ENTENDIENDO DELTA DESDE LAS HUMANIDADES

UNDERSTANDING DELTA FROM THE HUMANITIES

JOSÉ CALVO TELLO
UNIVERSIDAD DE WÜRZBURG

ARTÍCULO RECIBIDO: 16-03-2016 | ARTÍCULO ACEPTADO: 17-05-2016

RESUMEN:
La estilometría es una de las áreas de investigación en las Humanidades Digitales con mayor desarrollo. Sin embargo pocos estudios han trabajado hasta hace poco con textos en español y menos aún se han desarrollado en países hispanohablantes. El objetivo de este artículo es presentar en español y sin presuponer conocimientos estadísticos por parte del lector uno de los principales métodos utilizados en la estilometría: la medida de distancia textual de Burrows llamada Delta. El artículo explica este algoritmo usando un corpus mínimo de refranes y posteriormente comprueba los resultados en un corpus de novelas españolas. Tanto los datos como los archivos de programación Python están a disposición de la comunidad mediante GitHub, comentados paso por paso para que se pueda reproducir y visualizar cada paso.
.

ABSTRACT:
Stylometry is one of the research areas in greater development within Digital Humanities. However, few studies have worked until recently with texts in Spanish and even less so from Spanish-speaking countries. The aim of this paper is to present in Spanish, and without prior statistical knowledge from the reader, one of the main methods used in stylometry, the measure of textual distance Burrows' Delta. This paper explains this measure using a very small corpus of proverbs and then checks the results in a corpus of Spanish novels. Both data and Python scripts are available to the community through GitHub, commented step by step so that you can play and visualize each step.

Caracteres. Estudios culturales y críticos de la esfera digital
Vol. 5 (1), 2016: 140-176
ISSN: 2254-4496
http://revistacaracteres.net

PALABRAS CLAVE:
Estilometría, Delta, atribución de autoría, algoritmos, Python

KEYWORDS:
Stylometry, Delta, authorship attribution, algorythms, Python

---

José Calvo Tello. Está realizando su tesis doctoral dentro del grupo de investigación *Estilística computacional del género literario* (CliGS, por sus siglas en alemán) en la Universidad de Würzburg. En él investiga los subgéneros de novelas y cuentos españoles de la Edad de Plata mediante métodos cuantitativos como la estilometría. Además participa de otros proyectos de investigación y edición como la colección de eBooks *Clásicos Hispánicos*.

1. Introducción

Delta es uno de los métodos más aplicados e investigados sobre atribución autorial con textos literarios de las últimas décadas. La rama a la que pertenece, la estilometría, es una de las más sólidas y florecientes de las Humanidades Digitales. Como muestra de ello, la prestigiosa revista *Digital Scholarship in the Humanities*, publicada por Oxford University Press, indicó que en 2015 el principal tema de sus artículos fueron sobre estilometría y atribución autorial[1].

En este artículo veremos, en primer lugar, los hitos más importantes de Delta desde su nacimiento hasta la actualidad. El principal bloque de este artículo aplica Delta con un corpus mínimo de textos cortísimos, con numerosas visualizaciones de los textos y las tablas para que el lector entienda mejor paso por paso qué se están realizando. Delta es implementado en un pequeño programa en Python <https://github.com/morethanbooks/publications/tree/master/understanding_delta>, por lo que el lector puede emplearlo para reproducir con los mismos textos las pruebas por sí mismo. Tras este bloque, investigo si el mismo método para clasificar un corpus mínimo puede utilizarse para responder preguntas reales sobre estilometría y atribución de autoría en español. Para ello utilizo un corpus de novelas españolas entre finales del siglo XIX y comienzos del siglo XX. Finalmente concluiré con las principales ideas señaladas por este artículo.

[1] Cuenta de Twitter de *DSHjournal*, <https://twitter.com/DSHjournal/status/717983932021612544> (07-04-2016).

Mi idea de esta publicación nació durante los encuentros prácticos del grupo de investigación *Estilística computacional del género literario* (CliGS, por sus siglas en alemán), que está financiado por el Ministerio de Educación e Investigación alemán (BMBF), ubicado en la cátedra de Filología Computacional en la Universidad de Würzburg, dirigido por Christof Schöch, en las que también mis colegas Ulrike Henny y Daniel Schlör tomaron parte activa. El grupo comenzó su fase inicial en 2014 y su objetivo es investigar el género literario utilizando métodos computacionales en varias lenguas romances, como en el caso del francés en Schöch (2014). Para mayor información puede consultarse la página web del grupo <http://cligs.hypotheses.org/> o Calvo Tello *et al*. (2015).

Schöch, al que agradezco sus aclaraciones y orientación en general y sus comentarios a este artículo en particular, señaló la ventaja de implementar uno mismo los métodos principales con los que trabajamos para una mejor comprensión de ellos. Algunos de los artículos publicados explicando y proponiendo mejoras sobre Delta están escritos por investigadores que provienen de la informática, la estadística o la lingüística computacional, por lo que en sus publicaciones se presuponen una serie de conocimientos tanto de lingüística de corpus como de estadística. No solo eso, las investigaciones en este campo utilizando textos en español son escasas, y aún más escasas son aportaciones de este tipo de investigadores que trabajen en países hispanohablantes o que se publiquen en español. Durante la ejecución práctica y la lectura de los artículos percibí el esfuerzo que entraña para un humanista hispánico, en concreto un filólogo, entender y utilizar conceptos estadísticos en lenguas extranjeras como el inglés. El impulso final para este artículo fue dado por Burr (2015: web), quien en la clausura del *1st Day of DH*, en Madrid, subrayó la importancia de publicar también en lenguas romance.

2. Delta: origen, familia, infraestructura y uso

En esta sección del artículo abordo brevemente los principales hitos en el desarrollo de Delta y sus versiones, desde su propuesta hasta la actualidad. Esta sección debe servir como panorámica general y no tiene como objetivo ni ser exhaustivo ni aclarar en qué consiste cada detalle o cada modificación de este método. Precisamente el objetivo del resto de la publicación es aportar luz básicamente sobre el siguiente párrafo.

Delta fue propuesta por John Burrows en 2002 como "measure of stylistic difference" (267) que pudiese utilizarse en casos complejos de atribución de autoría. Burrows presenta resultados sorprendentemente exitosos de atribución de autoría con numerosos autores posibles. La idea en la que se basa Delta es que la variación de frecuencia de las palabras más frecuentes en un texto permite reconocer autoría. Su propia definición de Delta es: "the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text" (2002: 271).

Numerosos autores han señalado la falta de justificación teórica (p.ej.: Evert et al., 2015: 79) de Delta. Hoover fue uno de los primeros al decir que "Delta is a relatively simple measure of difference, but its calculation and interpretation are not very transparent" (2004b: 454). Siguiendo a Burrows y sus pruebas con textos poéticos en inglés, Hoover prueba en dos artículos de 2004 el mismo método, con tecnología similar (utilizando hojas de cálculo Excel) pero con un género literario diferente (novelas), de época más tardía y editando el texto de numerosas maneras; los resultados de Delta en su investigación resultan aún más sólidos.

Tras las propuestas de Hoover han seguido otra serie de propuestas y trabajos que han centrado su investigación principalmente (aunque no únicamente) en mejorar, controlar y entender mejor:

- la estandarización de las frecuencias de las palabras en los textos (Evert et al., 2015)
- uso de diferentes medidas de distancia[2] (Argamon, 2008), (Smith y Aldridge, 2011)
- la cantidad, tipos y calidad filológica de palabras que se deben asumir (Rybicki y Eder, 2011), (Eder, 2012), (Eder, 2013)

Tras los avances teóricos y de investigación se han desarrollado diferentes iniciativas de desarrollo de software para su implementación que permiten a otros investigadores trabajar en estilometría[3]. Entre ellas la herramienta más utilizada por investigadores es *stylo* (Eder, Kestemont y Rybicki, 2016), desarrollada en R desde 2013, gratuita, de código abierto, sencilla de utilizar, bien documentada y con una numerosa comunidad de usuarios. Más recientemente se han implementando en el lenguaje de programación Python otras iniciativas como *Pystyl* (Kestemont y Karsdorp, 2014) o *Pydelta* (Jannidis y Vitt, 2015). Estos

[2] Es decir, en qué grado deben aportar la frecuencia de cada palabra al conjunto del texto.

[3] Lamentablemente estas iniciativas no son sostenidas por presupuestos ni instituciones encargados de infraestructuras. Sorprende que una de las mayores ramas de las Humanidades Digitales esté siendo sostenida en último término por la generosidad, tiempo libre y paciencia de unos pocos investigadores que preparan, documentan y ponen a disposición de la comunidad su código.

proyectos utilizan librerías de Python como *pandas*, *NumPy* o S*ciPy*[4].

En cuanto a su uso con textos literarios en diferentes lenguas, la investigación en inglés ha sido la primordial. Recordemos que Burrows es angloparlante y anglista. Además, esa lengua cuenta con una larga tradición de codificación y publicación de textos literarios en formato electrónico. En relación a otras lenguas, los investigadores citados aquí hasta el momento han trabajado también con otras lenguas europeas como alemán, francés, polaco, holandés, latín o húngaro.

En el caso del español, las publicaciones que utilizan Delta son escasas[5], se han presentado en el contexto internacional ciertos trabajos sobre el *Quijote de Avellaneda* (Rißler-Pipka, 2016), la *Conquista de Jerusalén* de Cervantes (Calvo Tello y Cerezo Soler, 2016) o sobre literatura sapiencial medieval en varios idiomas, entre ellos el castellano (Wrisley, 2016).

