Propuesta recibida el octubre 2018 y aceptada para publicación el junio 2019.

SAUTEE: un recurso en línea para análisis estilométricos

SAUTEE: an online resource for stylometric analysis

Fernanda López-Escobedo
Universidad Nacional Autónoma de México
flopeze@unam.mx

Gerardo Sierra
Universidad Nacional Autónoma de México
gsierram@iingen.unam.mx

Julián Solórzano
Universidad Nacional Autónoma de México
jsolorzanos@iingen.unam.mx

Resumen

La estilometría es la cuantificación del estilo por medio de la búsqueda de rasgos textuales que sean medibles y representativos del estilo de un autor. No existen muchas aplicaciones dirigidas al público en general que permitan realizar estudios de esta naturaleza, y las que existen son relativamente limitadas o no necesariamente amigables al usuario. En este artículo presentamos una aplicación web para análisis estilométrico. La aplicación está respaldada por un gestor de corpus, es de fácil manejo y presenta los resultados de manera intuitiva, sin dejar de lado la visión de ofrecer un catálogo exhaustivo de marcadores estilométricos y métodos de análisis.

Palabras clave

estilometría, atribución de autoría, lingüística forense

Abstract

Stylometry is a method that quantifies writing styles by isolating and counting distinctive and measurable textual features of an individual's style. Currently, there are few software applications, aimed at a wide user base, capable of performing such an analysis. Of those readily available, most suffer from limited computing power and/or are not user-friendly. In contrast, our web-based stylometric application, backed by a robust corpus manager, is easy to use, offers a thorough catalogue of stylometric markers and analytic methods to choose from, and produces an intuitive readout of the results.

Keywords

stylometry, authorship attribution, forensic linguistics

1 Introducción

Tradicionalmente se han aplicado los estudios del estilo literario a problemas cronológicos y obras de autoría disputada, como por ejemplo el caso de las obras de Shakespeare. Desde finales del siglo XIX se han intentado establecer métodos numéricos y estadísticos que permitan medir el estilo de un autor y hasta hace poco esta tarea era vista principalmente como una ayuda complementaria en estudios de humanidades. Sin embargo, desde la segunda mitad del siglo XX estas ideas empezaron a ser de interés para el ámbito legal, por ejemplo véase Svartvik (1968), en donde se usaron técnicas estadísticas para demostrar que las supuestas confesiones de Timothy John Evans —hombre acusado de asesinar a su esposa e hija y condenado a muerte en 1950— fueron alteradas por la policía. Esta evidencia fue importante para su perdón póstumo (Nieto et al., 2008). En los 90's el lingüista Malcom Coulthard dio forma a la subdisciplina que hoy en día se conoce como lingüística forense. El testimonio presentado por Coulthard en el caso de otro hombre acusado de asesinato, Derek Bentley, demostró, similar a lo ocurrido en el caso Evans, que la supuesta confesión grabada había sido fabricada (Coulthard, 1994).

La estilometría es una línea de investigación dentro del ámbito de la lingüística forense que tiene como objetivo cuantificar el estilo o, en otras palabras, analizarlo estadísticamente. Para ello se busca identificar ciertos rasgos que sean comunes en el lenguaje pero que sean característicos de cada autor. Es decir, se basa en el supuesto de que hay un factor inconsciente, pero distintivo y medible, en el estilo de escritura de cada persona (Holmes, 1998).

Se considera que uno de los primeros trabajos en este ámbito es el realizado por el físico Thomas Mendenhall en 1887, quien analizó en las obras de Shakespeare la distribución de la frecuencia de las palabras de acuerdo con su longitud (Mendenhall, 1887). Aunque no demostró nada contundente, surgió un interés por el te-

DOI: 10.21814/lm.11.1.270

ma y en las décadas subsecuentes se propusieron otros marcadores estilométricos, con éxito moderado. Más tarde, en 1964, el trabajo de Mollester y Wallace acerca de la autoría de "The Federalist Papers" (Mosteller & Wallace, 1964) se posicionó como un parteaguas en el área debido a sus convincentes resultados y su entonces novedosa técnica Bayesiana. En los años posteriores se adoptaron técnicas de estadística multivariada y varios tipos de análisis usando algoritmos de aprendizaje de máquina (*machine learning*), como por ejemplo máquinas de soporte de vectores (Diederich et al., 2003) y redes neuronales artificiales (Tweedie et al., 1996).

La estilometría y la atribución de autoría siguen siendo objeto de estudio y polémica, ya que no se ha podido definir un conjunto de marcadores estilométricos universales que puedan consistentemente identificar a cualquier autor en cualquier situación; ni tampoco se ha aceptado una metodología universal para llevar a cabo una tarea de atribución de autoría en el ámbito forense. Hasta el día de hoy se lucha por encontrar un protocolo estándar, por ejemplo véase la reciente propuesta de Juola (2015).

Con el afán de incrementar la investigación en el área y hacerla más conocida, es deseable contar con herramientas utilizables por usuarios finales que les permitan conocer y evaluar las diversas técnicas existentes. Esto es, por ejemplo, acercar al área a lingüistas que no necesariamente están familiarizados con la estadística o la computación, a profesionistas del ámbito legal que deseen evaluar la confiabilidad de los resultados, y otras personas interesadas que no sean expertos en análisis cuantitativo. De hecho, actualmente no existen muchas opciones de software que cumplan con estas características, por lo menos no para el uso del público en general.

El presente artículo tiene como objetivo presentar un nuevo sistema que cumple con las características de ser amigable al usuario sin dejar de lado el poder de sus análisis. Además, tiene la particularidad de ser una aplicación web, con miras a explotar ventajas como son el hecho de no requerir instalación, de poder usarse desde cualquier computadora en cualquier momento y de estar respaldado por un gestor de corpus colaborativo.

En la siguiente sección se presentan las herramientas existentes más conocidas, discutiendo brevemente sus ventajas y desventajas. En la Sección 3 se describe el marco del proyecto en el que surge SAUTEE. En la Sección 4 se establecen las bases teóricas de la metodología con la cual opera el sistema, específicamente la selección de marcadores estilométricos, el cálculo de la similitud entre documentos y la técnica de visualización de los resultados. En la Sección 5 se describe a detalle el funcionamiento del sistema desde el punto de vista de la interfaz de usuario, seguido de un pequeño ejemplo de uso en la Sección 6. Finalmente, en la última sección se presentan conclusiones.

