ÐLESPAÑOL
REVISTA DE LENGUA

Poisoned Texts

UN EJEMPLO DE DESARROLLO INFORMÁTICO PARA LA DETECCIÓN DEL PLAGIO Y LA CARACTERIZACIÓN DEL IDIOLECTO DISCURSIVOA

Poisoned Texts

AN EXAMPLE OF SOFTWARE DEVELOPMENT FOR PLAGIARISM DETECTION AND TEXTUAL IDIOLECT CHARACTERIZATION

Adrián Cabedo Nebot
Universitat de València
adrian.cabedo@uv.es
https://orcid.org/0000-0002-3881-9308

Resumen: Este artículo presenta una investigación de caso relacionada con dos áreas normalmente abordadas desde la Lingüística forense: la detección de plagio y la caracterización del estilo idiolectal. Desarrollado en el contexto universitario, el uso de un programa informático, *Poisoned texts*, sirve tanto para analizar el plagio como el estilo de un conjunto de 13 trabajos de estudiantes universitarios que son sospechosos de haber copiado un mismo texto fuente. Los resultados muestran que 6 estudiantes plagiaron un amplio número de largas secuencias textuales (n-gramas). En una segunda capa de análisis, el texto tomado como fuente manifiesta un estilo discursivo más completo y depurado que el utilizado por los estudiantes. En definitiva, este artículo, desde un estudio de caso real, expone una metodología de análisis del plagio y del idiolecto que combina tanto el apartado computacional como el interpretativo.

Palabras clave: Plagio académico, idiolecto, programa informático.

Abstract: This article presents a case study relating two areas of Forensic Linguistics: plagiarism detection and characterization of idiolectal style. Based on a real case study, a method of analysis of plagiarism and idiolect is presented that combines both the computational and interpretive approaches. To do so, the software Poisoned texts is used to detect and analyze the style and the eventual plagiarism in a set of 13 academic works delivered by university students who are suspected of having copied the same source text. The results on plagiarism and style reveal, respectively, (a) that 6 students plagiarized a significant number of textual sequences (n-grams), and (b) that the source text has a more complete and refined discursive style than the students' assignments.

Keywords: Academic plagiarism, idiolect, software.

Del Español: Revista de Lengua, 1, 2023, pp. 233-258
ISSN: 3020-2434, https://doi.org/10.33776/dlesp.v1.7922

Recibido: 11/10/2022
Aceptado: 12/12/2022

1. Introducción: el plagio como principio

El plagio constituye un ámbito de estudio muy general en nuestros días. Tradicionalmente, la Lingüística forense ha separado como dos ramas separadas la detección del plagio y la atribución de autoría, de tal manera que los programas informáticos que han sido desarrollados en los últimos tiempos se dirigen hacia una u otra según el objetivo concreto de análisis. Algunas propuestas, sin embargo, se presentan como soluciones integradoras; de este modo, un mismo acercamiento computacional permite al investigador realizar estudios tanto de plagio como de autoría. En esa línea, se sitúan programas como *Copycatch* (Woolls, 2010) o *Wcopyfind* (Bloomfield, 2016), que basan sus resultados en la cuantificación de frecuencias léxicas y en el porcentaje de similitud textual de caracteres y/o palabras.

Siguiendo la perspectiva de análisis establecida por estos programas, el objetivo de este artículo es presentar un caso de análisis práctico que sirva de ejemplo tanto para la detección del plagio entre textos como para una caracterización estilística básica; para poder realizar esa tarea, se parte del diseño y uso de un entorno computacional desarrollado con lenguaje de programación R, *Poisoned Texts* (Cabedo, 2022). Esta herramienta permite observar el grado de coincidencia textual entre documentos basándose en las concurrencias de n-gramas definidos previamente por el investigador; al mismo tiempo, los textos se han etiquetado gramaticalmente mediante el uso del paquete *UdPipe* (Wijffels, 2022); este etiquetado permite observar singularidades estilísticas como, por ejemplo, qué categorías gramaticales o combinación de categorías son más frecuentes al inicio de una oración.

No obstante, abordar el plagio y la caracterización discursiva del idiolecto es una tarea complicada si se tiene en cuenta la definición operativa de los conceptos teóricos que subyacen o a la metodología de trabajo que se aplica (como veremos en la sección 2 y la sección 3); no se trata únicamente de utilizar programas informáticos que realicen cómputos estadísticos, sino que el papel del investigador será crucial en la interpretación valorativa de estos.

Inicialmente, la definición de plagio es aparentemente muy básica: «at its simplest, plagiarism, or more accurately the type of plagiarism linguists are competent to deal with, is the theft, or unacknowledged use, of text created

by another» (Coulthard, Johnson, Kredens y Wools, 2010: 523). Según Cicres i Bosch y Gavaldà (2014: 68)[1]:

> El plagio consiste en la apropiación de ideas y palabras escritas por otra persona y, posteriormente, hacerlas pasar por propias. Así, podemos establecer dos tipos de plagio: el plagio de ideas, que se produce cuando una persona utiliza ideas de otra sin hacer ningún reconocimiento explícito, o bien el plagio lingüístico, que sucede cuando, además de copiar las ideas, se utilizan las mismas palabras o estructuras lingüísticas que el autor del texto original. Esta práctica constituye un delito contra la propiedad intelectual y, en litigios de este tipo, el perito lingüista puede ayudar a establecer el nivel de similitud textual entre dos o más textos y establecer la probabilidad de que no hayan sido producidos independientemente y, por tanto, que haya habido plagio lingüístico. [La traducción es nuestra].

Habitualmente, otros factores relacionados con la atribución de plagio deben tomarse con reserva: por ejemplo, Coulthard, Johnson, Kredens y Wools (2010) indican que la voluntariedad de quien plagia no puede tenerse en cuenta por el investigador, ya que sobrepasa sus competencias y capacidades, mientras que el plagio más habitual, el *verbatim* o la copia directa literal, solo consiste en una de las maneras en las que puede realizarse la copia.

Al mismo tiempo, el plagio se vincula a cuestiones de ámbito jurídico, más concretamente a situaciones en las que pueda existir un beneficio por parte de quien plagia, bien en forma de remuneración económica, bien en forma de consolidación profesional (plazas públicas, concursos...). Mientras la documentación jurídica[2] prohíbe expresamente la reproducción total o

[1] Texto original: «El plagi consisteix en l'apropiació d'idees i paraules escrites per una altra persona i fer-les passar per pròpies. Així, podem establir dos tipus de plagi: el plagi d'idees, que es produeix quan una persona utilitza idees d'una altra sense fer-ne cap reconeixement explícit, o bé el plagi lingüístic, que succeeix quan, a més de copiar les idees, utilitza les mateixes paraules o les mateixes estructures lingüístiques que l'autor del text original. Aquesta pràctica constitueix un delicte contra la propietat intel·lectual, i en litigis d'aquest tipus, el pèrit lingüista pot ajudar a establir el nivell de similitud textual entre dos o més textos i establir la probabilitat que no hagin estat produïts independentment, i per tant, que hi hagi hagut plagi lingüístic».

[2] Nos referimos concretamente al Artículo 270 del Código Penal (Ley Orgánica 10/1995, de 23 de Noviembre 1995), por el que «será castigado con la pena de prisión de seis meses a cuatro años y multa de doce a veinticuatro meses el que, con ánimo de obtener un beneficio económico directo o indirecto y en perjuicio de tercero, reproduzca, plagie, distribuya, comuni-

parcial de los textos que se comercializan, existen otros documentos jurídicos, basados en sentencias judiciales o en la *Ley sobre la propiedad intelectual*, que sí facilitan un cierto grado de coincidencia textual[3] y que también permiten la cita directa siempre que esta se establezca como tal y se haga una referencia a la fuente original[4]. Por otro lado, un factor clave que facilite la mencionada «copia en forma de cita» es el potencial permiso concedido por los autores originales, elemento que suele ser muy habitual en el entorno de la investigación científica.