Sobre las causas de su menor uso en el ámbito hispánico, creo que es útil subrayar que este método se ha utilizado poco tanto por investigadores hispanohablantes como por hispanistas de otros países. Lo primero puede ser explicado por la dificultad que representa acceder a herramientas, investigación y documentación en inglés. Pero el inglés no puede ser la causa por la que hispanistas y romanistas alrededor del mundo tampoco hayan trabajado con textos en español. En mi opinión una de las principales dificultades para trabajar en diferentes métodos de Humanidades Digitales con

[4] Librerías también utilizadas en el programa escrito para este artículo aunque sin la pericia de los investigadores citados ni un objetivo de investigación.
[5] No pretendo aquí recoger diferentes trabajos estilométricos usando textos en español en general, ya que queda fuera de los objetivos de este artículo, sino exclusivamente aquellos que han utilizado Delta.

textos en español es precisamente la falta de textos literarios disponibles en formatos aceptables. Un ejemplo paradigmático de esto, aunque no el único, es la *Biblioteca Virtual Miguel de Cervantes*, un proyecto que ha editado miles de textos en formato XML-TEI, lo que merece un enorme reconocimiento, pero que no solo no lo publica en ese formato, sino que tampoco lo facilita a investigadores que lo solicitan. Esto hace que cada investigador tenga que encargarse de conseguir un texto digno a partir de las obras troceadas en diferentes páginas HTML en formato hoy en día obsoleto. Por supuesto comenzar el trabajo desde ese HTML siempre es mejor que tener que comenzarlo desde el escaneo del libro.

3. Delta de un micro corpus de refranes

3.1. Metodología, datos e hipótesis

En esta parte del artículo quiero realizar un experimento con un doble objetivo: en primer lugar, esclarecer cada paso de los análisis que se realizan al utilizar Delta. En segundo lugar, quiero probar la hipótesis de si Delta ordena textos muy pequeños de una manera similar a como la introspección de un nativo lo haría.

Para este pequeño experimento he elegido una serie de refranes en español con la ayuda del portal *Refranario* (Calvo Tello, 2012). Para la elección de refranes se ha intentado elegir aquellos refranes que compartan alguna palabra entre sí. Además se ha intentado que el refrán mismo contenga alguna palabra repetida

para que los valores de frecuencia de cada una de las palabras por refrán no resulten idénticos en todos los casos[6].

Nuestro corpus (es decir, colección de textos) está compuesto por los siguientes refranes:

1. *lo hecho hecho está*
2. *lo pasado pasado está*
3. *a lo hecho pecho*
4. *si no lo veo no lo creo*
5. *una golondrina no hace verano*

Por supuesto los ejemplos son sencillos y están escogidos a propósito para que tengamos una idea intuitiva de las relaciones que debería tener. Esperamos encontrar los textos 1 y 2 juntos ya que comparten varias palabras; el 3 también relacionado con los dos primeros al compartir con ellos tanto *lo* como *hecho*; y los otros dos más o menos alejados de ellos. Utilizaré esta numeración durante el resto del artículo como identificadores (o ID) numéricos de cada refrán.

Estos refranes son considerados en este experimento como textos y durante el artículo utilizaré ambas palabras como sinónimos, aunque preferiré *textos* al ser más general. Recordemos que en los experimentos estándar de estilometría no se analizan refranes, sino que se utilizan textos largos como poemas, novelas, obras de teatro, etcétera.

[6] Es decir, el experimento no tiene como objetivo demostrar que Delta organice refranes de manera correcto según algún criterio paremiológico. Recordemos que el principal objetivo de este artículo es entender Delta.

El programa utilizado es una sencilla implemetanción propia de Delta en Python y durante el resto del artículo me referiré a él simplemente como el programa o los archivos de programación. Todos los archivos de programación y los corpus utilizados en este artículo están publicados en GitHub <https://github.com/morethanbooks/publications/tree/master/understanding_delta> y pueden utilizarse libremente. El pequeño programa cuenta en concreto de dos archivos:

- understanding_delta.py: archivo con las funciones básicas que en principio no debemos modificar
- understanding_delta_workflow.py: archivo que utiliza las funciones del anterior archivo y donde señalamos dónde se encuentran los textos con los que queremos trabajar

Ambos archivos están comentados de manera profusa para que se entienda en cada paso qué se está haciendo. Además hay numerosas líneas en understanding_delta.py que, al descomentarlas, el usuario podrá ver el estado de los datos exactamente en ese punto. Además durante este artículo iré haciendo menciones a las funciones para hacer explicita la relación entre teoría y práctica. La utilización de ambos archivos requiere tener instalado Python, las librerías que se importan y exige del usuario conocimiento básico de Python. Estos archivos están estrechamente relacionados con la *toolbox* de CliGS (Schöch *et al*., 2014).

El objetivo de esta implementación de Delta no es en ningún caso ofrecer una herramienta para trabajar realmente en estilometría, como las arriba mencionadas. Su objetivo es desglosarlo en pasos sencillos y apoyar al investigador que quiera entenderlos con ejemplos pequeños.

3.2. Frecuencia y longitud de palabras y corpus

Antes de comenzar con los pasos específicos de Delta, necesitamos conseguir algunos datos estadísticos básicos sobre los textos, en nuestro caso los refranes. Esta sección y la siguiente son realizadas por la función *countWordfrequencies*. Los textos han de ser divididos en palabras individuales[7] o *tokens*, proceso al que se le suele llamar *tokenizar*[8]. Se debe extraer la longitud total del refrán en palabras y cuántas veces aparece en ese refrán cada palabra.

ID	Texto	Longitud de texto en palabras	Palabra y frecuencia
1	lo hecho hecho está	4	'hecho': 2, 'lo': 1, 'está': 1
2	lo pasado pasado está	4	'pasado': 2, 'lo': 1, 'está': 1
3	a lo hecho pecho	4	'pecho': 1, 'hecho': 1, 'lo': 1, 'a': 1
4	si no lo veo no lo creo	7	'lo': 2, 'no': 2, 'creo': 1, 'si': 1, 'veo': 1
5	una golondrina no hace verano	5	'una': 1, 'verano': 1, 'hace': 1, 'golondrina': 1, 'no': 1

Tabla 1: Textos y datos básicos sobre frecuencia

[7] A partir de este punto introduciré el concepto *token*, palabra de la tradición de lingüística de corpus y que representa la idea intuitiva de que una palabra es una cadena de caracters alfabéticos rodeado de espacio o puntuación. Para evitar repeticiones, *palabra* y *token* se usarán en el resto del artículo como sinónimos.

[8] Aunque el proceso de tokenización puede parecer inocuo, en realidad modifica notablemente la cantidad y las características de las unidades con las que se trabajará posteriormente. La estilometría suele trabajar borrando puntuación, guarismos y pasando todo el texto a minúsculas. En la tokenización se toman decisiones además sobre qué se realizan en casos como *tic-tac* o *e-mail;* en otros idiomas como el inglés o el francés estas decisiones resultan más importantes y complejas que para el español, debido a las contracciones como *j'ai* o *I'm.*

Comparando los diferentes valores por texto, podemos extraer la lista completa de tokens que componen nuestro corpus. En nuestro caso nuestros cinco refranes se componen de un total de 14 palabras diferentes. Si a estos les añadimos la frecuencia de esa palabra en el corpus y ordenamos los datos según ese criterio, el resultado sería el siguiente[9]:

Token	**Frecuencia. en corpus**
lo	5
hecho	3
no	3
está	2
pasado	2
a	1
creo	1
golondrina	1
hace	1
pecho	1
si	1
una	1

[9] Nuestro programa asume como cantidad máxima 5 000 palabras por texto. Tanto en artículos de investigación como en la documentación de los programas se suele hablar de *rasgos* (en inglés *features*) en caso de que se trabaje con tokens (lo usual), por lo que es usual también el término *most frequent words* o *MFW*. Los primeros trabajos con Delta trabajaban con algunas centenas de palabras, y la cantidad ha ido aumentando progresivamente. Por supuesto nuestro corpus de refranes está muy alejado de esa cifra, pero posteriormente veremos un ejemplo donde esto tendrá importancia.

veo	1
verano	1

Tabla 2: Frecuencia total de tokens en el corpus

Delta, como otros métodos estilométricos, analiza los textos comparándolos con otros textos. Por lo tanto cualquier resultado debe ser interpretado dentro del corpus utilizado; si ingresásemos otro texto más o sustituyésemos uno por otro texto, los resultados podrían cambiar.

En la tabla anterior no se visualizaba la frecuencia de cada token en cada texto, un valor que necesitaremos para nuestro análisis. En la siguiente tabla cada columna representa uno de los tokens del corpus completo y cada fila cada uno de los textos. En caso de que ese token no aparezca en ese texto, se señala como 0. En caso contrario, se coloca la cantidad de veces que aparece en total en ese texto[10]:

Texto	**a**	**creo**	**está**	**golondrina**	**hace**	**hecho**	**lo**	**no**	**pasado**	**pecho**	**si**	**una**	**veo**	**verano**
1_lo hecho hecho está	0	0	1	0	0	2	1	0	0	0	0	0	0	0
2_lo pasado pasado está	0	0	1	0	0	0	1	0	2	0	0	0	0	0
3_a lo hecho pecho	1	0	0	0	0	1	1	0	0	1	0	0	0	0

[10] En esta tabla se puede observar mejor por qué era interesante tener refranes con palabras repetidas. De no ser así, la tabla solo tendría como valores 0 y 1, lo que podría ser confuso en cuanto al tipo de datos.