2 Recursos existentes

Con el fin de presentar los recursos existentes es importante tomar en cuenta que para hacer un análisis estilométrico de textos se siguen, en general, tres etapas: preprocesamiento, determinación de marcadores estilométricos y análisis estadístico. Este *pipeline* es muy común en tareas de procesamiento de textos y Juola et al. (2006) las describen para un sistema de atribución de autoría:

- **Preprocesamiento:** Son todas aquellas modificaciones que se hacen al texto antes de su procesamiento. Juola maneja esta fase bajo el nombre de canonización.
- **Determinación de marcadores estilométricos:** Es la especificación, lo que se va a medir en los textos. Por ejemplo, palabras, n-gramas de palabras, etc. Juola llama a esta fase selección del conjunto de eventos.
- **Selección del método de análisis:** Es la selección del método por el cual se hará el análisis estadístico para presentar conclusiones, resultados, gráficas, etc.

Bajo estos tres puntos se analizan 3 herramientas de análisis estilométricos, además de presentar sus ventajas y desventajas. Todas estas herramientas son gratuitas y están disponibles en la web para su descarga.

2.1 Signature

The Signature Stylometric System[1] es una aplicación de escritorio desarrollada por el profesor Peter Millican de la Universidad de Leeds.

2.1.1 Pipeline

Preprocesamiento. Signature hace un preprocesamiento mínimo de los textos. Uno de ellos es convertir todo el texto a mayúsculas. Además permite combinar varios textos en uno solo.

[1] Disponible en `http://www.philocomp.net/humanities/signature.htm`

Determinación de marcadores estilométricos. El programa realiza el conteo de unas características predeterminadas que son: distribución de longitud de palabras, oraciones y párrafos, frecuencia de letras y uso de signos de puntuación. El usuario puede adicionalmente ingresar una lista de palabras para incluirlas en el análisis. Fuera de esto no hay ninguna otra opción o parámetro relativo a esta etapa.

Selección de método de análisis. El método de análisis por defecto es la visualización de las frecuencias de los marcadores estilométricos en una gráfica del tipo histograma. Los corpus o documentos seleccionados aparecen cada uno con un color diferente. No hay otra manera de visualización, salvo la elección de ver la gráfica en 2D o 3D. El otro método de análisis es la prueba de la Chi cuadrada. Para esto se solicita que el usuario elija dos documentos (o un documento y un conjunto de documentos combinado). El programa determinará automáticamente si los rasgos presentes en los documentos permiten hacer la prueba (ya que la prueba requiere que los valores de las frecuencias rebasen cierto umbral). Si el análisis procede, el programa reportará el p-valor derivado de la prueba y su interpretación tradicional.

2.1.2 Ventajas

La aplicación ofrece la funcionalidad de "combinar archivos en corpus" por medio de la cual un conjunto de textos pueden ser combinados en uno solo. De esta manera se pueden combinar en un solo corpus todos los textos de un mismo autor, permitiendo comparar un documento individual (dubitado o texto de autoría desconocida) contra todos los demás textos de un determinado autor.

La gráfica resultante muestra de manera intuitiva las diferencias de cada marcador entre los distintos corpus. Además, Signature ofrece la prueba de la χ^2 para determinar la similitud entre dos textos de una manera cuantitativa.

2.1.3 Desventajas

Para visualizar los resultados, el programa crea una gráfica para palabras, otra para puntuación, otra para letras, etc. La gráfica generada no puede mostrar el acumulado de las diferencias de todos los marcadores al mismo tiempo, lo cual es una limitante. Por otro lado, el tipo de gráfica es impráctico para una lista de palabras de extensión considerable, puesto que termina haciéndose muy larga en el eje horizontal.

Además, el catálogo de marcadores estilométricos no es muy extenso y no tiene los más usuales, que son n-gramas de palabras y de caracteres (para $n > 1$).

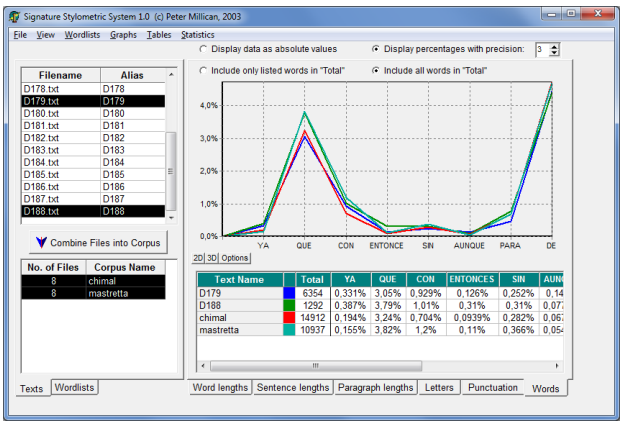

Figura 1: Interfaz de Signature.

2.2 JGAAP

JGAAP[2] es un proyecto desarrollado por Patrick Juola de la Universidad Duquesne (Juola, 2009), está diseñado para permitir a no-expertos en el área de aprendizaje de máquina un acercamiento a este tipo de técnicas, así como para facilitar la comparación entre la efectividad de varios métodos.

La interfaz gráfica (Figura 2) de este programa es muy organizada puesto que se presentan diferentes pestañas, cada una de las cuales indica las opciones para cada fase del pipeline.

2.2.1 Pipeline

Preprocesamiento. Ya que Juola maneja esta etapa como una parte esencial del análisis, la aplicación contiene una serie de opciones dedicadas únicamente a la misma. El usuario puede determinar exactamente qué tipo de procesamiento se hará al texto, como por ejemplo, eliminar caracteres especiales, eliminar signos de puntuación, entre otros.

Determinación de marcadores estilométricos. El programa cuenta con un extenso catálogo de diversos marcadores. El usuario puede elegir uno o más, y además especificar los parámetros de cada uno, de ser necesario. Por ejemplo, al elegir n-gramas se pedirá que se especifique el valor de n. De esta manera, el usuario puede pedir que se analicen bigramas o trigramas de palabras

[2]Disponible en https://github.com/evllabs/JGAAP

o bigramas o trigramas de etiquetas POS (*Part of Speech*), entre otros.

Selección del método de análisis. De manera similar a la determinación de marcadores, al usuario se le presenta una extensa lista de métodos de análisis, entre los que se encuentran, Análisis de Componentes Principales, Análisis Discriminante Lineal y Máquinas de Soporte de Vectores.

2.2.2 Ventajas

Contiene un extenso catálogo tanto de marcadores estilométricos como de métodos de análisis, los cuales son parametrizables.