Así mismo, existen ciertas zonas de vaguedad que no siempre se explicitan convenientemente; por ejemplo, en cuanto a las citas, no siempre se alude al porcentaje lícito de caracteres o palabras que pueden ser citados en un mismo bloque. Estas citas, por tanto, pueden comprender desde una o varias líneas hasta varias páginas.

En general, por tanto, la detección de plagio consiste en averiguar el porcentaje de similitud entre dos textos y determinar si ese porcentaje y la manera en la que se ha cometido la copia puedan ser motivo de delito y, por tanto, de denuncia. Mientras que la copia directa es más fácilmente detectable mediante programas informáticos (Barrón Cedeño, 2014; Kiss, 2013; Sun, 2013; Woolls, 2010), otro tipo de copias requiere de la participación activa de los investigadores y entra más en el ámbito de la estilística textual y la llamada atribución de autoría (McMenamin, 1993).

---

que públicamente o de cualquier otro modo explote económicamente, en todo o en parte, una obra o prestación literaria, artística o científica, o su transformación, interpretación o ejecución artística fijada en cualquier tipo de soporte o comunicada a través de cualquier medio, sin la autorización de los titulares de los correspondientes derechos de propiedad intelectual o de sus cesionarios».

[3] La *Sentencia número 1204/2008, Sala 1ª, de lo Civil, 18 de diciembre de 2008* (STS 1204/2008, 18 de diciembre de 2008, 2009), contempla que «el sentido general del plagio se centra en la copia sustancial, como actividad material mecanizada y poco intelectual y menos creativa, carente de toda originalidad [...] y no constituye plagio cuando son dos obras distintas y diferenciables aunque tengan puntos comunes de exposición».

[4] El Artículo 32 de la *Ley de Propiedad Intelectual* («Ley de propiedad intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia (*Real Decreto Legislativo* 1/1996, de 12 de Abril)» 1996) indica que «es lícita la inclusión en una obra propia de fragmentos de otras ajenas de naturaleza escrita, sonora o audiovisual, así como la de obras aisladas de carácter plástico o fotográfico figurativo, siempre que se trate de obras ya divulgadas y su inclusión se realice a título de cita o para su análisis, comentario o juicio crítico. Tal utilización solo podrá realizarse con fines docentes o de investigación, en la medida justificada por el fin de esa incorporación e indicando la fuente y el nombre del autor de la obra utilizada».

2. La intertextualidad en el marco académico

En la actualidad, existe un enorme interés por estudiar el plagio en el ámbito académico. De hecho, son innumerables las referencias hacia el plagio académico y su tipología general en el ámbito de la Educación Secundaria (Morey López , Sureda Negre, Oliver Trobat y Comas Forgas, 2013; Díaz Arce, 2016) o de la educación universitaria (Alfaro Torres y Juan Juárez, 2014; Anson, 2022; Cayuela Mateo, Tauste Francés, Seguí Crespo, Esteve Faubel y Ronda Pérez, 2015; Eret y Gokmenoglu, 2010); pero también hacia el perfil del estudiante plagiario (Law, Ting y Jerome, 2013; Ochoa y Cueva, 2016; Sureda, Comas y Morey, 2009).

El plagio, aunque resulte llamativo, no siempre es reconocido como una actividad punible por parte del estudiante. Concretamente en la Educación universitaria, un estudio centrado en 305 estudiantes de las titulaciones de Óptica, Magisterio y Relaciones Laborales (Ronda Pérez, Seguí Crespo, Cayuela Mateo, Tauste Francés, Lumbreras Lacarra y Esteve Faubel, 2016) detecta, a partir de un cuestionario de actitudes y creencias, que la mayor parte de los alumnos desconocen qué es el plagio y consideran inevitable su uso. No obstante, la formación previa al alumnado sobre el plagio consigue que exista un 30 % menos de probabilidades de que se cometa.

Al mismo tiempo, los argumentos que justifican la copia en el marco académico son múltiples y variados: inexperiencia, comodidad, pereza, búsqueda de precisión, hábito, desconocimiento de las normas de redacción, búsqueda de mejores notas... (Sureda Negre, Comas Forgas y Morey López, 2009; Sureda Nergre y Comas Forgas, 2006). Como ejemplo bastante representativo, en el año 2000, se realizó una encuesta a 1189 estudiantes de la Escuela de Informática de la universidad australiana de Monash (Sheard, Markham y Dick, 2003). En él se descubrieron distintas motivaciones, entre las que sobresalían, además de las ya mencionadas, otras como la falta de tiempo, la dificultad de una asignatura o un trabajo, pérdida de clases por enfermedad, ayudar a un amigo, presión familiar, deseo de no fallar (o, curiosamente, la indiferencia ante el fallo), etc.

Como comentábamos en la introducción (sección 1), no queda claro el porcentaje de similitud entre dos textos que pueda considerarse plagio y, por lo tanto, en el ámbito docente, suele ser el profesor el que decide qué límite es el máximo de coincidencia aceptable para poder aprobar o superar una tarea. En general, una copia superior al 20 % o 30 %, indica una similitud muy am-

plia que difícilmente puede explicarse por citación directa. En casos en los que hubiera una similitud de 20 % a 50 %, incluso con una cita adecuada mediante comillas o sangría y reducción de tamaño de letra, depende siempre del profesor establecer qué parte ha sido redactada, o al menos parafraseada, y qué parte es simplemente una transliteración. Evidentemente, el esfuerzo de transliterar o parafrasear no suele ser el mismo, por lo que ambas acciones se evalúan normalmente con evaluaciones diferentes, es decir, con una sanción para la primera y una mejor consideración para la segunda.

3. Sobre el estilo discursivo idiolectal

Una de las primeras definiciones creadas sobre el concepto de *idiolecto* es la desarrollada por Bloch (1948: 7), para quien el concepto consiste en:

> The totality of possible utterances of one speaker at one time in using a language to interact with one other speaker is an idiolect… Our definition implies (a) that an idiolect is peculiar to one speaker, (b) that a given speaker may have different idiolects at successive stages of his career, and (c) that he may have two or more different idiolects at the same time. (Bloch, 1948: 7)

En cuanto al objeto de análisis lingüístico, se podría considerar que, en la práctica, la selección de los métodos para identificar el idiolecto o el estilo discursivo depende principalmente de dos factores: la tarea que se desarrolla y la metodología del experto, en la medida en que este puede determinar las variables de análisis a partir de las características del texto concreto o, de distinto modo, puede realizar el análisis partiendo de unas características (marcas) establecidas previamente (Grant, 2010: 521). En la actualidad, este análisis combina el conocimiento lingüístico del experto con el uso de software.

Las tareas de atribución de autoría y de identificación de estilo discursivo consisten en analizar y/o comparar conjuntos de textos escritos u orales (o varios, dependiendo de las necesidades de la investigación policial, judicial o académica), con el objetivo de analizar si los grupos de textos comparten una serie de características lingüísticas distintivas; se analiza cuáles son las características lingüísticas individualizadoras o idiolectales de cada grupo de textos y si tales características se reproducen en los grupos de textos que se comparan. Al realizar esta actividad analítica, siempre a través de la lupa valorativa del investigador, se delimita un potencial perfil discursivo del autor

y, por tanto, de su idiolecto. Grant (2010) desarrolla una visión del idiolecto que integra concepciones teóricas cognitivas y estilísticas para servir de base a una atribución de autoría:

> My exposure to a certain variety of language containing one set of collocates would be different from my neighbour's and this personalisation would gradually cause individual differences in our language production. Idiolectal consistency and variation would draw on the resource of my cognitive capacity for language production and also draw on the complexity of my personal sociolinguistic history. According to this potential theory of idiolect, the cognitive capacity is itself structured but malleable and the sociolinguistic history is realised in incremental changes to that neuro-cognitive capacity (Grant, 2010: 514).