4_si no lo veo no lo creo	0	1	0	0	0	0	2	2	0	0	1	0	1	0
5_una golondrina no hace verano	0	0	0	1	1	0	0	1	0	0	0	1	0	1

Tabla 3: Frecuencia de cada token en cada texto

Sin embargo la longitud de los textos es variable, por lo que la frecuencia de los tokens debería ser relativa. Es decir, si tenemos un texto de 4 palabras y otros de 4000 y en ambos la palabra *golondrina* aparece una sola vez, aceptamos intuitivamente que en el primer texto la palabra *golondrina* es más importante que en el segundo. Por eso dividimos cada uno de los valores de la frecuencia de cada palabra por texto entre la longitud del texto (valor obtenido en la tabla 1):

Texto	**a**	**creo**	**está**	**golondrina**	**hace**	**hecho**	**lo**	**no**	**pasado**	**pecho**	**si**	**una**	**veo**	**verano**
1_lo hecho hecho está	0	0	0.25	0	0	0.5	0.25	0	0	0	0	0	0	0
2_lo pasado pasado está	0	0	0.25	0	0	0	0.25	0	0.5	0	0	0	0	0
3_a lo hecho pecho	0.25	0	0	0	0	0.25	0.25	0	0	0.25	0	0	0	0
4_si no lo veo no lo creo	0	0.142857	0	0	0	0	0.285714	0.285714	0	0	0.142857	0	0.142857	0
5_una golondrina no hace verano	0	0	0	0.2	0.2	0	0	0.2	0	0	0	0.2	0	0.2

Tabla 4: Frecuencia relativa de cada token en cada texto

Fijémonos en la palabra *hecho*. Esta palabra representa el 50% del texto 1, mientras que representa el 25% del texto 3. Otro ejemplo interesante es *lo*: aunque aparece 2 veces en total en el texto 4, como ese texto es más largo que el resto, la frecuencia relativa de *lo* en todos es bastante similar.

3.3. Palabras como dimensiones

Ya hemos convertido los textos en números, por lo que ya podemos empezar a comparar los diferentes textos por sus resultados numéricos. Por ejemplo, en la siguiente gráfica observamos que el texto 1 y 2 (azul y rojo) comparten exactamente el mismo valor en *está*; también en *lo*; y además son los únicos textos que consigue adquirir un valor de 0.5 (en *hecho* y *pasado* respectivamente)[11]. Es decir, la similitud que veíamos en el el texto del refrán, también lo observamos en los valores de frecuencia relativa.

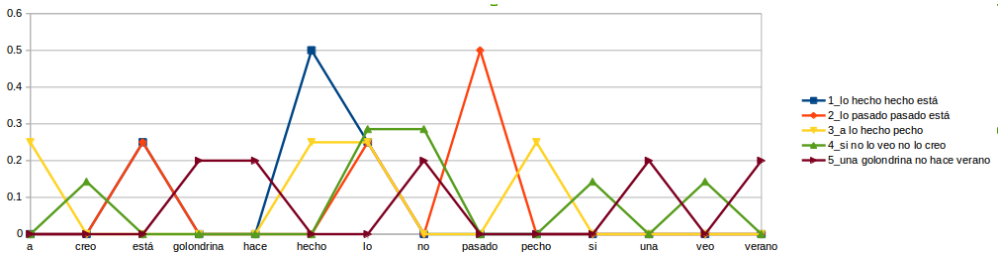

Ilustración 1: Visualización de valores relativos de frecuencia, siendo el eje horizontal cada una de las palabras

Pensemos en otra posible visualización como el típico eje cartesiano con un eje horizontal y un eje vertical al que le podemos añadir una dimensión más:

[11] Aunque este último aspecto no será tenido en cuenta por Delta.

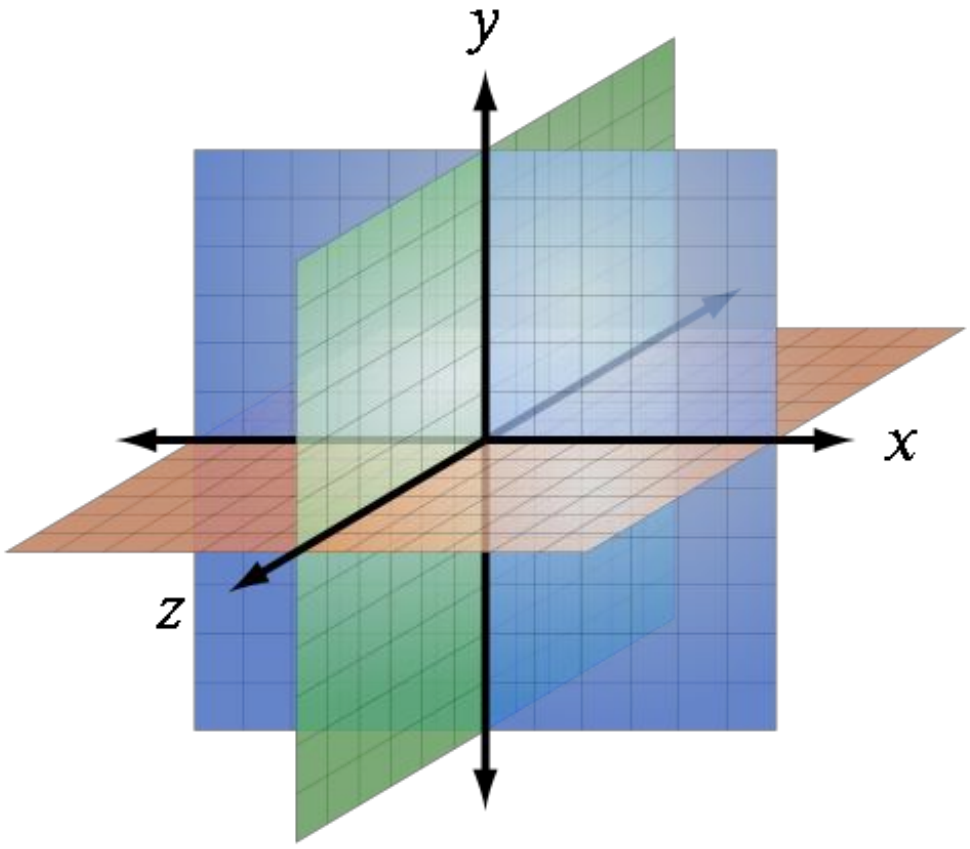

Ilustración 2: ejes cartesianos. *Wikimedia*, <https://commons.wikimedia.org/wiki/File:Ejes_cartesianos_7.jpg> (05-04-2012)

Imaginemos que queremos proyectar un texto sobre estos ejes. Cada palabra es una dimensión y su frecuencia relativa dentro del texto es su valor. Digamos por ejemplo que queremos representar de esta manera el texto *lo hecho hecho está*, que tiene 3 palabras diferentes. A cada una le adjudicamos una dimensión. Por ejemplo *lo* es el eje horizontal (eje *x*) y tiene un valor de 0.25; *está* es el eje vertical o eje *y,* y tiene un valor de 0.25; y *hecho* en el tercer eje,

eje *z*, y tiene un valor de 0.5. Podemos representarlo de la siguiente manera:

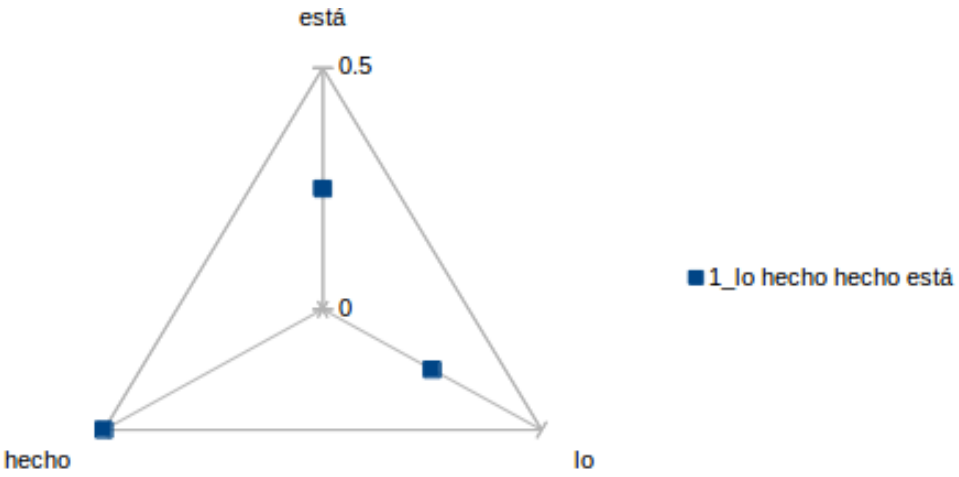

Ilustración 3: Representación de lo hecho hecho está como tres dimensiones

En este punto puede entenderse mejor la ventaja de trabajar con textos tan cortos como los refranes y en concreto con refranes que solo tienen tres palabras diferentes ya que nuestras limitaciones humanas circunscriben algunas de nuestras capacidades a tres dimensiones. Visualizaciones de cuatro dimensiones nos permitirían mostrar las cuatro dimensiones de las palabras de *a lo hecho pecho*; con visualizaciones de 5 dimensiones permitirían mostrar *si no lo veo no lo creo*. Incluso podríamos proyectar un texto de miles de palabras y visualizarlo como miles de dimensiones. El hecho de que no podamos ni visualizarlo ni imaginarlo no significa que no se pueda calcular y que podamos utilizar esos valores para compararlos, ya que los números son más fácilmente comparables que las palabras.