2.2.3 Desventajas

La descripción de los marcadores estilométricos no es muy informativa.

La salida del programa es únicamente texto. No muestra ninguna gráfica ni diagrama, ni es posible exportar los datos a una hoja de cálculo.

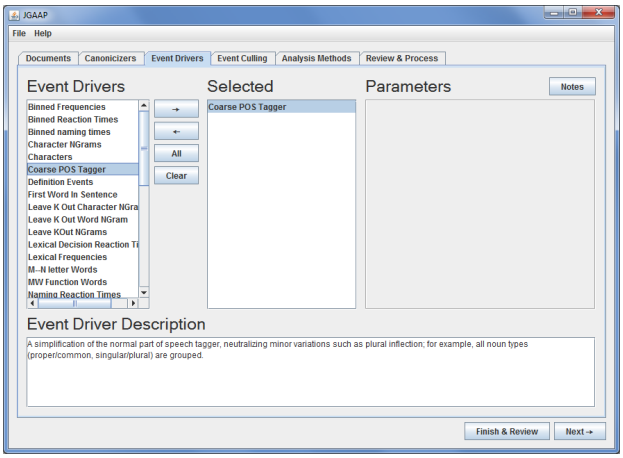

Figura 2: Interfaz de JGAAP.

2.3 Stylo

Stylo es un paquete diseñado para el entorno estadístico R (R Core Team, 2008). Ofrece una interfaz gráfica de usuario (Figura 3), de manera que no es necesario escribir un programa para poder usar sus funcionalidades.

2.3.1 Pipeline

Preprocesamiento. El procesamiento es mínimo: las únicas opciones que el usuario puede seleccionar son si desea que se preserven las mayúsculas (ya que por omisión este programa convierte todo el texto a minúsculas), y si desea que se eliminen los pronombres.

Determinación de marcadores estilométricos. Únicamente hay dos tipos de marcadores: palabras y caracteres. Sin embargo, hay varios parámetros manipulables por el usuario. Se puede elegir el tamaño de los n-gramas, así como el número de palabras que entrarán dentro del análisis. Es decir, se puede determinar que solo se usen las 100 palabras más frecuentes, o 50 o las que se deseen.

Selección del método de análisis. Se presenta al usuario las opciones para llevar a cabo el análisis estadístico, específicamente el tipo de análisis y el tipo de distancia.

2.3.2 Ventajas

Ofrece varias opciones de visualización de resultados, incluyendo escalamiento multidimensional, análisis de componentes principales y análisis de clusters. Asimismo, se puede hacer uso de varios métodos de clasificación como vecinos más cercanos, Bayes ingenuo, SVM, entre otros.

Además, gracias a que corre dentro del ambiente R, se puede hacer cualquier otro tipo de análisis con los datos generados, siempre que el usuario conozca el uso de este lenguaje.

2.3.3 Desventajas

Solo hace análisis con n-gramas de palabras y de caracteres, no hay ningún otro marcador estilométrico disponible.

La carga del corpus no es tan fácil como en las otras aplicaciones, pues se basa en preparar una estructura de carpetas determinada y seguir una convención para los nombres de los archivos.

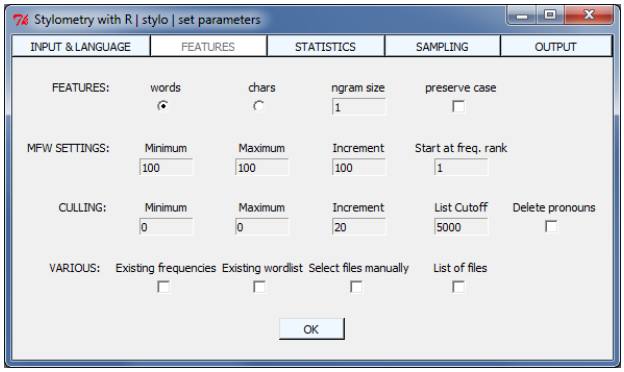

Figura 3: Interfaz de stylo.

3 Marco del proyecto

Una de las tareas del procesamiento automático de lenguaje natural que ha cobrado mayor importancia a últimas fechas es la detección de similitud textual. Esta labor responde a diversas necesidades, tales como la clasificación textual, la identificación de autoría, el análisis de reutilización de textos, la detección de paráfrasis y la detección de plagio. Con el fin de contribuir en el desarrollo de las metodologías existentes para la medición de similitud textual en documentos utilizando diferentes enfoques, tanto lingüísticos como estadísticos, se propuso la creación de un recurso lingüístico.

El objetivo fue generar una herramienta pública que sea de utilidad a los usuarios y fomente la colaboración. En este sentido, resultaba crucial la creación de un repositorio central de documentos, o gestor de corpus, en el que cualquier persona interesada pueda registrarse para poder cargar sus propios documentos. Este sistema tiene por nombre GECO: Sistema de Gestión de Corpus (Sierra et al., 2017). Los textos cargados pueden ser, opcionalmente, puestos a disposición de todos los demás usuarios de la plataforma, de manera que cada vez se puedan tener corpus más robustos.

El sistema presentado en este artículo, SAUTEE (Sistema Automático para Estudios Estilométricos)[3], es un sistema en web accesible desde cualquier computadora con conexión a Internet, el cual permite al usuario analizar la aparición de diversos marcadores estilométricos en un conjunto de documentos. El SAUTEE se alimenta de los documentos de los corpus creados a través de GECO, y es la primera de varias herramientas planeadas para sacar provecho de este repositorio de documentos. Cada herramienta tendrá un fin en específico, que en el caso del SAUTEE es el análisis estilométrico.

4 Fundamentos teóricos del sistema

El análisis realizado por el SAUTEE se basa en el cálculo de distancias entre los documentos. Para calcular la distancia que hay entre dos textos es preciso primeramente representar cada texto de una forma numérica, es decir, vectorizarlos. Una vez hecho esto y calculadas las distancias, se aplica un método de visualización de datos por medio del cual es posible apreciar la similitud entre cada documento. El SAUTEE realiza entonces las siguientes tres tareas:

- La vectorización de los textos, es decir, la extracción de marcadores estilométricos y su cuantificación.
- El cálculo de una distancia entre cada par de documentos.
- La generación de un conjunto de puntos en 2 dimensiones, por medio del cual se pueden visualizar en una gráfica las distancias calculadas.