En la medida en que el uso de un rasgo lingüístico puede caracterizar las producciones de un número de hablantes indefinido, resulta clave la percepción de esas características habituales como conjunto: «In forensic authorship attribution, the linguist's task is to identify the writer's habitual choices and define them as a set» (McMenamin, 1993: 158).

Así pues, como puede observarse, tanto la detección de plagio como la identificación del estilo de autor son actividades emparentadas y el investigador puede cruzar constantemente entre ambas áreas sin ser plenamente consciente de ello. Es por ello por lo que los métodos informáticos, convenientemente trazados, deben facilitar esa transición y facilitar el enfoque concreto del analista.

4. Métodos informáticos para la detección del plagio y la definición del estilo discursivo

En cuanto a los programas informáticos diseñados para estudiar el plagio o caracterizar el estilo discursivo, la mayor parte se basan en la búsqueda de coincidencias formales entre los textos, normalmente en forma de caracteres, palabras o n-gramas (Wright, 2017). En cualquier caso, estos programas presentan tanto fortalezas como debilidades:

> The strengths are clearly the size of data and the speed of processing of the texts, without mental fatigue and with consistent application of the rules. The weaknesses come from the complexity of the

> concept of similarity and the fact that any computer program can only be an approximation of what human readers can recognise and handle with ease (Woolls, 2010: 590).

Algunos programas como *Antconc* (Anthony, 2022) han sido diseñados para observar concordancias y frecuencias entre textos sin una voluntad forense, pero pueden utilizarse para este fin con una debida adaptación de los datos por parte del investigador. Sí existen, sin embargo, programas dirigidos al procesamiento y la búsqueda del plagio, como *Copycatch* (Woolls, 2010) o *Wcopyfind* (Bloomfield, 2016). En el caso de *Copycatch*, se trata de un programa no gratuito, desarrollado hace más de veinte años por David Wools, mientras que *Wcopyfind*, de uso gratuito y desarrollado por Lou Bloomfield, no se actualiza desde el año 2016. En ambos casos, los programas observan concordancias entre textos y generan un porcentaje de coincidencia basado en frecuencias léxicas, es decir, en la cantidad de palabras que se comparten.

En este sentido, es importante la inclusión de factores como los *hapax legomena,* secuencias que ocurren una única vez en un texto. Si hay una coincidencia de una o varias palabras que solo se repitan una vez entre dos textos, se considera un elemento valioso en la determinación del plagio, ya que es una coincidencia difícilmente explicable por el azar (Grant, 2010; Woolls, 2010). En el caso de *Copycatch*, además, puede utilizarse una lista de palabras que, por la naturaleza de los textos, se determine que no son susceptibles de ser valoradas en la coincidencia entre textos. Por ejemplo, en un trabajo sobre dinosaurios del Cretácico es posible que palabras como *años, miles, periodo, Cretácico, dinosaurio, época*... aparezcan de forma ineludible; lo mismo sucedería con algunas categorías gramaticales, como los determinantes o las preposiciones.

5. Metodología

Con ánimo de exponer un caso de identificación del plagio y de caracterización de los rasgos estilísticos de uno o varios autores, exponemos en este artículo el estudio de un conjunto de textos procedentes de una prueba universitaria. La metodología elegida en esta investigación, y que coincide con lo señalado por la bibliografía (Lukashenko, Graudina y Grundspeņķis 2007; Grant 2010; Cicres i Bosch y Gavaldà, 2014), utiliza la gestión computacional como recurso de apoyo para el análisis realizado por el especialista. Al fin y

al cabo, más allá de lo extenso o pormenorizado del estudio cuantitativo, será siempre la interpretación del investigador la que determine en qué medida dos o más textos son similares u originales y, también, cuáles son los rasgos más característicos del estilo discursivo o idiolectal de los autores

5.1 Corpus de análisis

El corpus de textos seleccionado para el análisis procede de un entorno docente real. En el curso académico 2021-2022, como prueba final de la asignatura Dialectología y Sociolingüística Españolas, se indicó a los estudiantes que debían presentar la redacción de un texto de 1000 a 2000 palabras en el que se ampliara y comentara un párrafo[5] de un artículo de Fernández Ordóñez (2016), titulado «Dialectos del español peninsular». Para esta investigación, se seleccionaron aleatoriamente 13 de esas pruebas escritas; su distribución queda reflejada en la **Tabla 1**, donde el texto «dialectos.txt» hace referencia al artículo de Fernández Ordóñez (2016) y el resto, nombrado con letras del alfabeto, a cada uno de los alumnos.

[5] El párrafo en cuestión era el siguiente: «El español hablado en la Península Ibérica está articulado en una doble división dialectal: por un lado, en áreas dialectales que se disponen de norte a sur y que parecen relacionarse con el proceso de ocupación del territorio durante la Edad Media, tal como la disposición vertical norte-sur de las lenguas romances peninsulares [...]. Con frecuencia los varios rasgos lingüísticos que caracterizan a estas áreas no coinciden en sus límites geográficos ni en su distribución social, de forma que se plantea inevitablemente el problema de decidir qué rasgos deben prevalecer en la caracterización...» (Fernández Ordóñez, 2016: 387).

texto	párrafos	oraciones	tokens
A.txt	15	85	1948
B.txt	17	48	1209
C.txt	17	48	1209
D.txt	9	56	1019
dialectos.txt	96	546	8566
E.txt	25	119	2018
F.txt	16	60	1492
G.txt	16	63	1812
H.txt	9	45	1098
I.txt	16	63	1812
J.txt	22	60	1274
K.txt	14	44	900
L.txt	6	56	1459
M.txt	7	54	1468

Tabla 1. Resumen de los textos sometidos a análisis

Llaman la atención distribuciones como las de L o M, en las que hay 6 y 7 párrafos respectivamente y, sin embargo, la cantidad de párrafos es más o menos comparable con la de otros textos que tienen más párrafos e, incluso, con aquellos que tienen más párrafos con menos oraciones, como B o C. El texto con menos palabras, K (900), contrasta con el que más palabras presenta, E (2018).

Tanto los 13 archivos como el artículo de Fernández Ordóñez se cotejaron unos con otros. Con este procedimiento, los textos no se comparan solo con el sospechoso de haber sido copiado (*dialectos.txt*, en este caso), sino también entre ellos; de este modo, no solo puede observarse si algunos estudiantes puedan haber compartido la redacción de la prueba, sino que posiblemente han tomado como fuente de copia una misma obra de consulta.