Imaginemos que cada una de las 14 palabras de nuestro corpus de refranes es una dimensión y que cada texto tiene un valor (su frecuencia relativa) en cada dimensión. *Lo hecho hecho está* y *una*

golondrina no hace verano no comparten ninguna palabra. Eso quiere decir que en la dimensiones donde *lo hecho hecho está* tiene valores positivos, *una golondrina no hace verano* siempre tiene 0, y viceversa. Pero algunos de los textos de nuestro corpus sí que comparten valores positivos en algunas de las dimensiones; por ejemplo, cuatro de los 5 textos tienen la palabra *lo*, por lo que podemos comparar los valores de esos textos en esa dimensión. Por supuesto no solo debemos tener en cuenta si ambos tienen valores positivos o no, sino si esos valores tienen valores similares. Y al tener muchas dimensiones, seguimos sin conseguir realmente comparar textos.

Si modificamos los ejes de la visualización 1 y pasamos los textos al eje horizontal, observaremos lo siguiente:

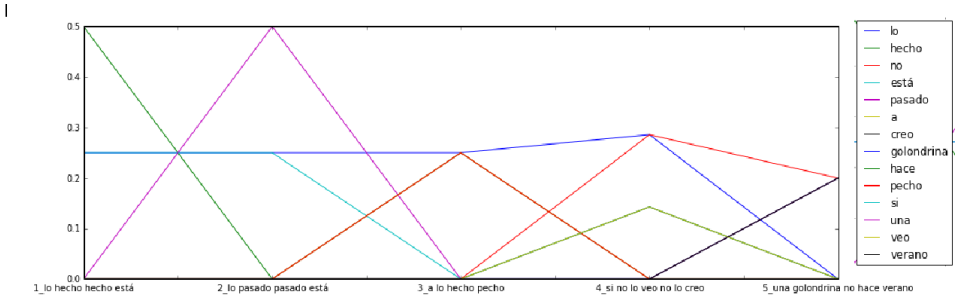

Ilustración 4: Visualización de valores relativos de frecuencia, siendo el eje horizontal cada uno de los textos

Como vemos, los valores de algunos tokens varían menos: por ejemplo *lo* (una de las líneas azules) cuyo valor mínimo es 0 y cuyo valor máximo es 0.28; sin embargo hay otras palabras, como *hecho*, (línea verde), cuya variedad pasa de 0.5 en el texto 1, a 0 en el texto 2.

Esto también ocurre cuando comparamos textos largos entre ellos, aunque no con porcentajes tan altos. Un nombre como *Juan* puede representar un porcentaje relativamente alto en una novela

cuyo protagonista se llama así, mientras que en otra novela puede representar un 0% del texto al no haber nadie que se llame así.

Frente a esto, tenemos el caso de *lo*. En los textos 1, 2 y 3 tiene un valor de 0.25; en el texto 4 tiene un valor de 0.28. Esas 3 centésimas pueden parecer poco. Pensemos en un grupo de unos 20 textos: 19 tienen una frecuencia relativa entre 0.21 y 0.22 para una palabra concreta; para la misma palabra, el texto que queda tiene una frecuencia relativa de 0.25. Aunque pueda parecer pequeña la diferencia, puede contener más significado del que cabría esperarse.

3.4. Estandarización mediante z-scores o democratización léxica

Es decir, aunque las frecuencias relativas son menos problemáticas que las frecuencias totales, seguimos teniendo el problema de que algunas palabras siguen mostrando una gran variación y que por lo tanto tienden a diferenciar más los textos que no aquellas diferencias sutiles de palabras muy frecuentes. Como señala Burrows "the object is to treat all of these words as markers of potentially equal power in highlighting the differences between one style and another" (2002: 171). La solución propuesta por Delta es utilizar los *z-scores*[12]. La idea detrás de esto es: ¿cómo de diferente es la frecuencia de esta palabra en comparación con la frecuencia media en el corpus? ¿Y en qué medida afectan a estos valores la desviación típica de los valores en el corpus. En el

[12] Recientemente Evert et al. (2016) han señalado que los z-scores no consiguen que cada palabra aporte al texto de manera igual; en su lugar han propuesto otra manera de estandarizar la frecuencia léxica mediante valores ternarios cuyos resultados cuyos resultados no solo son mejores sino también más robustos antes diferentes cantidades de palabras.

programa que acompaña este artículo, esto es realizado por la función *getZscore*.

Para ello en primer lugar, necesitamos calcular la media de la frecuencia relativa de cada token en el corpus. En comparación con la tabla anterior, esta tabla solo aporta las dos últimas filas, además de cambiar la ordenación de los tokens a la que utiliza el programa:

	lo	**he-cho**	**no**	**está**	**pasa-do**	**a**	**creo**	**golon-drina**	**hace**	**pecho**	**si**	**una**	**veo**	**verano**
1_lo hecho hecho está	0.25	0.5	0	0.25	0	0	0	0	0	0	0	0	0	0
2_lo pasado pasado está	0.25	0	0	0.25	0.5	0	0	0	0	0	0	0	0	0
3_a lo hecho pecho	0.25	0.25	0	0	0	0.25	0	0	0	0.25	0	0	0	0
4_si no lo veo no lo creo	0.285714	0	0.285714	0	0	0	0.142857	0	0	0	0.142857	0	0.142857	0
5_una golon-drina no hace verano	0	0	0.2	0	0	0	0	0.2	0.2	0	0	0.2	0	0.2
Suma de las frecuen-cias relativas	1.035714	0.75	0.485714	0.5	0.5	0.25	0.142857	0.2	0.2	0.25	0.142857	0.2	0.142857	0.2
Media	0.207143	0.15	0.097143	0.1	0.1	0.05	0.028571	0.04	0.04	0.05	0.028571	0.04	0.028571	0.04

Tabla 5: Frecuencia relativa de cada token en cada texto

Si nos fijamos en los valores de *está*, había dos textos que tenían una frecuencia relativa de 0.25. Su suma es de 0.5, dividido entre la cantidad total de textos, cinco, el resultado es 0.1.

Además necesitamos en primer lugar la media de la frecuencia relativa de la palabra en el corpus y la desviación típica de cada valor. Una vez tenemos ambos valores, se aplica la siguiente fórmula, extraída de la definición de Delta que señalábamos como aportación de Burrows:

z-score = (frecuencia relativa - media de la palabra en el corpus) / desviación típica

De esta manera obtenemos los z-scores por cada palabra en cada texto. Visualicemos los valores en concreto:

	lo	hecho	no	está	pasado	a	creo	golondrina	hace	pecho	si	una	veo	verano
1_lo hecho hecho está	0.366851	1.565248	-0.712052	1.095445	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214
2_lo pasado pasado está	0.366851	-0.670820	-0.712052	1.095445	1.788854	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214	-0.447214
3_a lo hecho pecho	0.366851	0.447214	-0.712052	-0.730297	-0.447214	1.788854	-0.447214	-0.447214	-0.447214	1.788854	-0.447214	-0.447214	-0.447214	-0.447214
4_si no lo veo no lo creo	0.672560	-0.670820	1.382218	-0.730297	-0.447214	-0.447214	1.788854	-0.447214	-0.447214	-0.447214	1.788854	-0.447214	1.788854	-0.447214
5_una golondrina no hace verano	-1.773112	-0.670820	0.753937	-0.730297	-0.447214	-0.447214	-0.447214	1.788854	1.788854	-0.447214	-0.447214	1.788854	-0.447214	1.788854

Tabla 6: z-scores de cada palabra en cada texto

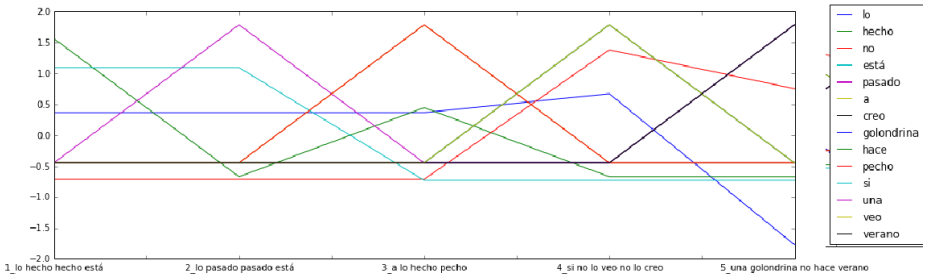

Ilustración 5: visualización de z-scores de cada palabra en cada texto

Veamos de nuevo el caso concreto de *lo*. En los tres textos que contenían este token una sola vez, el z-score es de 0.366851 y ese valor prácticamente se dobla (0.672560) en el texto 4, que contenía dos veces la palabra. Hasta ahora estos valores representan de manera similar los valores de frecuencia ya vistos. Pero si nos fijamos en texto 5, el único que no tenía esa palabra, veremos que su valor queda muy por debajo que el resto: -1.773112. De esta manera el z-score recoge lo anormal que resulta dentro de este corpus que un texto no tenga la palabra *lo* cuando todos los otros textos sí la tenían.