4.1 Extracción de marcadores estilométricos y Vectorización

4.1.1 Marcadores de estilo

La estilometría se basa en identificar rasgos que puedan ser cuantificados y que sean característicos del estilo del autor. Bailey (1979) sugiere que estos rasgos deben ser:

- salientes;
- estructurales;
- frecuentes y fácilmente cuantificables;
- relativamente inmunes al control consciente.

Reciben en la literatura muchos nombres, uno de los más usuales es "marcadores de estilo". Normalmente este término hace referencia a la categoría general de la característica de la cual se está hablando. Por ejemplo, si se determina que la frecuencia de uso de ciertos signos de puntuación es una característica distintiva, entonces se dice que "signos de puntuación" es un marcador estilométrico y, dentro de esta categoría, existirán características específicas como lo son punto, coma, punto y coma, etcétera. En SAUTEE adoptamos el término "marcadores estilométricos" para referirnos a estas categorías.

A lo largo del tiempo se han identificado e investigado cientos de marcadores estilométricos — Rudman (1997) estima 1000 — sin embargo no se ha encontrado un conjunto que funcione para cualquier situación. Muchas veces el dominio en el que se lleva a cabo la investigación dictará el tipo de marcadores que se requieren. Por ejemplo, el tipo de marcadores estilométricos no será el mismo para analizar una novela que para analizar una publicación en una red social. Para hacer un recuento de los marcadores que son comúnmente utilizados en el área, se puede hacer una clasificación de acuerdo con diversos criterios. Por ejemplo, la presentada por Stamatatos (2009) en su revisión del tema, divide los marcadores en las siguientes categorías:

[3]Disponible en `http://www.corpus.unam.mx/saute`

- **Léxicos.** Aquí se encuentran todas aquellas medidas a nivel palabra, por ejemplo n-gramas de palabras, longitud de palabras y oraciones, riqueza de vocabulario, etc.
- **Caracter.** Son aquellas medidas a nivel caracter, por ejemplo n-gramas de caracteres y conteos de caracteres específicos (por ejemplo dígitos).
- **Sintácticos.** Primordialmente se incluyen aquí las etiquetas POS (Part of speech) que identifican a cada palabra con su parte de la oración. Estas etiquetas pueden contener tanta información como sea necesario, por ejemplo, para el caso de los verbos: persona gramatical, número, tiempo verbal, etc. La estructura sintáctica de una frase puede ser representada por n-gramas de etiquetas de parte de la oración, aunque también pueden usarse soluciones más complejas como parsers o chunkers.
- **Semánticos.** Statamatos cita tres principales: dependencias semánticas, análisis de sinónimos e hiperónimos, y "características funcionales". Estas últimas asignan un rol a las palabras dentro del discurso, por ejemplo "elaboración", "clarificación", etc.
- **Específicos a la aplicación.** Son aquellas características que dependen del dominio en el que se está llevando a cabo el experimento y que pueden ser estructurales, específicas del contenido o específicas del lenguaje. Por ejemplo, al analizar textos de internet podemos extraer características como nombres de usuario, etiquetas HTML, entre otras.

Para lograr extraer cada tipo de característica se necesitan herramientas diferentes. Por ejemplo, para poder hacer análisis usando información de etiquetas POS, se necesita una herramienta capaz de hacer este etiquetado. En el caso de SAUTEE, esta herramienta es Freeling (Padró & Stanilovsky, 2012). Freeling es una suite de análisis del lenguaje desarrollada en la Universidad Politécnica de Cataluña bajo la dirección de Lluís Padró.

4.1.2 Catálogo de Marcadores del SAUTEE

A continuación se describen los marcadores estilométricos con los que cuenta SAUTEE al momento de escribir este artículo. Todos ellos se basan en el conteo de las apariciones de ciertas características en el texto. Además, ya que cada texto tiene diferente longitud, se lleva a cabo una normalización de manera que las frecuencias utilizadas son relativas.

Signos de puntuación. Se contabilizan los signos de puntuación del texto con base en el etiquetado de Freeling (es decir, se toma como signo de puntuación todo lo que Freeling etiqueta como tal). Cada frecuencia se divide entre el número total de signos de puntuación en el texto.

Distribución de longitud de oraciones y palabras. Se contabilizan las frecuencias de aparición de las siguientes categorías de palabras: palabras de 1 letra, palabras de 2 letras, palabras de 3 letras, sucesivamente hasta 20 letras. Cada frecuencia se divide entre el número total de palabras en el texto. Respecto a la longitud de las oraciones se tienen las categorías: menos de 10 palabras, de 11 a 20, de 21 a 30, de 31 a 40, de 41 a 50, y más de 51. Cada frecuencia se divide entre el número total de oraciones en el texto.

Categoría gramatical al inicio de la oración. A partir de las etiquetas POS generadas por Freeling, se contabiliza el número de veces que cada categoría gramatical aparece al inicio de una oración. Por ejemplo, cuántas veces un verbo inicia la oración, cuántas veces un sustantivo, y así sucesivamente. Cada frecuencia es dividida entre el número total de palabras al inicio de la oración (o lo que es lo mismo, entre el número total de oraciones del texto).

Categoría gramatical al final de la oración. Lo mismo que la anterior, pero considerando las palabras al final de la oración.

Unigramas de palabras funcionales. Se contabiliza la aparición de las palabras funcionales (de acuerdo con la lista de palabras funcionales cargada actualmente en el sistema). La frecuencia de cada palabra se divide entre el total de apariciones de palabras funcionales contabilizadas en el texto.

Bigramas de palabras funcionales. Lo mismo que la anterior, pero considerando bigramas, es decir todas aquellas apariciones de dos palabras funcionales seguidas. Por ejemplo, tomando como palabras funcionales los artículos y las conjunciones, en el segmento "el niño y la niña" se contabilizaría como bigrama de palabras funcionales "y la". La frecuencia de cada bigrama es dividida entre el total de bigramas de palabras funcionales contabilizadas en el texto.

Trigramas de palabras funcionales. Lo mismo que la anterior pero tomando en cuenta

las apariciones de tres palabras funcionales seguidas. La frecuencia de cada trigrama es dividida entre el total de trigramas de palabras funcionales contabilizadas en el texto.

Bigramas de palabras funcionales con hasta 2 huecos. En este caso se contabilizan las apariciones de dos palabras funcionales que no están contiguas, sino que están separadas a lo más por otras dos palabras. Por ejemplo, en la frase "el niño y la niña", se contabilizaría el bigrama "el y", cuyos elementos están a una palabra de separación (en este caso, "niño"). Cada frecuencia se divide entre el total de bigramas de palabras funcionales con hasta 2 huecos contabilizadas en el texto.

