos utilizado.

Por último, queremos hablar de los posibles usos de esta Estilometría geo-variacional. Como se ha dicho anteriormente, los métodos estilométricos se utilizan en la investigación del estilo autorial, tanto para estudios literarios como forenses. Este enfoque estilométrico podría tener otros usos. Para empezar, esta puede ser una herramienta complementaria de los trabajos dialectométricos, actuando sobre aquellos corpus menos procesados computacionalmente. Además, la Lingüística Forense también podría verse beneficiada, ya que, con la mejora de estas técnicas, el establecimiento de estándares y el tratamiento de más datos, podría desarrollarse aún más el procesamiento computacional del perfilamiento lingüístico.

REFERENCIAS BIBLIOGRÁFICAS

Alvar López, M. (1996-2016). Andaluz. En M. Alvar López (Dir.), *Manual de dialectología hispánica el español de España* (pp. 245-256). Ariel.

Aurrekoetxea Olabarri, G. (2019). Sobre el valor de la dialectometría en la delimitación de las distancias lingüísticas. *Glosema: Revista Asturiana de Llingüística*, *1*, 19-39. https://doi.org/10.1093/llc/fqt066

Calvo Tello, J. (2016). Entendiendo Delta desde las Humanidades. *Caracteres: Estudios culturales y críticos de la esfera digital*, *5*(1), 140-176. http://revistacaracteres.net/revista/vol5n1mayo2016/entendiendo-delta/

Calvo Tello, J. (2019). Stylometric Classification of Periods and Groups of His Novels. *Romanische Studien*, *6*, 151-163. http://www.romanischestudien.de/index.php/rst/article/view/625

Crespo Miguel, M. (2017). PRESEEA y su aporte a la creación de perfiles lingüísticos en Lingüística forense. *Linred: Lingüística en la Red*, *15*. https://doi.org/10.1093/llc/fqt066

Cuéllar González, A. (2018). La necesidad de la validación cruzada en Stylo y cómo programarla. *Caracteres: Estudios culturales y críticos de la esfera digital*, *7*(2), 301-320. http://revistacaracteres.net/wp-content/uploads/2018/11/Caracteresvol7n2noviembre2018-validacion.pdf

Dunn, J., Argamon, S., Rasooli, A., & Kumar, G. (2016). Profile-Based Authorship Analysis. *Digital Scholarship in the Humanities*, *31*(4), 689-710. https://doi.org/10.1093/llc/fqv019

Eder, M. (2015). Does Size Matter? Authorship Attribution, Small Samples, Big Problem. *Digital Scholarship in the Humanities*, *30*(2), 167-182. https://doi.org/10.1093/llc/fqt066

Eder, M., & Rybicki, J. (2013). Do Birds of a Feather Really Flock Together, or how to Choose Training Samples for Authorship Attribution. *Digital Scholarship in the Humanities*, *28*(2), 229-236. https://doi.org/10.1093/llc/fqs036

Eder, M., Rybicki, M., & Kestemont, J. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, *8*(1), 107-120. https://doi.org/10.32614/RJ-2016-007

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding and Explaining Delta Measures for Authorship Attribution. *Digital Scholarship in the Humanities*. 32(2), ii4-ii16. https://doi.org/10.1093/llc/fqx023

Fernández-Ordoñez Hernández, I. (Dir.). (2005). *Corpus Oral y Sonoro del Español Rural.* www.corpusrural.es

Fradejas Rueda, J. M. (2016). El Análisis Estilométrico Aplicado a la Literatura Española: las Novelas Policiacas e históricas. *Caracteres: Estudios culturales y críticos de la esfera digital*, *5*(2), 196-245. http://revistacaracteres.net/revista/vol5n2noviembre2016/analisis-estilometrico/

García de Diego, V. (1959). *Manual de dialectología española*. Ediciones de Cultura Hispánica.