Si observamos el resto de tokens y sus valores, todos están entre 2 y -2. Para entender la ventaja de trabajar con z-scores y no

con frecuencias relativas, comparemos ambos valores en los casos de *lo* y *hecho*:

- El valor más alto de lo está ligeramente por encima del 0.5 y el valor más bajo que tiene es algo inferior a -1.5. Es decir, la diferencia entre sus valores máximo y mínimo es aproximadamente de 2.
- El valor más alto de *hecho* está ligeramente por encima de 1.5, y su valor más bajo ronda el -0.5. Como vemos, la diferencia entre sus valores máximo y mínimos también es de aproximadamente 2.

Comparemos la diferencia entre valores máximos y mínimos tanto de la frecuencia relativa como de los z-scores en la siguiente tabla:

Medidas	*lo*	*hecho*
Frecuencia relativa máxima	0.285714	0.5
Frecuencia relativa mínima	0	0
Diferencia de frecuencia relativa	**0.285714**	**0.5**
Z-score máximo	0.672560	1.565248
Z-score mínimo	-1.773112	-0.670820
Diferencia de z-score	**2.445672**	**2.236068**

Tabla 7: comparación de valores de frecuencia relativa y z-scores de dos tokens

Como vemos, los valores de las diferencias de z-score (2.445672 y 2.236068) en ambas palabras son muy similares. Sin embargo los valores de la diferencia de frecuencia relativa (0.285714 y 0.5) de ambas palabras palabras prácticamente se doblan. Es decir, hemos reducido notablemente la variación entre los posibles resultados pasando de frecuencia relativa a z-scores. Esto es un aspecto fundamental de Delta: dar una importancia similar a todas las palabras.

3.5. Calculando la distancia de dimensiones para Delta

Una vez tenemos los z-scores de cada token en este corpus, podemos finalmente utilizar Delta para comparar los textos. Esta sección es realizada por la función *delta* en los archivos de programación. La idea a seguir ahora es:

Paso 1: Cojo un primer texto

Paso 2: Cojo un segundo texto

Paso 3.1: Cojo una palabra

Paso 4.1: Recojo el valor de esta palabra para el primer texto y recojo el valor de esta palabra para el segundo texto

Paso 4.2: Resto sus valores absolutos (usando la medida de distancia Manhattan)[13]

Paso 4.3: Voy sumando lo obtenido en 4.2 a la distancia entre estos textos

Vuelta al Paso 3.1, utilizando una nueva palabra

Paso 3.2: Cuando termino las palabras, vuelvo al paso 1 con el siguiente texto

Una vez el programa ha realizado todos los pasos, obtenemos la siguiente tabla, llamada matriz Delta, con los *delta scores* para cada pareja de textos:

[13] Esta medida de distancia fue la que Burrows implementó en su versión de Delta. En Argamon (2008) propuso utilizar en su lugar la medida de distancia euclidiana y en recientes trabajos se ha utilizado el coseno (Smith y Aldridge, 2011), que hasta ahora se muestra como la más robusta. Para una comparación de las diferentes propuestas puede consultarse Evert et al. (2015).

	1_lo hecho hecho está	2_lo pasado pasado está	3_a lo hecho pecho	4_si no lo veo no lo creo	5_una golondrina no hace verano
1_lo hecho hecho está	0	4.472136	7.415912	13.16999	16.61203
2_lo pasado pasado está	4.472136	0	9.65198	13.16999	16.61203
3_a lo hecho pecho	7.415912	9.65198	0	14.69835	18.14039
4_si no lo veo no lo creo	13.16999	13.16999	14.69835	0	18.72643
5_una golondrina no hace verano	16.61203	16.61203	18.14039	18.72643	0

Tabla 8: Matriz Delta de nuestro corpus

Asegurémonos que entendemos esta tabla, que representa la distancia entre los diferentes textos. Se puede interpretar de manera similar a las tablas de distancias entre ciudades que aparecen en los atlas:

Distancias en kilómetros	*Cdad. de Bs. As.*	*Córdoba*	*Corrientes*	*Formosa*	*La Plata*
Cdad. de Bs. As.		646	792	933	53
Córdoba	646		677	824	698
Corrientes	792	677		157	830
Formosa	933	824	157		968
La Plata	53	698	830	968	

Ilustración 6: tabla de distancia entre ciudades argentinas

De una manera similar al hecho de que la distancia entre Córdoba y la Plata es de 698 (kilómetros), la distancia entre el texto 1 y el 4 es de 16.61203. Por supuesto, la distancia entre Córdoba y Córdoba es 0. Los atlas tienden a no poner ese valor, sino que dejan la casilla en gris. En nuestra tabla sí que aparece el 0, en diagonal a lo largo de la tabla.

Podemos visualizar la matriz Delta como *heatmap* para que su interpretación resulte más intuitiva:

	1_lo hecho hecho está	2_lo pasado pasado está	3_a lo hecho pecho	4_si no lo veo no lo creo	5_una golondrina no hace verano
1_lo hecho hecho está	0	4.472136	7.415912	13.16999	16.61203
2_lo pasado pasado está	4.472136	0	9.65198	13.16999	16.61203
3_a lo hecho pecho	7.415912	9.65198	0	14.69835	18.14039
4_si no lo veo no lo creo	13.16999	13.16999	14.69835	0	18.72643
5_una golondrina no hace verano	16.61203	16.61203	18.14039	18.72643	0

Tabla 9: Matriz Delta de nuestro corpus como *heatmap*

En este caso cuanto más fuerte sea el rojo, mayor es la distancia. Como vemos los textos 1 y 2 tienen un color pálido y es que efectivamente los textos *lo hecho hecho está* y *lo pasado pasado está* son muy similares. De hecho estos dos son los textos más cercanos de nuestro corpus. Frente a ellos están los textos 5 y 4, tienen la mayor distancia entre ellos.

3.6. Suma de matrices

Antes de continuar adelante, volvamos un paso atrás en el cálculo de la matriz Delta, exactamente al paso 4.3. del *pseudo-código* de la anterior sección, antes de que sumemos los valores de las dimensiones por cada texto. El objetivo de esta vuelta es, ahora que sabemos interpretar una matriz Delta, entendamos mejor cómo cada dimensión está afectando el cálculo final. En el programa que acompaña este artículo las siguientes tablas pueden visualizarse mediante la función *delta_word*, aunque no es necesario realizarlo. Por ejemplo, vamos a ver algunas tablas de distancias Delta para las palabras *lo* y *hecho*:

```
In [36]: delta_word_matrix # Tabla Delta para la palabra "lo"
Out[36]:
```

	1_lo hecho hecho está	2_lo pasado pasado está	3_a lo hecho pecho	4_si no lo veo no lo creo	5_una golondrina no hace verano
1_lo hecho hecho está	0	0	0	0.3057089	2.139962
2_lo pasado pasado está	0	0	0	0.3057089	2.139962
3_a lo hecho pecho	0	0	0	0.3057089	2.139962
4_si no lo veo no lo creo	0.3057089	0.3057089	0.3057089	0	2.445671
5_una golondrina no hace verano	2.139962	2.139962	2.139962	2.445671	0

```
In [41]: delta_word_matrix # Tabla Delta para la palabra "hecho"
Out[41]:
```

	1_lo hecho hecho está	2_lo pasado pasado está	3_a lo hecho pecho	4_si no lo veo no lo creo	5_una golondrina no hace verano
1_lo hecho hecho está	0	2.236068	1.118034	2.236068	2.236068
2_lo pasado pasado está	2.236068	0	1.118034	0	0
3_a lo hecho pecho	1.118034	1.118034	0	1.118034	1.118034
4_si no lo veo no lo creo	2.236068	0	1.118034	0	0
5_una golondrina no hace verano	2.236068	0	1.118034	0	0

Tabla 10. Matriz Delta para las palabras *lo* y *hecho*

Como vemos la distancia entre el texto 1 y el texto 2 en *lo* es de 0, palabra que ambos textos contienen de la misma forma. Si nos fijamos, el texto 4 y el texto 5 tienen también una distancia de 0 en *hecho*, palabra que ninguno de los textos contiene. Es decir, Delta no distingue que una palabra esté en dos textos en la misma manera

o que esa palabra no esté en los dos textos; a los dos casos le asigna el mismo valor. Como Burrows mismo señaló "an expression of difference, pure difference, is what we seek" (2002: 269).