Trigramas de palabras funcionales con hasta 2 huecos. Lo mismo que la anterior pero considerando apariciones de tres palabras funcionales. No necesariamente tiene que haber el mismo número de huecos entre la primera y la segunda palabra funcional que entre la segunda y la tercera. Por ejemplo, la primera y la segunda palabra funcional pueden estar a una separación de una palabra, y la segunda y la tercera a una distancia de dos. Cada frecuencia se divide entre el total de trigramas de palabras funcionales con hasta 2 huecos contabilizadas en el texto.

Unigramas de etiquetas POS. Se contabiliza la aparición de las etiquetas POS tal como Freeling las genera. La frecuencia de cada etiqueta se divide entre el número de palabras en el texto.

Bigramas de etiquetas POS. Se contabilizan las apariciones de dos etiquetas POS contiguas. La frecuencia de cada bigrama se divide entre el total de bigramas de etiquetas POS contabilizadas en el texto.

Trigramas de etiquetas POS. Lo mismo que la anterior pero considerando tres etiquetas contiguas. La frecuencia de cada trigrama se divide entre el total de trigramas de etiquetas POS contabilizadas en el texto.

Unigramas de etiquetas POS no fino. Lo mismo que unigramas de etiquetas POS pero en vez de contabilizar la frecuencia de las etiquetas tal cual las genera Freeling, se toma en cuenta únicamente el primer caracter de la etiqueta que corresponde a la categoría gramatical más general. Por ejemplo, la etiqueta vmii3s0 (verbo principal indicativo imperfecto tercera persona de singular) y la etiqueta vsip3p0 (verbo semiauxiliar imperfecto tercera persona del plural) se agrupan bajo una misma etiqueta "v", verbo. La frecuencia de cada una de estas etiquetas simplificadas se divide entre el número total de palabras.

Bigramas de etiquetas POS no fino. Lo mismo que la anterior pero contabilizando las apariciones de dos etiquetas seguidas. La frecuencia de cada bigrama se divide entre el total de bigramas de etiquetas POS no fino contabilizadas en el texto.

Trigramas de etiquetas POS no fino. Lo mismo que la anterior pero contabilizando las apariciones de tres etiquetas seguidas. La frecuencia de cada trigrama se divide entre el total de trigramas de etiquetas POS no fino contabilizadas en el texto.

Bigramas de caracteres. Se contabilizan las apariciones de dos caracteres seguidos. Los espacios se consideran caracteres. Por ejemplo en el segmento "el niño", los bigramas de caracteres son "el", "l_", "_n", "ni", "iñ", "ño" (el guión bajo representa un espacio). La frecuencia de cada bigrama se divide entre el total de bigramas de caracteres contabilizados en el texto.

Trigramas de caracteres. Lo mismo que la anterior pero considerando tres caracteres seguidos. La frecuencia de cada trigrama se divide entre el total de trigramas de caracteres contabilizados en el texto.

4.1.3 Vectorización del texto

Una vez seleccionados los marcadores se crea un vector por cada documento. A continuación se presenta un ejemplo de la creación de estos vectores. Sea el texto (1),

> *La cantante de ópera deleitó al público en la función de anoche.*

El preprocesamiento hecho por Freeling obtiene el lema y la etiqueta POS de cada palabra del texto, como se puede ver en el cuadro 1.

De acuerdo a los marcadores estilométricos elegidos por el usuario, se hacen los respectivos conteos de aparición en el texto. Por ejemplo, sea el marcador elegido "Unigramas de etiquetas POS no fino", las características obtenidas

Palabra	Lema	POS
La	el	DA0FS0
cantante	cantante	NCCS000
de	de	SPS00
ópera	ópera	NCFS000
deleitó	deleitar	VMIS3S0
a	a	SPS00
el	el	DA0MS0
público	público	NCMS000
en	en	SPS00
la	el	DA0FS0
función	función	NCFS000
de	de	SPS00
anoche	anoche	RG

Cuadro 1: Análisis de Freeling para el texto (1).

usando este marcador y sus respectivos valores se pueden ver en el cuadro 2.

Característica	Frecuencia relativa
D	23.08 %
N	30.77 %
S	30.77 %
V	7.69 %
R	7.69 %

Cuadro 2: Frecuencias relativas para el texto (1).

Sea el texto (2),

> *El mesero trajo la sopa fría y una hora tarde.*

Sus características y frecuencias se muestran en el cuadro 3. Aquí debemos notar que hay dos características que no están presentes en el texto anterior: la conjunción (C) y el adjetivo (A), y que le falta una característica que sí está presente en el primero, la preposición (S). Por lo tanto para que ambos vectores sean susceptibles de ser comparados, deben modificarse para que contengan el mismo número de características. Las características que el otro texto no comparte se llenan con ceros. En el cuadro 4 se puede ver cómo quedan ambos vectores después de hacer las modificaciones necesarias. La siguiente etapa en el análisis es medir la similitud entre los textos mediante la aplicación de una función de distancia a estos vectores.

4.2 Medidas de distancia

Una medida de distancia o métrica es un valor que se define para cada par de elementos de un conjunto, en este caso el conjunto de textos.

Característica	Frecuencia relativa
D	30.00 %
N	30.00 %
V	10.00 %
A	10.00 %
C	10.00 %

Cuadro 3: Frecuencias relativas para el texto (2).

	Texto 1	Texto 2
D	23.08	30
N	30.77	30
S	30.77	0
V	7.69	10
R	7.69	10
A	0	10
C	0	10

Cuadro 4: Vectores para el texto (1) y el texto (2).

Busca cuantificar la disimilitud entre cada par de elementos de manera que la medida tenga una magnitud mayor para elementos no semejantes. Es una función que toma como entrada dos elementos y da como resultado un único número. Si el resultado de aplicar la función a los vectores de dos textos es un número pequeño significa que éstos poseen un estilo similar. Desde luego el significado de pequeño variará de acuerdo al tipo de medida seleccionada, y dependerá del valor de los demás elementos del corpus, es decir, es un valor relativo. A continuación se presentan las distancias con que el sistema cuenta actualmente.