Goeble, H. (1981). Éléments d'analyse dialectométrique (avec application à l'AIS). *Revue de Linguistique Romane*, *45*, 349-420.

González Ollé, F. (1996-2016). *Navarro*. En M. Alvar López (Dir.), *Manual de dialectología hispánica el español de España* (pp. 305-316). Ariel.

Hernández Campoy, J. M. (1993). Dialectología Tradicional, Sociolingüística Laboviana y Geolingüística Trudgilliana: tres aproximaciones al estudio de la variación. *ELUA Estudios de lingüística Universidad de Alicante*, *9*, 151-181. http://hdl.handle.net/10045/6470

Hernández Campoy, J. M. (1999). La geolingüística: consideraciones sobre la dimensión espacial del lenguaje. *ELUA Estudios de lingüística Universidad de Alicante*, *13*, 65-88. http://dx.doi.org/10.14198/ELUA1999.13.03

Hernández Campoy, J. M. (2008). Principios básicos para el estudio geolingüístico de la variación. *Estudios Románicos, 17*(2), 515-528. https://revistas.um.es/estudiosromanicos/article/view/94981/91351

Hernández Lorenzo, L. (2019). Poesía áurea, Estilometría y fiabilidad: métodos supervisados de atribución de autoría atendiendo al tamaño de las muestras. *Caracteres: Estudios culturales y críticos de la esfera digital, 8*(1), 189-228. http://revistacaracteres.net/wp-content/uploads/2019/06/Caracteresvol8n1mayo2019-estilometria.pdf

Jannidis, F., Pielstrom, S., Schöch, C., & Vitt, T. (2015). Improving Burrows' Delta - an Empirical Evaluation of Text Distance Measures. Digital Humanities 2015: Conference Abstracts.

Juola, P. (2009). JGAAP: A System for Comparative Evaluation of Authorship Attribution. *JDHCS*, *1* (1). https://knowledge.uchicago.edu/record/117/files/4-173-1-PB.pdf?download=1

Julià Luna, C. (2020). *Geolingüística digital: proyecto de un corpus de atlas lingüísticos*. Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020. Universidad de Antioquía, Medellín, 21-23 octubre de 2020. 226-229. https://dlcc20.files.wordpress.com/2020/11/libro-de-resumenes-actas-iii-dlcc-2020-y-v-wopatec-2020-virtual.pdf

Martín Zorraquino, M. A., & Fort Cañellas, M. R. (1996-2016). La frontera catalano-aragonesa. En

M. Alvar López (Dir.). *Manual de dialectología hispánica el español de España* (pp. 293-304). Ariel.

Moreno Fernández, F. (1998-2009). *Principios de sociolinguística y sociología del lenguaje*. Ariel.

Moreno Fernández, F. (2003). Los estudios dialectales sobre el español en España (1979-2004). *Lingüística Española Actual*, *25*, 1-36. https://www.researchgate.net/publication/282737387_Los_estudios_dialectales_sobre_el_espanol_de_Espana_1979-2004

Muñoz Acebes, J. (2018). De la Glotocronología a la Filogenética: estado de la cuestión y los nuevos desarrollos de la metodología de clasificación lingüística. *Revista de Investigación Lingüística*, *21*, 170-184. https://orcid.org/0000-0002-0641-0727

Queralt Estévez, S. (2020). El uso de recursos tecnológicos en lingüística forense. *Pragmalinguistica*, *28*, 212-237. https://orcid.org/0000-0002-0641-0727

Smith P., & Aldridge W. (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, C. (2008). Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating. *Literary and Linguistic Computing*, *23*(2), 181-199. https://doi.org/10.1093/llc/fqm029

Zamora Vicente, A. (1960-1985). *Dialectología española*. Gredos.