3.7. Representación jerárquica arbórea de los resultados

Aunque el principal resultado de Delta es la matriz arriba mostrada, en las Humanidades Digitales es más frecuente mostrar e interpretar estos datos jerarquizados como estructuras arbóreas o dendogramas, principalmente si el objetivo de la investigación es atribución de autoría. Para ello se utilizan métodos de aprendizaje automático no supervisado (*clustering*) que eligen qué valores son más similares entre ellos y que establezca una jerarquía. Hay numerosos métodos diferentes, y uno de los más utilizados en la estilometría es Ward[14]. Esta sección es implementada por las funciones *create_dendogram* y *visualize_dendogram*. Al utilizarlo, el resultado es:

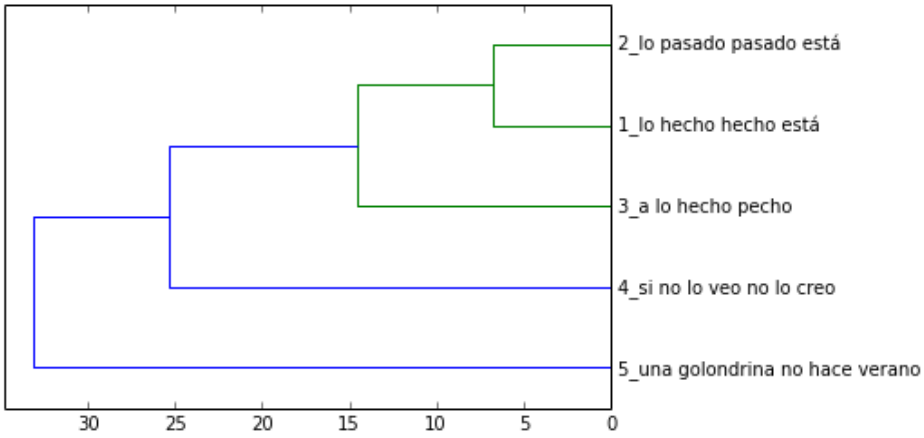

Ilustración 7: dendograma de la matriz Delta de nuestro corpus de refranes

[14] Por defecto en *stylo*.

En esta visualización, la relación entre los textos o los grupos de textos se crea mediante los nodos de ramas. El eje horizontal representa la distancia, es decir, la relación, de los textos. El tamaño del eje vertical no tiene más significado que permitirnos ver la estructura correctamente; su orientación es aleatoria, es decir, podríamos estar viendo que el refrán 1 quedase el primero y el texto 2 el segundo, pero su relación permanecería.

De esta manera este dendograma se interpreta como que los textos 1 y 2 cuelgan de una misma rama y esta es la más cercana a los textos, por lo que ambos textos son los más similares (algo ya visto en la matriz Delta). El nodo que une ambos textos se une con el texto 3, por lo que este refrán sería el siguiente más similar. El siguiente texto más similar sería el 4, mientras que el 5 sería el texto más diferente del grupo[15].

Si recordamos lo señalado en la presentación de este experimento, dijimos que "esperamos encontrar los textos 1 y 2 juntos ya que comparten varias palabras; el 3 también relacionado con los dos primeros al compartir con ellos tanto *lo* como *hecho*; y los otros dos más o menos alejados de ellos". Es decir, el método ha reproducido nuestra instrospección como hablantes.

4. Delta con un corpus de novelas españolas

En esta sección del artículo quiero reproducir el experimento con un corpus similar a los utilizados en trabajos estilométricos y

[15] Las diferentes opciones de visualización permiten utilizar colores tanto a los nombres de los archivos de texto (como hace *stylo*) como a las ramas (como ocurre en este caso). En este caso, el cambio de color de las ramas diferencia aquellas ramas cuya distancia es menor y por lo tanto su relación puede ser significativa a las que no.

comentar por qué el archivo Python aquí presentado aporta resultados pobres. Para ello voy a utilizar un corpus de 24 novelas (Calvo Tello, 2015) publicado bajo licencia Creative Commons en XML-TEI y texto plano como parte de mi actividad en el grupo de investigación CliGS de la Universidad de Würzburg (Schöch *et al*., 2015). Este corpus contiene tres obras de siete autores: Galdós, Bazán, Blasco Ibáñez, Valera, Pereda, Picón y Clarín; todos los textos fueron publicados desde finales del siglo XIX hasta comienzos del siglo XX y está específicamente modelado para realizar experimentos estilométricos de autoría en español.

El resultado del archivo Python para este corpus es el siguiente:

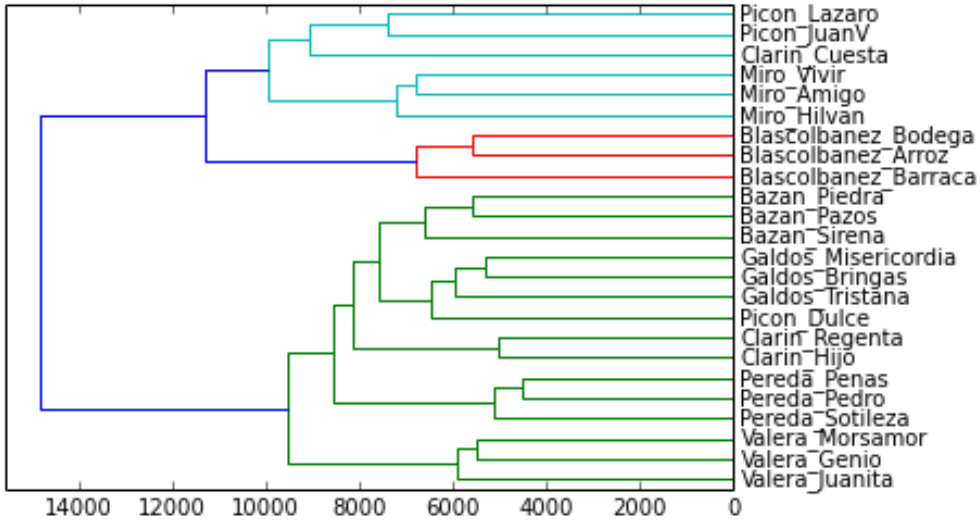

Ilustración 8: dendograma de 24 novelas españolas realizado con el archivo Python

Como se puede observar, 22 de los 24 textos han sido organizados correctamente por autoría. Los únicos dos casos

erróneos son *Cuesta*[16] de Clarín y *Dulce*, de Picón. Si se observa el valor del nodo que une las obras de Picón y *Cuesta* de Clarín, su posición en el eje horizontal es superior (más a la izquierda) que el resto de nodos que señalan autoría, por lo que cabría dudar de su valor para señalar autoría. No es así en el caso de Dulce de Picón y los textos de Galdós, aunque en este caso sí que se observa que los textos de Galdós mantienen una relación más estrecha, al que se le añade el texto de Picón. Como explicación de la errónea ordenación de estos dos casos, se puede señalar que Cuesta es el único de los tres textos de Clarín escrito en primera persona, mientras que en el caso de Picón la novela Dulce es la más tardía de las tres y es la única de las tres de contenido erótico (las otras dos se ordenan en los movimientos realista y naturalista).

Para confirmar los resultados, repitamos el experimento con los mismos textos y los mismos parámetros en un software utilizado realmente por la comunidad: stylo. Para ello utilizamos los parámetros: 5000 tokens (o MFW), Classic Delta (la versión propuesta por Burrows), diferenciación entre mayúscula y minúscula. El resultado es el siguiente:

[16] Para que la relación entre las ilustraciones y las explicaciones sea más clara, citaré los textos mediante nombres abreviados. Es usual utilizar versiones abreviadas tanto de los nombres de los autores como de los textos escritos sin acentos, para evitar problemas de procesamiento de caracteres y visualización.

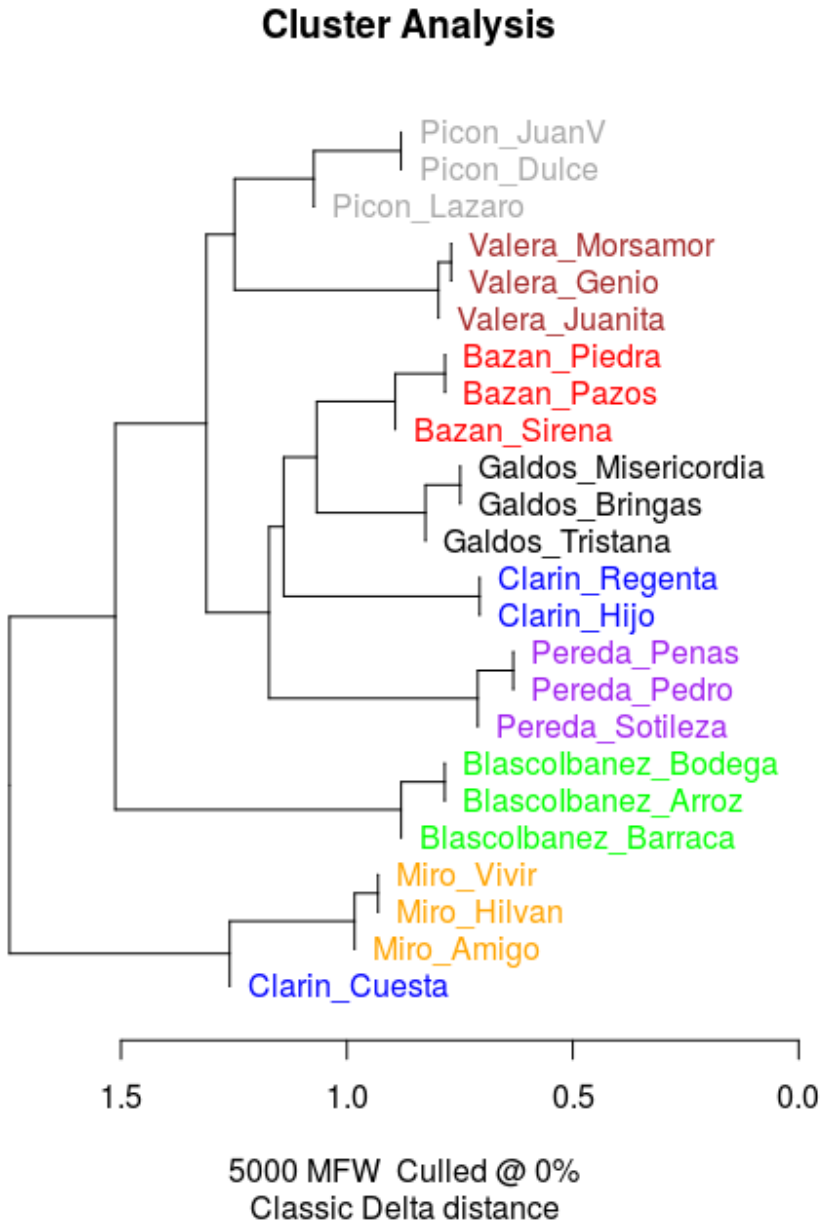

Ilustración 9: dendograma del corpus de novelas, realizado en *stylo* con la Delta de Burrows

En este caso 23 de los 24 textos han sido organiados correctamente por autor. Los textos de Picón han sido organizados correctamente, mientras que *Cuesta* de Clarín sigue siendo ordenado con otros textos, en concreto con los de Miró (cuyos

textos comparten la característica de no ser autodiegéticos, es decir, que como en *Cuesta* el narrador está en primera persona). La diferencia de los resultados entre los dos dendogramas debe ser a causa de algún paso concreto en Delta, probablemente en el proceso de tokenización, del que ya había comentado que no era inocuo.