5.2. Poisoned texts: una herramienta informática de apoyo

En el epígrafe de este apartado se usa una palabra clave, como es «apoyo», es decir, los recursos de computación, los programas informáticos en general, son en la mayor parte de las veces herramientas de soporte o apoyo que facilitan la labor de investigación de los especialistas. En el caso del plagio o de la atribución de autoría, los programas proporcionan valores clave, como las frecuencias léxicas, el tipo de categorías gramaticales utilizadas, las posiciones de las palabras, las colocaciones... Todos ellos son recursos que sirven a los especialistas para mejorar sus informes lingüísticos, pero el valor del investigador será siempre fundamental y jerárquicamente más importante que los resultados arrojados por cualquier entorno computacional. Como señalan Lukashenko, Graudina y Grundspenkis (2007: 5):

> Analysis of the known plagiarism detection tools shows that although these tools provide excellent service in detecting matching text between documents, even advanced plagiarism detection software can't detect plagiarism so good as human does. They have several drawbacks and, so manual checking and human judgment is still needed. Human brain is universal plagiarism detection tool, which is able to analyze document using statistical and semantical methods, is able to operate with textual and non-textual information (Lukashenko, Graudina y Grundspeņķis 2007: 5)

En la actualidad, acercarse al estudio cuantitativo de la similitud textual puede realizarse con programas especializados (previamente, en la sección 4, hemos mencionado *Copycatch* o *Wcopyfind*), pero también con programas dirigidos al tratamiento textual más genérico (*Antconc*). Estos programas tienen a veces algunos problemas, normalmente relacionados con el sistema operativo en el que se ejecutan, que pueden dificultar su uso: *Copycatch*, por ejemplo, es una herramienta desarrollada en lenguaje de programación *Java* (por tanto, de uso general en cualquier sistema operativo), pero es un programa comercial; *Wcopyfind*, por su parte, es gratuito, pero solo funciona en sistema operativo *Windows* y, a fecha de realización de este artículo, no ha sido actualizado desde 2016. Finalmente, *Antconc* es la solución más operativa, desarrollada en *Java*, actualizado regularmente y gratuito, pero no enfocado precisamente a la detección de plagio o a la creación de informes de similitud lingüística o de descripción de estilo.

Es en este contexto en el que hemos considerado útil crear un programa en lenguaje de programación R (R Core Team, 2022) y en un entorno de ejecución *Shiny* (Chang y otros 2021) que es gratuito, multiplataforma y completamente ampliable o extensible con un mínimo de conocimiento de R (Cabedo. 2021). Actualmente, este programa, llamado *Poisoned Texts* (Cabedo, 2022), está disponible en https://github.com/acabedo/poisoned-texts y puede ejecutarse directamente mediante una versión demo o descargarse e instalarse en el propio ordenador. Aunque no se almacenan los textos que se procesan en la versión demo del navegador, podría darse el caso de que el especialista no se sintiera cómodo subiendo los textos de análisis a un entorno web, así que el programa puede descargarse en el sistema operativo nativo del investigador y ser ejecutado en modo offline, sin ningún tipo de filtración a la web. La interfaz general del programa puede verse en la **Figura 1**:

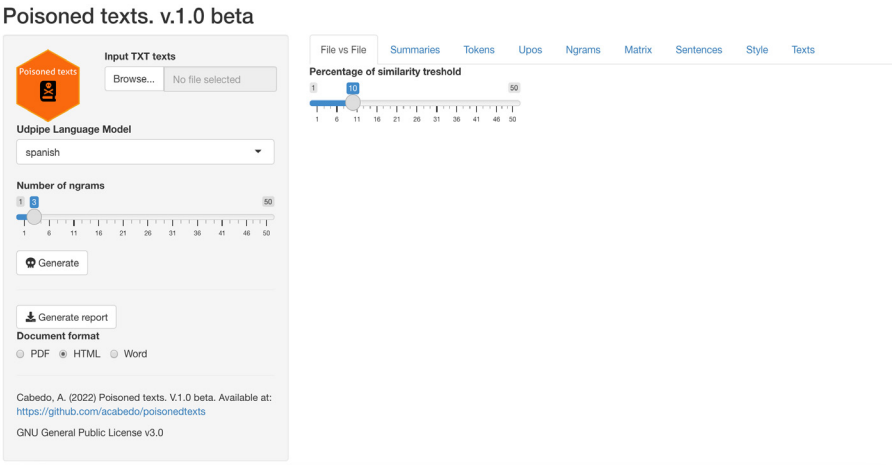

Figura 1. Imagen general del programa Poisoned Texts.

R es actualmente uno de los lenguajes de programación más utilizados en la comunidad científica a nivel internacional. En primer lugar, se trata de un lenguaje más accesible para investigadores que no necesariamente procedan del sector de la Informática; en segundo lugar, los programas desarrollados con R son fácilmente escalables, por lo que siempre pueden añadirse nuevas funcionalidades sin necesidad de modificar en gran medida el código fuente del programa. Finalmente, un último factor no menos importante es que precisamente la gran comunidad de usuarios que existe se preocupa cons-

tantemente de mejorar y potenciar, mediante librerías, las funcionalidades, ya de por sí ingentes, del propio lenguaje.

En general, la codificación de programas por parte de los investigadores tiene la opción de estar al servicio de los intereses particulares de quien los diseña, pero también pueden ofrecerse de modo gratuito a la comunidad científica, tanto para que esta los utilice sin más, como para que los descargue y modifique a partir del código fuente. En el caso de *Poisoned Texts,* el programa puede descargarse desde la plataforma *Github*, cuyo objetivo es la de permitir a los desarrolladores, y a otros investigadores, compartir programas creados normalmente con código de libre acceso.

El enfoque de *Poisoned Texts* parte en cierta medida de otros programas citados previamente como *Copycatch* o *Wcopyfind*, pero presenta otras novedades que desarrollamos en la siguiente lista:

- Uso del etiquetado gramatical de las palabras, realizado automáticamente mediante el uso de la librería *UdPipe*, en el que se incluyen categorías gramaticales tradicionales (sustantivo, adjetivo, adverbio...), pero también números o puntuación.
- Frecuencias léxicas absolutas y relativizadas de los n-gramas cuyo tamaño ha sido seleccionado por el analista (con un máximo de 50 palabras consecutivas).
- Para aproximarse al estilo del autor, los textos pueden caracterizarse a partir de una tabla en el que se calcula la media de la cantidad de palabras, sustantivos, adjetivos, comas... que aparecen por oración en los textos analizados. Estas frecuencias se relativizan además sobre una base de 100 palabras.
- Los datos anteriores se proyectan para mejorar su visibilidad en un mapa de calor en el que los valores se han relativizado por columna.
- Los n-gramas se etiquetan, de manera similar a como hace *Copycatch*, en aquellos que aparecen más de una vez (*shared*), los que se aparecen una única vez y coinciden una sola vez con otro o más textos (*once*) y los que aparecen solo en un único texto, sin coincidir con ningún otro texto (*only*).
- Es posible exportar un informe general a html, Word o PDF, en el que se añade más información, como las palabras más frecuentes y n-gramas de 2, 3 y 4 palabras que aparecen a principio de oración.

- En el informe que se genera puede observarse también las categorías gramaticales más frecuentes, pero también las combinaciones de 2, 3 y 4 categorías más frecuentes que aparecen a principio de oración. De este modo, siempre pueden observarse características particulares del idiolecto o estilo de los sujetos en un ámbito más profundo que el meramente superficial de la coincidencia formal de los tokens o formas léxicas en general. Imagínese, por ejemplo, que una secuencia DET NOUN VERB ADJ (por ejemplo, subyacente a oraciones como *el hombre viene ligero*) fuera la más frecuentes a inicio de oraciones y coincidiera solo entre dos textos de una comparación realizada entre 30 textos. No tendría que concluirse necesariamente que esos dos textos pertenecieran al mismo autor, pero sí podría ser un argumento de apoyo a un análisis más amplio y detallado.

Al haber sido diseñado con R, *Poisoned Texts* puede ampliarse con nuevas funciones para la explotación estadística, bien descriptiva (gráficos de frecuencias, diagramas de caja…), bien inferencial (árboles de decisiones, regresiones, análisis factorial, análisis discriminante…).

6. Análisis de caso

A partir del uso de *Poisoned Texts* hemos podido realizar dos tipos de análisis: el primero, centrado en la similitud textual entre los textos analizados; el segundo, por su parte, nos ha permitido explorar algunas características discursivas de los autores de los textos, con la voluntad de observar cuáles se asemejan, más allá de una mera coincidencia de secuencias textuales. De esta manera, con este estudio de caso, ejemplificamos tanto la detección de plagio como caracterización de estilo o idiolecto.