4.2.1 Distancia Euclidiana

Para todo par de puntos a y b, la distancia euclidiana representa el camino más corto entre ellos, es decir una línea recta. Es la distancia que puede resultar más intuitiva puesto que para el caso de 2 y 3 dimensiones, es equivalente a nuestra idea de distancia en el mundo real; sin embargo, puede ser generalizable a cualquier número de dimensiones. En este caso, las dimensiones corresponden a cada una de las características contabilizadas. Matemáticamente, la distancia euclidiana es igual a la raíz cuadrada de la suma del cuadrado de las diferencias de cada dimensión y se expresa por medio de la siguiente fórmula:

$$\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

En donde estamos hablando de que hay n características en cada documento, y que X y Y representan los vectores de los dos documentos, es decir, que X_i representa el valor de la i-ésima característica en el primer documento y Y_i representa el valor de esa misma característica en el segundo documento.

4.2.2 Distancia Manhattan

La distancia Manhattan es también llamada distancia de taxista, haciendo referencia a la distancia que un vehículo tendría que recorrer para llegar de un punto a otro en una cuadrícula. Matemáticamente es igual a la suma de las diferencias absolutas entre cada dimensión, como lo muestra la siguiente ecuación:

$$\sum_{i=1}^{n} |X_i - Y_i|$$

4.2.3 Delta de Burrows

La Delta de Burrows es un método específicamente desarrollado para medir la diferencia estilística entre un conjunto de documentos. Originalmente publicado en 2002, se ha convertido en un referente de los estudios de autoría. Se basa en contar la frecuencia de un conjunto de palabras en un texto y calcular el z-score de cada una, es decir, su número de desviaciones estándar sobre la media. La Delta, tal y como fue definida por Burrows, es el promedio de las diferencias absolutas entre los z-scores de un conjunto de palabras en un grupo de textos y los z-scores del mismo conjunto de palabras en el texto objetivo (Burrows, 2002).

Stein & Argamon (2006) demuestram, que matemáticamente esta definición es equivalente a una distancia Manhattan ponderada, en donde el peso en cada dimensión corresponde a la desviación estándar de esa dimensión. Específicamente, es igual a:

$$\sum_{i=1}^{n} \frac{|X_i - Y_i|}{\sigma_i}$$

En donde X son las palabras correspondientes al primer texto y Y las palabras correspondientes al segundo texto, n es el total de palabras y σ es la desviación estándar de la frecuencia de cada palabra. SAUTEE hace el cálculo de la Delta usando esta formula.

4.2.4 Distancia Canberra

La distancia Canberra es otro ejemplo de una distancia Manhattan ponderada. En este caso, la diferencia absoluta entre las variables es dividida entre la suma de los valores absolutos de las mismas. Es igual a:

$$\sum_{i=1}^{n} \frac{|X_i - Y_i|}{|X_i| + |Y_i|}$$

Esta distancia tiene la bondad de ser sensible a valores cercanos a cero, por lo cual es útil si el conjunto de datos contiene, tanto valores pequeños, como valores muy grandes.

4.3 Escalamiento multidimensional

El escalamiento multidimensional (MDS por sus siglas en inglés) es una técnica en el área de visualización de datos que permite apreciar las distancias existentes entre un conjunto de objetos. Su objetivo es asignar a cada punto de un conjunto de datos de n-dimensiones, una coordenada en un espacio de menor dimensión (comúnmente 2), de tal modo que las distancias entre los puntos en este nuevo espacio se mantengan lo más parecidas posible a las distancias originales. De esta manera los puntos pueden ser graficados en un plano cartesiano donde se puede ver la relación que existe entre cada uno respecto a los demás. En este caso cada punto representará un documento y su cercanía o lejanía con otros documentos indicará qué tan similar es el estilo de ambos.

El MDS recibe como entrada una matriz de disimilaridad cuyos elementos d_{ij} representan la distancia que hay entre el objeto i y j. El objetivo en concreto del escalamiento multidimensional es encontrar un conjunto de vectores $x_1, ..., x_n$, $x \in \mathbb{R}^N$ tal que $D(x_i, x_j) \approx d_{ij}$ donde N es la nueva dimensionalidad deseada y D es normalmente la distancia euclidiana. De esta manera, si $N = 2$, los vectores resultantes serán coordenadas de dos dimensiones que pueden ser graficadas sin problema en un plano cartesiano o gráfico de dispersión. Por ejemplo, si los datos de entrada fueran las distancias existentes entre un conjunto de ciudades europeas, tras llevar a cabo el MDS se obtendría un conjunto de puntos $\mathbb{R}^2$ cuya gráfica en el plano cartesiano se asemejaría a un mapa de Europa.

Una de las ventajas de hacer un análisis con MDS es que se puede visualizar el efecto que tienen simultáneamente todos los marcadores estilométricos elegidos. Por lo tanto, se pueden hacer

experimentos utilizando diversas combinaciones de estos.

Es importante poder visualizar este efecto simultáneo de un conjunto determinado de marcadores puesto que los marcadores que son discriminativos para un autor no necesariamente lo serán para otro, y por lo tanto resulta ilustrativo hacer varias pruebas con conjuntos de marcadores diferentes.

5 Funcionamiento del sistema

En la figura 4 se puede observar la interfaz principal del SAUTEE. Tiene un diseño basado en pestañas, cada una de las cuales representa un paso en la secuencia del proceso completo (similar al JGAAP). SAUTEE funciona de la siguiente manera: primeramente se solicita al usuario que seleccione los documentos que desea analizar. Después han de elegirse los marcadores estilométricos que serán tomados en cuenta en el análisis. Finalmente es necesario indicar qué tipo de distancia intertextual será calculada entre cada par de documentos. El tipo de distancia elegida puede depender del tipo de marcador estilométrico que se esté analizando (López-Escobedo et al., 2016). Tras llevar a cabo los pasos anteriores, el sistema genera una gráfica que muestra visualmente la distancia entre cada par de documentos, y da al usuario la opción de descargar las estadísticas en un formato de hoja de cálculo.

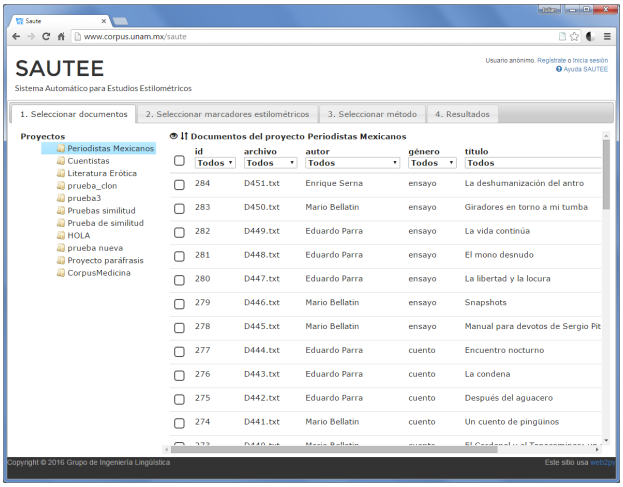

Figura 4: Interfaz del SAUTEE.