Aunque la calidad de los resultados es notable, estamos usando hasta ahora la versión Delta que Burrows propuso hace ya más de una década. Realicemos de nuevo el experimento utilizando la versión de Delta más actualizada entre las opciones de *stylo*, utilizando los mismos parámetros anteriores pero utilizando la versión Delta de Eder:

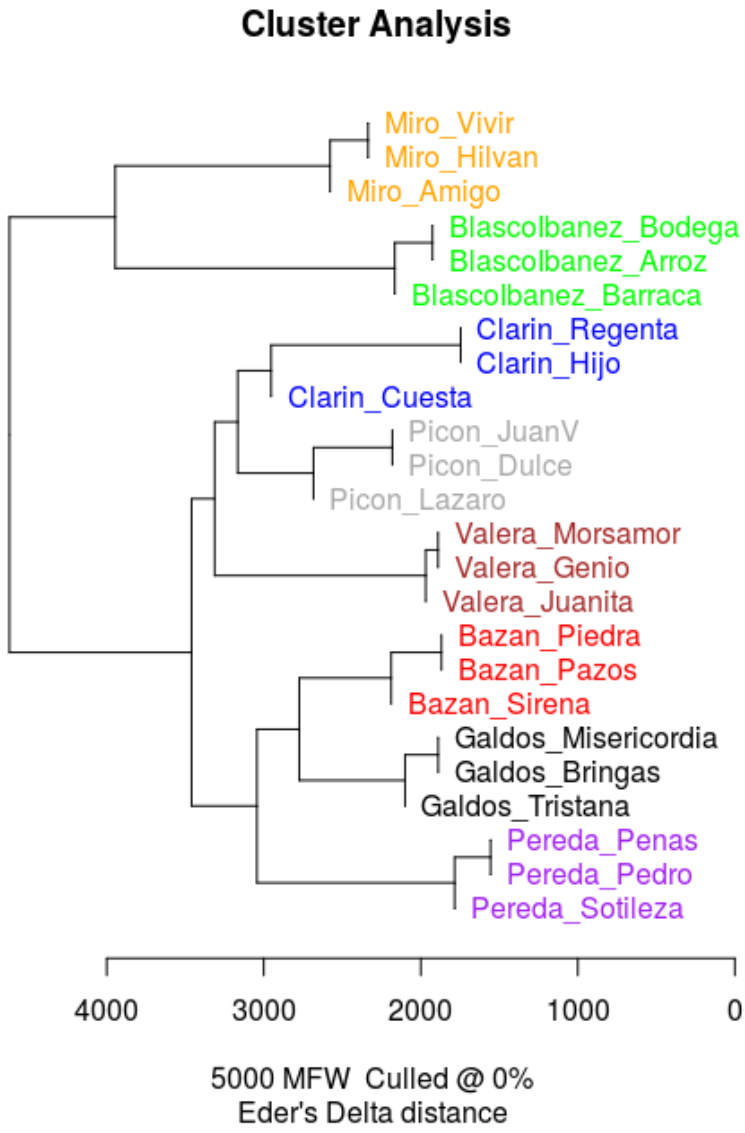

Ilustración 10: dendograma del corpus de novelas, realizado en *stylo* con la Delta de Eder

En este caso el texto de Clarín ha sido organizado también correctamente, por lo que esta versión de Delta con estos parámetros ha sido capaz de organizar correctemante todos los textos.

5. Conclusiones

Tras resumir los principales hitos en la historia, desarrollo, implementación y uso en español de Delta, este trabajo ha presentado los datos y el software, accesible en Internet, para realizar varios experimentos estilométricos con textos en español. El primero, realizado con un corpus mínimo de refranes, nos ha permitido entender mejor qué pasos ejecuta Delta. En concreto:

1. Los textos son tokenizados
2. Se extraen diferentes datos de frecuencia de los tokens en los textos del corpus
3. La frecuencia se estandariza utilizando z-scores
4. Se comparan los valores de cada dimensión por cada pareja de textos, sumándose sus valores para obtener los *delta-scores*
5. Estos valores de distancias se agrupan (*cluster*) jerárquicamente para poder realizar una representación arbórea

El uso de Delta con los refranes reproducía nuestra introspección de las relaciones que debía haber. Al repetir el experimento con un corpus de novelas españolas, casi todos los textos eran organizados correctamente por autoría. Al utilizar una implementación y una versión de Delta que representan el estado de la investigación actual, todos los textos eran ordenados correctamente por autoría.

Confío que este trabajo haya mostrado la capacidad de Delta de detectar similitud entre texto y justifique su uso para estudios textuales y de atribución de autoría, también con textos en español y que impulse su difusión, comprensión y aplicación. Quedo, junto

a mi grupo de investigación CliGS, a disposición de otros grupos de investigación o investigadores que estén interesados en colaboración o profundización en el tema.

6. Bibliografía

Argamon, Shlomo (2008). "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations". *Literary and Linguistic Computing* 23 (2): pp. 131-47.

Burr, Elisabeth (2015). "What can the Digital Humanities offer small linguistic and cultural communities?". *Day of DH*, Madrid. <http://e-spacio.uned.es/_congresosuned/index.php/eadh/EADHDay/paper/view/182/1> (20-03-2016)

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

Calvo Tello, José, y Juan Cerezo Soler (2016, en prensa). "La conquista de Jerusalén ¿de Cervantes? Análisis estilométrico sobre autoría en el teatro del Siglo de Oro español". *Digital Humanities Quarterly* 10.

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

Eder, Maciej (2012). "Mind Your Corpus: Systematic Errors in Authorship Attribution". En *Digital Humanities 2012: Conference Abstracts*. Hamburg: Hamburg Univ. Press. <https://sites.google.com/site/computationalstylistics/preprints/m-eder_mind_your_corpus.pdf> (20-03-2016).

Eder, Maciej (2013). "Does Size Matter? Authorship Attribution, Small Samples, Big Problem". *Digital Scholarship in the Humanities* 30 (2): pp. 167-82.

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

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Christof Schöch y Thorsten Vitt (2015). "Towards a Better Understanding of Burrows's Delta in Literary Authorship Attribution". En *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver CO: Association for Computational Linguistics. pp. 79-88.

Rißler-Pipka, Nanete (2016). "Der falsche Quijote? Autorschaftsattribution für spanische Prosa der frühen Neuzeit". En *DHd 2016 Modellierung, Vernetzung, Visualisierung.* Leipzig: DHd. pp. 212-217. <http://dhd2016.de/boa.pdf>(20-03-2016)

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

Schöch, Christof (2014). "Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik". En *Literaturwissenschaft im digitalen Medienwandel 7:* pp. 130-157. <http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf> (20-03-2016)

Smith, Peter W. H., y W. Aldridge (2011). "Improving Authorship Attribution: Optimizing Burrows' Delta Method". *Journal of Quantitative Linguistics* 18 (1): pp. 63-88.

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

Recursos

Biblioteca Cervantes Virtual. Alicante: Universidad de Alicante. <www.cervantesvirtual.com> (1999)

Calvo Tello, José (coord.), *Refranario*, Madrid: Molino de Ideas. <www.refranario.com> (2012)

Calvo Tello, José. *Corpus of Spanish Novel from 1880-1940*. Sp. Würzburg: University of Würzburg. <https://github.com/cligs/textbox/tree/master/es/novela-espanola> (2015)

Eder, Maciej, Jan Rybicki, y Mike Kestemont.. *stylo*. Kraków. <https://sites.google.com/site/computationalstylistics/stylo> (2013)

Jannidis, Fotis, y Thorsten Vitt. *Pydelta*. Würzburg: University of Würzburg. <https://github.com/fotis007/pydelta> (2014)

Kestemont, Mike, y Folgert Karsdorp. *Pystyl*. Brussels: University of Antwerp. <https://github.com/mikekestemont/pystyl> (2014)

Schöch, Christof, Ulrike Henny, y José Calvo Tello. *The CLiGS toolbox*. Würzburg: University of Würzburg. <https://github.com/cligs/toolbox> (2014)

Schöch, Christof, Ulrike Henny, José Calvo Tello, y Stefanie Popp. *The CLiGS textbox*. Würzburg: University of Würzburg. <https://github.com/cligs/textbox> (2015)

Este mismo texto en la web
http://revistacaracteres.net/revista/vol5n1mayo2016/entendiendo-delta/

SOBRE LOS AUTORES

SOBRE LOS AUTORES

Clara Isabel Arribas Cerezo

Licenciada en Bellas Artes y Máster en Estudios Avanzados en Historia del Arte por la Universidad de Salamanca. Es artista y comisaria independiente, destacando su labor como directora de la Feria de Arte Contemporáneo de Arévalo. Ha coordinado las Jornadas de Profesionalización en Arte Emergente para el Servicio de Actividades Culturales de la Universidad de Salamanca. Su obra artística gira en torno a diferentes formas de representación digital llevadas al terreno de lo artístico.