6.1 Porcentaje de plagio

En la **Figura 2** se expone el porcentaje de similitud entre los textos que más coincidencia han presentado en el análisis del programa; en este caso, los n-gramas tomados en consideración han sido de tamaño 10, dado que la bibliografía ha establecido que, en general, secuencias que superen 8 o 9

palabras consecutivas deberían ser idiolectales; esto lo sugiere, por ejemplo, Grant (Grant 2010), que toma como base un experimento previo realizado en Google por Malcolm Coulthard.

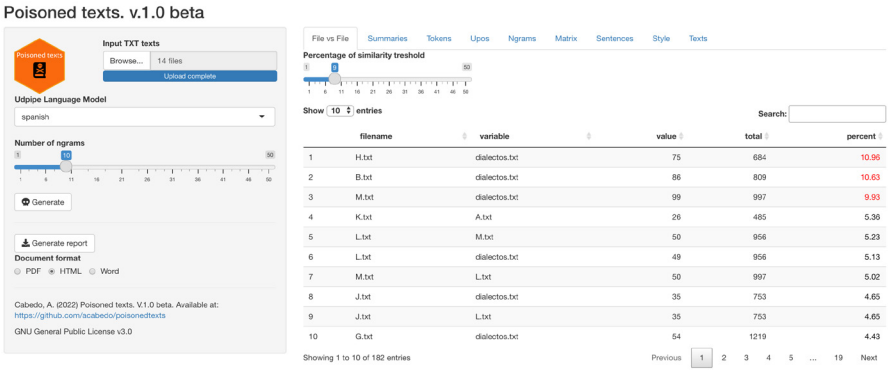

Figura 2. Porcentajes de similitud más altos en n-gramas de tamaño 10 en los textos del corpus.

En la **Figura 2**, por tanto, se observa que hay tres textos (H, B y M) que presentan un 10 % de coincidencia de n-gramas tamaño 10 con el archivo *dialectos.txt*, que incluye el texto del artículo de Fernández Ordóñez (2016). Por ejemplo, en el caso del estudiante H, se observan 75 secuencias de n-gramas tamaño 10 que coinciden con *dialectos.txt*. Si ese tamaño de n-gramas se reduce a dos, o incluso a uno, el porcentaje de copia asciende necesariamente; de hecho, aunque no aportamos imagen de ello, las coincidencias ascienden en ese último caso a 42-44 % en los mismos textos.

Por lo tanto, en este caso el resultado no solo establece un porcentaje de similitud como elemento que señale el plagio del estudiante (de hecho, un 10 % puede no parecer demasiado), sino que es precisamente el tipo de similitud el que señala en la dirección de una copia *verbatim* o literal con secuencias que, como señalábamos con el ejemplo de Coulthard, supera el umbral habitual en construcciones que deberían ser propias de los estudiantes.

Si ahondamos en esas secuencias de 10 palabras consecutivas, podemos ver en la Figura 3 una matriz de coincidencias, en la que aparece una columna con los distintos n-gramas de 10 detectados en los textos analizados y, extendidos a la derecha, la cantidad de esos n-gramas por texto.

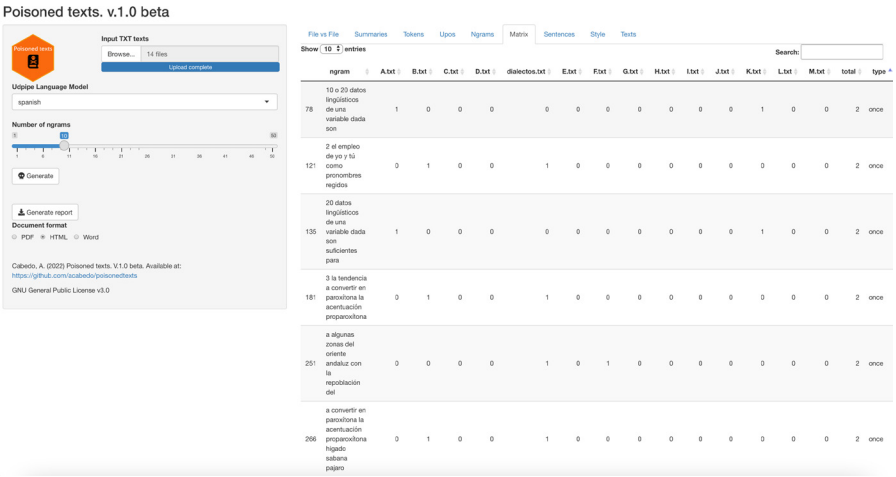

Figura 3. Ejemplo de coincidencia de n-gramas de 10 entre los textos.

En el caso de la **Figura 3**, por tanto, hemos seleccionado aquellos n-gramas que coinciden una única vez entre los textos (*once*), es decir, que aparecen una vez, por lo menos, en uno de los textos y se repiten en otro. Este fenómeno, denominado *hapax legomena*, se utiliza también como criterio habitual en la determinación de plagio (Woolls 2010). Por ejemplo, el n-grama *a algunas zonas del oriente andaluz con la repoblación del* aparece una única vez en el texto *dialectos.txt* y en el texto del estudiante F.

Todos los n-gramas se resumen también visualmente como se muestra en la **Figura 4**, en la que, para cada texto, se indica qué distribución por texto se determina para el conjunto de n-gramas que se repiten más de una vez (*shared*), los que aparecen una sola vez y se repiten con otro texto (*once*) y los que son exclusivos de cada texto (*only*).

Así pues, la **Figura 4**, muestra las distribuciones de los textos que, en principio, no han copiado tanto secuencias de n-gramas tamaño 10, como sería el caso de los estudiantes G e I; por su parte, el estudiante H, aunque presenta un 88.89 % de contenido presuntamente original (o quizá mejor parafraseado), sí incluye el resto de su texto como copia en forma de *hapax legomena* o recurrente.

	ngram	type	freq	total
21	F.txt	shared	6	0.64
22	G.txt	once	49	4.02
23	G.txt	only	1153	94.59
24	G.txt	shared	17	1.39
25	H.txt	once	66	9.65
26	H.txt	only	608	88.89
27	H.txt	shared	10	1.46
28	I.txt	once	11	1.11
29	I.txt	only	977	98.59
30	I.txt	shared	3	0.3

Showing 21 to 30 of 42 entries Previous 1 2 3 4 5 Next

Figura 4. Ejemplo de vocabulario por frecuencia de aparición en alguno de los textos analizados.

Por su parte, en la **Figura 5**, a modo de visualización más atractiva, los distintos n-gramas de 10 que han sido identificados como coincidentes, se proyectan para poder ser utilizados en una posible justificación ante el alumno por haber recibido este algún tipo de sanción o reducción de nota en la evaluación final.

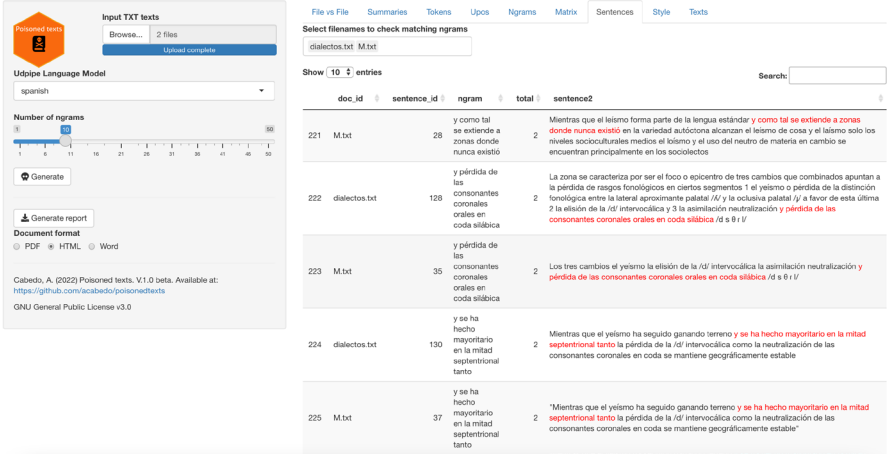

Figura 5. Ejemplo de concordancias coincidentes entre textos.