ANEXO I: ANÁLISIS DE GRUPOS OFRECIDO POR STYLO EN EL SEGUNDO ANÁLISIS

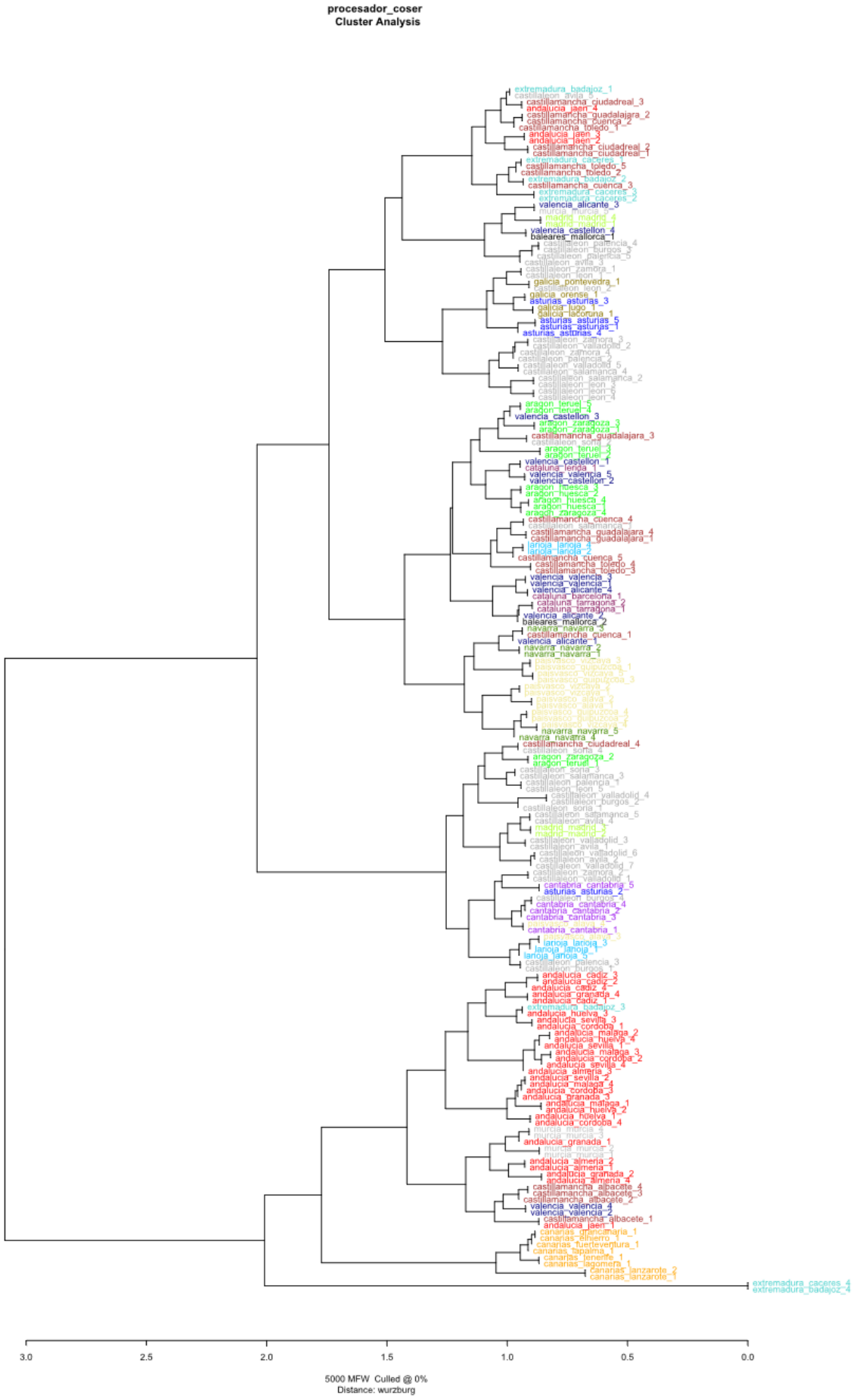


Figura A.1: Dendrograma del corpus completo. Fuente: elaboración propia.