5.1 Pipeline general

5.1.1 Preprocesamiento

El preprocesamiento no es llevado a cabo directamente por el SAUTEE. Como se ha mencionado, los textos con los que opera el SAUTEE deben ser cargados primero en la plataforma GECO. Una vez cargados ahí automáticamente pasan por un preprocesamiento que consiste en las siguientes tareas:

- Convierte el archivo a texto plano (en el caso de documentos subidos como PDF o DOC).
- Codifica el texto en UTF-8 (si tenía el texto una codificación distinta).
- Entrega los textos a Freeling para su lematización y etiquetado de partes de la oración (Part of Speech o POS).

Gracias a este preprocesamiento, el SAUTEE trabaja con textos que ya están separados por oraciones y palabras, y en los cuales cada palabra está anotada con su lema y su etiqueta POS. De estos textos preprocesados se extraen las características con las cuales se hará el análisis.

5.1.2 Determinación de características

El SAUTEE solicita al usuario que elija uno o varios marcadores estilométricos de entre los disponibles en el catálogo del sistema y que se describen en el apartado 4.1.2. Esta selección determinará los marcadors con los que se construirá el vector del documento. Si se selecciona más de un marcador, el vector contendrá características de cada uno de ellos.

5.1.3 Selección del método de análisis

El SAUTEE calcula una medida de distancia entre cada par de vectores generados en la etapa anterior. La medida de distancia puede ser especificada por el usuario. A partir de las distancias generadas se realiza el escalamiento multidimensional para producir una gráfica en 2 dimensiones que es presentada al usuario para su análisis.

5.2 Operación a detalle

5.2.1 Selección de documentos

En esta pantalla el sistema muestra un listado de los corpus disponibles. Al seleccionar un corpus el sistema muestra un listado de los documentos individuales que lo conforman. Estos documentos pueden, además, venir acompañados

de una serie de metadatos, como lo son, autor, género literario, entre otros. De esta lista de documentos se tienen que elegir por lo menos dos.

5.2.2 Selección de marcadores estilométricos

En este apartado se muestra un listado de los marcadores estilométricos actualmente disponibles para el análisis y que fueron enumerados en el apartado 4.1.2 de este documento. Como se mencionó en el marco del proyecto, el sistema está pensado para recibir actualizaciones frecuentemente de manera que esta lista de maracadores estilométricos se vea aumentada respondiendo a sugerencias de los usuarios o a avances en el estado del arte del área. Esta pantalla muestra también una pequeña descripción de cada marcador estilométrico, especificando exactamente la manera en que es calculado para un texto dado. El usuario debe seleccionar por lo menos un marcador, pero pueden seleccionarse cualquier número de ellos, incluso todos. El vector generado para calcular las distancias contendrá características de todos los marcadores seleccionados.

5.2.3 Selección de método

La selección de método hace referencia a la forma en que será calculada la distancia entre cada par de documentos. Se incluye en el sistema la explicación de cada método y la bibliografía correspondiente, en su caso. Una vez elegido el método todo está listo para comenzar el análisis. En esta misma pantalla se encuentra un botón que dispara el proceso, el cual puede durar desde unos pocos segundos a algunos minutos dependiendo del volumen de los textos, así como de la cantidad de marcadores estiolométricos elegidos. Una vez que el proceso concluye se muestra la sección de resultados.

5.2.4 Resultados

Esta es la última pantalla, en donde el usuario visualiza el resultado del análisis. Tras el proceso de generación de distancias intertextuales, el método de escalamiento multidimensional asigna a cada documento una coordenada en un espacio de dos dimensiones. Todos estos puntos son mostrados en un diagrama de dispersión de manera que la distancia aparente entre los puntos es proporcional a la distancia intertextual realmente calculada entre los documentos. De esta manera, el usuario puede visualizar la similitud estilística entre todos los documentos inmediatamente. Adicionalmente, los datos numéricos resultantes del proceso se preparan en dos archivos en formato CSV (visualizable en cualquier programa de hoja de cálculo) que el usuario puede descargar para análisis subsecuentes. Uno de estos archivos contiene la frecuencia de uso de cada característica perteneciente a los marcadores estilométricos seleccionados (es decir, los vectores utilizados para calcular las distancias). El otro archivo contiene las distancias intertextuales calculadas entre cada par de documentos (las distancias entre los vectores).

6 Caso de uso

Veamos un ejemplo de uso. Uno de los corpus disponibles por defecto es un corpus llamado Periodistas Mexicanos. Contiene artículos, ensayos y cuentos de 6 escritores mexicanos o residentes en México desde temprana edad: Alberto Chimal, Ángeles Mastretta, Enrique Serna, José de la Colina, Eduardo Parra y Mario Bellatín. Tiene 9 documentos de cada uno de estos autores, dando un total de 54 documentos. En este ejemplo se hará la comparación entre Ángeles Mastretta y José de la Colina.

Se empieza por seleccionar el corpus Periodistas Mexicanos de la lista de proyectos del lado izquierdo. Una vez seleccionado aparecerán del lado derecho los documentos que conforman este corpus. En la columna de autor abrimos el filtro y se selecciona Ángeles Mastretta. Para seleccionar todos los documentos de la autora se selecciona la casilla que se encuentra en el encabezado de la tabla en la primera columna. Se repite el mismo procedimiento pero ahora seleccionando en el filtro a José de la Colina.

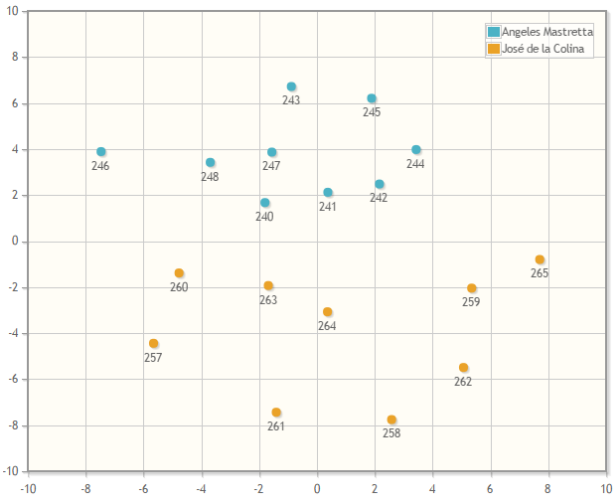

Figura 5: Análisis con unigramas POS. Coloreado por autor.