Concretamente, en la **Figura 5** se proyectan las coincidencias del estudiante M con el artículo de Fernández Ordóñez (2016). Podrían haberse seleccio-

nado más textos o incluso todos ellos, pero normalmente la funcionalidad de este apartado de *Poisoned Texts* está pensada para ser ejecutada en grupos de dos. Debe tenerse en cuenta que esta selección es una de las partes que, si se deciden exportar los resultados, aparecerá en el archivo de exportación final, bien sea en *PDF*, *HTML* o *Word*.

Así pues, en general, puede decirse que los estudiantes que mayor plagio han mostrado han sido H, B y M; por su parte, A, I y K también presentan coincidencias entre ellos. La revisión general de los textos muestra que no se han copiado realmente entre sí, sino que han tomado como base el texto de Fernández Ordóñez (2016), si bien en algunas ocasiones las labores de parafraseo han sido bastante compartidas. Finalmente, el resto de estudiantes como C, D, E, F, J y L presentan porcentajes de copia, al menos de n-gramas tamaño 10, menores e incluso residuales.

6.2 Estilos discursivos

En la sección anterior (6.1.) se observaba la similitud textual entre los documentos analizados. En el caso del presente apartado, más dirigido hacia el estudio del estilo discursivo, *Poisoned Texts* realiza una caracterización gramatical básica de los textos, tomando como base el etiquetado en *part of speech tags* realizado con *UdPipe* (Wijffels, 2022). El resumen de la media de categorías gramaticales y otras variantes por oración, con una relativización sobre 100 palabras, se puede observar en la **Figura 6**.

En el caso de esta funcionalidad, el investigador puede seleccionar las variables que quiere que aparezcan en el mapa de calor, con lo que este mapa puede ser más o menos amplio según los intereses particulares. Al mismo tiempo, como se señalaba en el apartado 5.2, *Poisoned Texts* puede ampliar sus variables de análisis con sencillas modificaciones del código fuente.

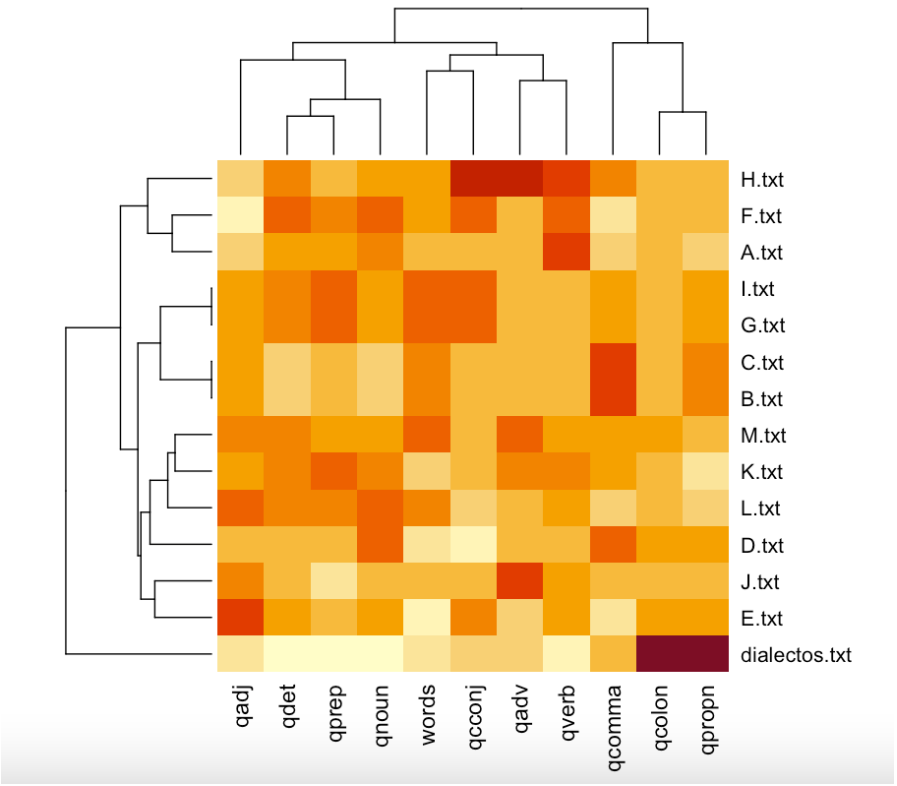

Figura 6. Mapa de calor con dendograma (análisis de clúster) que une en grupos a los textos analizados

La **Figura 6** incluye mucha información tanto en la disposición y el orden vertical en el que aparecen los textos, como en el color de las diferentes variables o el análisis de clúster de la parte izquierda del mapa. El análisis de clúster horizontal acumula las variables que presentan una mayor cercanía, como, por ejemplo, es el caso de la cantidad de adjetivos, determinantes, preposiciones y nombres que aparecen de media por oración. En cuanto a los colores, las tonalidades más oscuras (cercanos al rojo o marrón oscuro) indican que los documentos presentan un mayor valor de las variables correspondientes. Los colores, en realidad, representan las medias de las variables, que en *Poisoned texts*, aparecen en una tabla. A modo de ejemplificación, se incluyen en la **Figura 7**:

ISSN: 3020-2434, https://doi.org/10.33776/dlesp.v1.7922

	doc_id	words	qnoun	qadj	qverb	qdet	qpropn	qadv	qprep	qcconj	qcomma	qcolon
1	A.txt	26.1	20.9	7.2	10.3	14.5	1.7	2.9	12.9	3	4.2	0.3
2	B.txt	30.3	17.7	8.2	6.9	11.6	5.6	2.8	11.9	3	7.4	0.2
3	C.txt	30.3	17.7	8.2	6.9	11.6	5.6	2.8	11.9	3	7.4	0.2
4	D.txt	21.9	22	7.9	6.9	12.7	4.5	2.8	12.1	2.1	6.7	0.6
5	dialectos.txt	20.6	14.5	5.9	4.7	6.9	11.4	2.2	7.1	2.7	5.1	3.4
6	E.txt	19.8	20.1	10.6	7.5	13.9	4.6	2.1	11.3	3.5	3.5	0.7
7	F.txt	28	21.8	5.5	9	16.9	2.5	3.1	13.9	3.8	3.4	0.1
8	G.txt	32.7	20	8.6	7.1	15.6	3.9	3	15.4	3.8	5.8	0
9	H.txt	27.9	19.5	7.1	10.1	15.2	3.1	5.9	12.1	4.4	6.3	0
10	I.txt	32.7	20	8.6	7.1	15.6	3.9	3	15.4	3.8	5.8	0
11	J.txt	25.2	18.6	9.2	8.1	12.5	3.2	5.1	9.9	3.1	4.7	0.1
12	K.txt	23.9	20.9	8.2	8.3	15.4	0.8	4.4	14.7	2.9	5.7	0
13	L.txt	29.8	21.4	10	7.6	15.8	2	3	14.2	2.7	3.9	0.2
14	M.txt	31.3	19.5	8.7	7.5	15.4	2.4	4.9	13.1	3	5.6	0.5

Figura 7. Valores medio por oración de las variables introducidas en el mapa de calor.