María Jesús Bernal Martín

Licenciada en Filología Hispánica (2007) y en Teoría de la Literatura y Literatura Comparada (2010) por la Universidad de Salamanca. Certificado de Aptitud Pedagógica (2008). Periodo de Docencia del Programa de Doctorado "Vanguardia y Posvanguardia en España e Hispanoamérica", con estudios de Filosofía (Área de Estética y Teoría de las Artes) (2008). Máster Oficial "La enseñanza del español como lengua extranjera" (2011). Diploma de Estudios Avanzados en la Especialidad de Literatura (2012). En la actualidad (2016), está realizando su tesis doctoral en la Universidad de Salamanca, bajo la tutela de Francisca Noguerol Jiménez, dentro del Programa de Doctorado "Español: Investigación avanzada en Lengua y Literatura".

Núria Calafell Sala

Doctora en Teoría de la Literatura y Literatura Comparada y Licenciada en Filología Hispánica por la Universidad Autónoma de Barcelona, actualmente desarrolla un proyecto titulado "La resistencia de los cuerpos o el sabotaje de una práctica cultural" como Investigadora Asistente del CONICET (Argentina).

José Calvo Tello

Está realizando su tesis doctoral dentro del grupo de investigación *Estilística computacional del género literario* (CliGS, por sus siglas en alemán) en la Universidad de Würzburg. En él investiga los subgéneros de novelas y cuentos españoles de la Edad de Plata mediante métodos cuantitativos como la estilometría. Además participa de otros proyectos de investigación y edición como la colección de eBooks *Clásicos Hispánicos*.

Anabel Fernández Moreno

Doctora en Historia del Arte por la Universidad de Granada. Su línea de investigación transita el campo de la museología contemporánea, centrándose en aspectos como la comunidad, las nuevas tecnologías o la propiedad de los bienes culturales. Su publicación más reciente es una monografía para la editorial Trea titulada: *¿De quién es ese Rembrandt?* (2015).

Juan Gil Segovia

Realiza actualmente su tesis doctoral sobre la pervivencia de la fotografía química en la era digital. Ha obtenido la licenciatura en Bellas Artes y el Máster en Estudios Avanzados en Historia del

Arte en la Universidad de Salamanca, en la que ha impartido docencia en el área de Didáctica de la Expresión Plástica. Compagina la práctica artística con la gestión cultural: ha realizado multitud de exposiciones, ha obtenido diversos premios y ha comisariado varios proyectos artísticos.

Aitana Martos García

Doctora en Documentación por la Universidad de Extremadura con Premio Extraordinario de Doctorado. Ha desarrollado tareas profesionales en diversos archivos y centros de documentación y ha publicado artículos como "Prosopografías comparadas de lamias, sirenas y otros genios acuáticos" (2013). Está vinculada también a la Red de Universidades Lectoras, en especial al Centro de Documentación de Estudios de Lectura y Escritura, del cual es la documentalista técnica.

Eloy Martos Núñez

Coordinador General de la Red Internacional de Universidades Lectoras. Ha publicado artículos como "Las leyendas regionales como intangibles territoriales" (2015) o "De los espacios de lectura a los espacios letrados" (2012) y ha coordinado obras colectivas como el *Diccionario de nuevas formas de lectura y escritura* (2013) junto a María del Mar Campos Fernández-Fígares.

Candela Salgado Ivanich

Graduada en Estudios Franceses (2011-2015) por la Universidad de Salamanca. Actualmente, cursa el Máster de

Literatura Española e Hispanoamericana, Teoría de la Literatura y Literatura Comparada también en la Universidad de Salamanca.

Carmela Tomé Cornejo

Doctora en Lengua Española por la Universidad de Salamanca. Sus principales líneas de investigación son la disponibilidad léxica, el procesamiento léxico, la enseñanza de español como lengua extranjera y la enseñanza en línea. Actualmente, forma parte de la Unidad de I+D+i de Cursos Internacionales de la Universidad de Salamanca y es profesora del Departamento de Filología de la Universidad de Burgos. Cabe destacar su participación en la elaboración de la *Gramática de referencia para la enseñanza de español* (2013), coordinada por Julio Borrego Nieto, así como en la corrección del manual de la *Nueva gramática de la lengua española* de la RAE (2010) o en la preparación del Corpus del Español del Siglo XXI (RAE y ASALE). En el ámbito de la disponibilidad léxica, destaca la coautoría de "Cognitive factors of lexical availability in a second language", publicado por Springer en el volumen *Lexical Availability in English and Spanish as a Second Language,* coordinado por Rosa María Jiménez Catalán.

Este mismo texto en la web
http://revistacaracteres.net/revista/vol5n1mayo2016/sobre-los-autores/

PETICIÓN DE CONTRIBUCIONES – CALL FOR CONTRIBUTIONS

Caracteres. Estudios culturales y críticos de la esfera digital es una publicación académica independiente **en torno a las Humanidades Digitales** con un reconocido consejo editorial, especialistas internacionales en múltiples disciplinas como consejo científico y un sistema de selección de artículos de doble ciego basado en informes de revisores externos de contrastada trayectoria académica y profesional. **El próximo número (vol. 5 n. 2, noviembre 2016) está abierto a la recepción de colaboraciones**.

Los temas generales de la revista comprenden las disciplinas de Humanidades y Ciencias Sociales en su medicación con la tecnología y con las Humanidades Digitales. **La revista está abierta a recibir contribuciones misceláneas dentro de todos los temas de interés para la publicación.**

La revista está abierta a la recepción de artículos todo el año, pero hace especial hincapié en los tiempos máximos para garantizar la publicación en el número más próximo. Puede consultar las normas de publicación y la hoja de estilo a través de la sección específica de la web <http://revistacaracteres.net/normativa/>. Para saber más sobre nuestros objetivos, puede leer nuestra declaración de intenciones. **La recepción de artículos para el siguiente número se cerrará el 2 de octubre de 2016** (las colaboraciones recibidas con posterioridad a esa fecha podrían pasar a un número posterior). Los artículos deberán cumplir con las normas de publicación y la hoja de estilo. Se enviarán por correo electrónico a articulos@revistacaracteres.net.

Caracteres se edita en España bajo el ISSN 2254-4496 y está recogida en bases de datos, catálogos e índices nacionales e internacionales como **ERIH Plus, Latindex, MLA**, Fuente Académica Premier o DOAJ. Puede consultar esta información en la sección correspondiente de la web <http://revistacaracteres.net/bases-de-datos/>.

Le agradecemos la posible difusión que pueda aportar a la revista informando sobre su disponibilidad y periodo de recepción de colaboraciones a quienes crea que les puede interesar.

PETICIÓN DE CONTRIBUCIONES – CALL FOR CONTRIBUTIONS

Caracteres. Estudios culturales y críticos de la esfera digital is an independent **journal on Digital Humanities** with a renowned editorial board, international specialists in a range of disciplines as scientific committee, and a double blind system of article selection based on reports by external reviewers of a reliable academic and professional career. **The next issue (vol. 5 n. 2, Nov 2016) is now open to the submission of contributions**.

The general topics of the journal include the disciplines of Humanities and Social Sciences in its mediation with the technology and the Digital Humanities. **The journal is now open to the submission of miscellaneous contributions** within all the relevant topics for this publication.

While the journal welcomes submissions throughout the year, it places special emphasis on the advertised deadlines in order to guarantee publication in the latest issue. Both the publication guidelines and the style sheet can be found in a specific section of our webpage <http://revistacaracteres.net/normativa/> . To know more about our objectives, the declaration of principles of the journal can be consulted. **The deadline for the reception of papers is October 2nd, 2016** (contributions submitted at a later date may be published in the next issue). Articles should adhere to the publication guidelines and the style sheet, and should be sent by email to articulos@revistacaracteres.net.

Caracteres is published in Spain (ISSN: 2254-4496) and it appears in national and international catalogues, indexing organizations and databases, such as **ERIH Plus, Latindex, MLA**, Fuente Académica Premier or DOAJ. More information is available in the website <http://revistacaracteres.net/bases-de-datos/>.

We appreciate the publicity you may give to the journal reporting the availability and the call for papers to those who may be interested.

Caracteres. Estudios culturales y críticos de la esfera digital

http://revistacaracteres.net

Mayo de 2016. Volumen 5 número 1
http://revistacaracteres.net/revista/vol5n1mayo2016

Contenidos adicionales

Campo conceptual de la revista Caracteres
http://revistacaracteres.net/campoconceptual/
Blogs
http://revistacaracteres.net/blogs/

Síguenos en

Twitter
http://twitter.com/caracteres_net

Facebook

http://www.facebook.com/RevistaCaracteres