Una vez seleccionados los 18 documentos se procede a la siguiente pestaña para hacer la selección de marcadores estiolmétricos. Para el primer experimento se selecciona un único marcador: unigramas de etiquetas POS. En la siguiente pestaña se selecciona distancia euclidiana. Una vez terminado el procesamiento se obtiene la gráfica mostrada en la figura 5 (seleccionando la opción de colorear por autor). En la gráfica podemos observar que los dos grupos de textos se separan visiblemente. De hecho, todos los textos de Ángeles Mastretta quedan por arriba del eje x y todos los textos de José de la Colina por debajo. Se puede concluir que los unigramas de etiquetas POS es un buen marcador para diferenciar estos dos autores.

Para un segundo experimento, se seleccionan esta vez unigramas, bigramas y trigramas POS simultáneamente y se vuelve a seleccionar distancia euclidiana. La gráfica resultante se muestra en la figura 6 (esta vez se selecciona la opción de colorear por género literario). Se puede observar que en esta gráfica los dos grupos de textos que se forman claramente son uno conformado por cuentos, y otro conformado por los artículos y ensayos. Se puede concluir que la combinación de unigramas, bigramas y trigramas de etiquetas POS no es un buen marcador en este caso para diferenciar el autor de los textos, sin embargo es útil para diferenciar el estilo del género literario cuento.

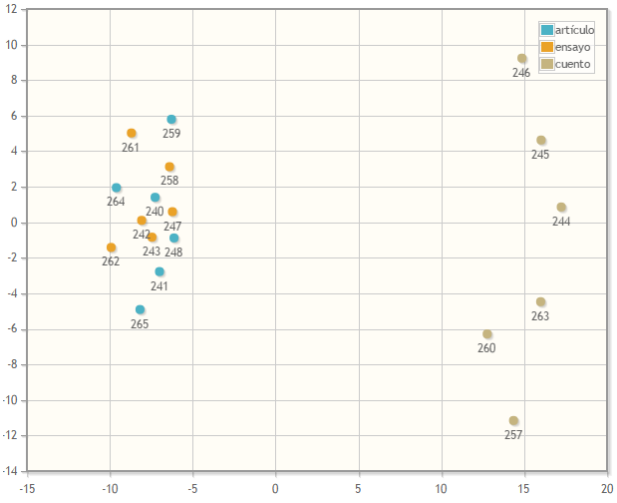

Figura 6: Análisis con unigramas, bigramas y trigramas POS. Coloreado por género literario.

7 Conclusiones

En este artículo se ha presentado un sistema que lleva a cabo extracción de características estilísticas en un corpus especificado por el usuario, que es cargado a un repositorio central en la nube por medio de una herramienta adicional llamada GECO. Es un sistema web, lo cual lo distingue de otras aplicaciones existentes, y permite que haya una mejor administración de los corpus. Toma en consideración un catálogo de marcadores estilométricos que no necesariamente es estático sino que puede ser ampliado por medio de parametrización. De cualquier modo, está pensado para añadir nuevos marcadores conforme se vaya necesitando y se propongan nuevos. Por otro lado, la presentación de los resultados en el sistema es por medio de una técnica de escalamiento multidimensional, lo cual permite intuitivamente apreciar la similitud y disimilitud entre todos los documentos del corpus, y usando una combinación de cualquier número de marcadores estilométricos.

Agradecimentos

Se agradece el apoyo recibido por el Consejo Nacional de Ciencia y Tecnología, proyecto 2016-01-2225 y a DGAPA-PAPIIT proyecto IA401517 y proyecto IA401419.

Referencias

Bailey, Richard W. 1979. Authorship attribution in a forensic setting. En *Advances in computer-aided literary and linguistic research*, 1–20. AMLC.

Burrows, John. 2002. Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3). 267–287. doi 10.1093/llc/17.3.267.

Coulthard, Malcolm. 1994. On the use of corpora in the analysis of forensic texts. *International Journal of Speech Language and the Law* 1(1). 27–43. doi 10.1558/ijsll.v1i1.27.

Diederich, Joachim, Jörg Kindermann, Edda Leopold & Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence* 19(1-2). 109–123. doi 10.1023/A:1023824908771.

Holmes, David I. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing* 13(3). 111–117. doi 10.1093/llc/13.3.111.

Juola, Patrick. 2009. JGAAP: a system for comparative evaluation of authorship attribution. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(1). doi 10.6082/M1N29V4Z.

Juola, Patrick. 2015. The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities* 30(suppl 1). i100–i113. 10.1093/llc/fqv040.

Juola, Patrick, John Sofko & Patrick Brennan. 2006. A prototype for authorship attribution studies. *Literary and Linguistic Computing* 21(2). 169–178. 10.1093/llc/fql019.

López-Escobedo, Fernanda, Julián Solorzano-Soto & Gerardo Sierra Martínez. 2016. Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution. *Journal of Quantitative Linguistics* 23(2). 154–176. 10.1080/09296174.2016.1142324.

Mendenhall, Thomas Corwin. 1887. The characteristic curves of composition. *Science* 9(214). 237–249.

Mosteller, Frederick & David Wallace. 1964. *Inference and disputed authorship: The federalist.* Addison-Wesley.

Nieto, Victoria Guillén, Chelo Vargas Sierra, María Pardiño Juan, Patricio Martínez Barco & Armando Suárez Cueto. 2008. Exploring state-of-the-art software for forensic authorship identification. *International Journal of English Studies* 8(1). 1–28.

Padró, Lluís & Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. En *Language Resources and Evaluation Conference (LREC)*, 2473–2479.

R Core Team. 2008. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Rudman, Joseph. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities* 31(4). 351–365. 10.1023/A:1001018624850.

Sierra, Gerardo, Julián Solórzano Soto & Arturo Curiel Díaz. 2017. GECO, un gestor de corpus colaborativo basado en web. *Linguamática* 9(2). 57–72. 10.21814/lm.9.2.256.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3). 538–556. 10.1002/asi.v60:3.

Stein, Sterling & Shlomo Argamon. 2006. A mathematical explanation of burrows's delta. En *Digital Humanities Conference*, 207–209.

Svartvik, Jan. 1968. *The evans statements: A case for forensic linguistics.* University of Gothenburg.

Tweedie, Fiona J., Sameer Singh & David I. Holmes. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities* 30(1). 1–10. 10.1007/BF00054024.