Por ejemplo, a partir de lo que se proyecta en la Figura 6 y en la Figura 7, el documento de Fernández Ordóñez (2016) se desliga del resto de los documentos en la parte final del mapa y se consolida como una rama única en el análisis de clúster de la parte izquierda. Si atendemos a las variables expuestas sobre el mapa, Fernández Ordóñez utiliza un mayor número de punto y comas (*qcolon [semicolon quantity]*) y de nombres propios (*qpropn [proper noun quantity]*). El uso del punto y coma denota un mayor cuidado por la expresión y la presentación textual, por lo que el registro formal del texto científico queda manifiesto; por su parte, los nombres propios están relacionados con gran citación de ciudades o de nombres relacionados con lugares o pertenencias geográficas (Barcelona, Madrid, Asturias, Vasco, Catalán...); también debe tenerse en cuenta que el artículo de Fernández Ordóñez (2016) es mucho más extenso que el resto de textos comparados y que, además, presenta un apartado de bibliografía en el que mayoritariamente abundan los nombres propios.

Al mismo tiempo, las variables que presentan valores más bajos son la cantidad de determinantes (*qdet*), preposiciones (*qprep*), verbos (*qverb*) y nombres comunes (*qnoun*); pero esto último tiene también sentido si se tiene en cuenta que la cantidad de palabras por oración es menor que en muchos de los textos de los estudiantes. La variable *words* manifiesta la media

de palabras por oración de Fernández Ordóñez (2016); así pues, el estilo de la autora es claramente mucho más preciso y cuidado, ya que no solo utiliza oraciones breves, las cuales mejoran la recepción por parte del lector en el marcado de aprendizaje académico, sino que combina este dato con el uso de punto y coma, signo de puntuación que, como hemos comentado, dosifica la información y estructura mejor el contenido de lo dicho.

Por su parte, el resto de las variables que caracterizan el estilo son la cantidad de comas (*qcomma*), la de conjunciones no subordinadas (*qcconj*) y la de adverbios (*qadv*), que tienen algo más de presencia que otras categorías gramaticales.

En cuanto al resto de estudiantes, los textos se agrupan de la siguiente forma:

- H, F y A. Tienen extensión similar de oraciones, con unos valores medio altos, y usan un gran número de verbos, nombres y conjunciones coordinantes, así como pocos adjetivos.
- I, G, C y B. Presentan una extensión mayor de oraciones, con pocos adverbios y verbos.
- M, K, L y D. Usan menos verbos y conjunción coordinantes, pero más determinantes y nombres. La extensión de las oraciones, en general, suele ser breve en este grupo.
- J y E. Estos dos textos presentan un estilo más compensado, con una distribución más o menos similar por cada una de las categorías gramaticales incluidas, además del uso de punto y coma, más usado en el caso de E. En todo caso, destaca el uso de adjetivos y de verbos.

Si recordamos los textos que presentaban una mayor cantidad de plagio (H, B y M), vemos que quedan repartidos en tres grupos distintos si atendemos al estilo discursivo. De este modo, mientras que el plagio puede ser más o menos constante en un estudio de similitud textual, la consideración de las características lingüísticas de los autores, desde un punto de vista intratextual, puede conllevar la creación de otros grupos diferentes.

Esto no debe llevar a confusión, dado que los presupuestos del plagio y de la caracterización de estilo no son exactamente los mismos. En los textos académicos de los estudiantes que hemos analizado hay marcas propias, más allá de las secuencias que pueden coincidir con el texto de Fernández Ordóñez (2016). Debemos recordar que, atendiendo a n-gramas de tamaño

10, las coincidencias eran de un 10 %, por lo que sigue existiendo un 90 % textual en el que los estudiantes utilizan sus propias palabras o, al menos, su manera idiosincrásica de expresarse; incluso en los casos en los que haya existido un parafraseo también directo del texto tomado como fuente.

7. Reflexiones finales

En este artículo hemos pretendido situar la detección del plagio académico en un entorno de análisis real, procedente del estudio de trece trabajos finales de una asignatura; algunos de estos trabajos copiaron en mayor medida un trabajo fuente sobre dialectología del español de Fernández Ordóñez (2016).

Precisamente, el utilizar un análisis de caso real nos ha permitido introducir brevemente, como marco teórico, elementos importantes en la identificación y determinación del plagio, como la legalidad de la acción de copia, su tipología, el valor que le conceden los estudiantes... Al mismo tiempo, hemos visto que la detección del plagio y la caracterización del estilo discursivo de los autores son actividades que pueden quedar integradas en un mismo bloque de análisis.

Por ello mismo, cuando un caso real de estudio requiere de la interpretación por parte del analista (o profesor, generalmente, en el ambiente académico), los problemas se articulan sobre todo en la medida en la que el análisis valorativo se realiza de un modo operativo, funcional y ágil. Es en este sentido en el que el entorno computacional cobra sentido, ya que las herramientas informáticas contribuyen en enorme medida a simplificar la tarea de análisis, siempre que ofrezcan una gestión visualmente simple, pero completa, del apartado cuantitativo.

Por esta razón, en este artículo introducimos el uso de un programa informático, *Poisoned Texts*, desarrollado con lenguaje de programación R, que permite distintas operaciones de análisis textual, en las que destacan no solo la búsqueda de intertextualidad a partir de la coincidencia de n-gramas de diferente tamaño, sino el etiquetado gramatical automático, la caracterización discursiva del estilo de los autores, basada en la proyección de un mapa de calor, etc. Este programa, disponible gratuitamente en *Github*, puede ser libremente utilizado, mejorado y ampliado por los investigadores siempre que sea la voluntad de estos incorporar nuevas funcionalidades; por ejemplo, el programa podrá mejorarse, en futuras versiones, con la añadidura de otros elementos de

estudio, como el análisis de sentimientos presente en los textos o el filtrado más consistente de secuencias establecidas claramente como citas.

En los trece estudiantes seleccionados de forma aleatoria para el estudio efectuado, se ha determinado que algunos de ellos (H, B y M, de manera más evidente; y A, I y K, de modo más secundario) copiaron de forma literal el texto de Fernández Ordóñez (2016); el resto de estudiantes, aunque también pudieron incluir algunos fragmentos coincidentes, no superaron un porcentaje llamativo, siempre teniendo en cuenta que en el ámbito académico, sobre todo en el desarrollo de trabajos o pruebas de evaluación, es común tomar un texto externo como fuente directa y, en tal sentido, es permisible un cierto grado de intertextualidad directa. Esta similitud, si bien puede realizarse lícitamente mediante citas directas, no siempre queda manifiesta de esa manera en el texto final redactado por los estudiantes. Las motivaciones, de carácter variado, se han comentado ya en este trabajo (sección 1.1).

Como reflexión general, podemos concluir que en actividades relacionadas con el análisis textual, orientado a detección de plagio o caracterización del estilo idiolectal, es importante que los investigadores posean conocimientos de programación para no solo poder ingeniar y desarrollar sus propios programas de análisis, sino también para poder utilizar programas ya creados por otros especialistas o, en el caso de que estos sean de libre uso, poder adaptarlos y ampliarlos según las necesidades particulares que un estudio concreto pueda requerir.

Bibliografía

Alfaro Torres, Paloma y Juan Juárez, Teresa de (2014): «El plagio académico: formar en competencias y buenas prácticas universitarias», *RUIDERAe: Revista de Unidades de Información*, 6, 1, pp. 1-20.

Anson, Daniel W. J. (2022): «Personas of plagiarism: the construction of the plagiarist in Australian university subreddits», *Linguistics and Education*, 69, pp. 1-12, https://doi.org/10.1016/j.linged.2022.101050.

Anthony, Laurence (2022): *AntConc*, Tokio, Waseda University, http://www.laurenceanthony.net/software/antconc/.

Barrón Cedeño, Alberto (2014): «Software para la detección de plagio académico», en R. Comas Forgas y J. Sureda Negre (coord.), *El plagio académico en Educación Secundaria: características del fenómeno y estrategias de intervención*, Palma, Grupo de investigación de la Universidad de las Islas Baleares, pp. 85-100.

Bloch, Bernard (1948): «A Set of Postulates for Phonemic Analysis», *Language* 24, 1, pp. 3-46, https://doi.org/10.2307/410284.

Bloomfield, Lou (2016): *Wcopyfind*, https://plagiarism.bloomfieldmedia.com/.

Cabedo Nebot, Adrián (2021): *Fundamentos de estadística con R para lingüistas*, Valencia, Tirant Lo Blanch.

Cabedo Nebot, Adrián (2022): *Poisoned Texts*, https://github.com/acabedo/poisonedtexts.

Cayuela Mateo, Ana; Tauste Francés, Ana; Seguí Crespo, Mar; Esteve Faubel, José María y Ronda Pérez, Elena (2015): «¿Cómo medir el plagio entre alumnos universitarios?: Revisión de instrumentos utilizados en artículos científicos», en María Teresa Tortosa Ybáñez, José Daniel Álvarez Teruel y Neus Pellín Buades (coords.), *XIII Jornadas de Redes de Investigación en Docencia Universitaria: nuevas estrategias organizativas y metodológicas en la formación universitaria para responder a la necesidad de adaptación y cambio*, Alicante, Universidad de Alicante, pp. 210-216.

Cebrián Robles, Violeta (2020): *Estudio sobre el plagio en las facultades de educación*. tesis doctoral dirigda por Manuela Raposo Rivas, Vigo, Universidad de Vigo.

Chang, Winston; Cheng, Joe; Allaire, JJ.; Sievert, Carson; Schloerke, Barret; Xie, Yihui; Allen, Jeff; McPherson, Jonathan; Dipert, Alan; y Borges, Barbara (2021): «Shiny: Web Application Framework for R», https://CRAN.R-project.org/package=shiny.

Cicres i Bosch, Jordi y Gavaldà, Núria (2014): «La lingüística forense: la llengua com a evidència», *Revista de llengua i dret*, 61, pp. 60-71.

Código Penal (Ley Orgánica 10/1995, de 23 de Noviembre), 1995, Noviembre. https://vlex.es/vid/ley-organica-codigo-penal-126987.

Coulthard, Malcolm; Johnson, Alison; Kredens, Krzysztof y Wools David (2010): «Plagiarism Four forensic linguists' responses to suspected plagiarism», en Malcolm Coulthard y Alison Johnson (eds.), *The Routledge Handbook of Forensic Linguistics*, Londrés, Routledge (col. Routledge Handbooks in Applied Linguistics), pp. 523-538.

Díaz Arce, Dariel (2016): «Plagio académico en estudiantes de bachillerato: ¿qué detecta TURNITIN?». *RUIDERAe: Revista de unidades de información*, 9, pp. 1-31.

Eret, Esra y Gokmenoglu, Tuba (2010): «Plagiarism in higher education: a case study with prospective academicians», *Procedia - Social and Behavioral Sciences, Innovation and Creativity in Education*, 2, 2, pp. 3303-3307, https://doi.org/10.1016/j.sbspro.2010.03.505.

Fernández-Ordóñez, Inés (2016): «Dialectos del Español Peninsular», en Javier Gutiérrez Rexach (ed.), *Enciclopedia de Lingüística Hispánica*, Londres, Routledge, pp. 387-404.

Grant, Tim (2010): «Text Messaging Forensics Txt 4n6: Idiolect Free Authorship Analysis?», en Malcolm Coulthard y Alison Johnson (eds.), *The Routledge Handbook of Forensic Linguistics*, Londres, Routledge (col. Routledge Handbooks in Applied Linguistics), pp. 508-522.

Kiss, András Károly (2013): «Loopholes of Plagiarism Detection Software», en Aytekin Isman, Colleen Sexton, Teresa Franklin, Ahmet Eskicumali (ed.), *Procedia - Social and Behavioral Sciences, 4th International Conference on New Horizons in Education*, 106 (diciembre), pp. 1796-1803, https://doi.org/10.1016/j.sbspro.2013.12.202.

Law, Lily; Ting, Su-Hie y Jerome, Collin (2013): «Cognitive dissonance in dealing with plagiarism in academic writing», en Chee Siong Teh, Hee-Rahk Chae, Shahren Ahmad Zaidi Adruce, Philip Nuli Anding, Chwen Jen Chen, Norazila Abd Aziz, Kock Wah Tan (ed.), *Procedia - Social and Behavioral Sciences, The 9th International Conference on Cognitive Science*, 97 (noviembre), pp. 278-284, https://doi.org/10.1016/j.sbspro.2013.10.234.

Ley de propiedad intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia (Real Decreto Legislativo 1/1996, de 12 de Abril)», 1996, abril. http://vlex.com/vid/regularizando-aclarando-armonizando-127567.

Lukashenko, Romans; Graudina, Vita y Grundspeņķis, Janis (2007): «Computer-Based plagiarism detection methods and tools: an overview», en Boris Rachev, Angel Smrikarov y Dimo Dimov (ed.), *Proceedings of the 2007 International Conference on Computer Systems and Technologies*, Rousse, Bulgaria, Association for Computing Machinery, pp. 1-6, https://doi.org/10.1145/1330598.1330642.

McMenamin, Gerald R. (1993): *Forensic stylistics*, Ámsterdam, Elsevier.

Morey López, Mercedes; Sureda Negre, Jaume; Oliver Trobat, Miquel Francesc y Comas Forgas, Rubén Lluc (2013): «Plagio y rendimiento académico entre el alumnado de Educación Secundaria Obligatoria», *ESE:*

Estudios sobre educación, 24, pp. 225-244.

Ochoa Sierra, Ligia y Cueva, Alberto (2016): «Percepciones de estudiantes acerca del plagio: datos cualitativos», *Encuentros* 14, 2, pp. 25-41.

R Core Team. 2022. *R: A language and environment for statistical computing*, Viena, Austria, R Foundation for Statistical Computing, https://www.R-project.org/.

Ronda Pérez, Elena; Seguí Crespo, Mar; Cayuela Mateo, Ana; Tauste Francés, Ana; Lumbreras Lacarra, Blanca y Esteve Faubel, José María (2016): «RedPlag: el plagio en los trabajos docentes de los estudiantes universitarios», en José Daniel Álvarez Teruel, Salvador Grau Company y María Teresa Tortosa Ybáñez (coords.), *Innovaciones metodológicas en docencia universitaria: resultados de investigación*, San Vicente de Raspeig, Universidad de Alicante, pp. 633-648.

Sheard, Judy, Markham, Selby y Dick, Martin (2003): «Investigating differences in cheating behaviours of IT undergraduate and graduate students: the maturity and motivation factors», *Higher Education Research & Development*, 22, pp. 91-108, https://doi.org/10.1080/0729436032000056526.

STS 1204/2008, 18 de diciembre de 2008. http://vlex.com/vid/intelectual-sistema-informatico-datos-52043815.

Sun, Yu-Chih (2013): «Do journal authors plagiarize? Using plagiarism detection software to uncover matching text across disciplines», *Journal of English for Academic Purposes*, 12, 4, pp. 264-272, https://doi.org/10.1016/j.jeap.2013.07.002.

Sureda Negre, Jaume y Comas Forgas, Rubén Lluc (2006): «Ciber-Plagio académico. Una aproximación al estado de los conocimientos», *Textos de la CiberSociedad*, 10, pp. 1-6.

Sureda Negre, Jaume; Comas Forgas, Rubén Lluc y Morey López, Mercedes (2009): «Las causas del plagio académico entre el alumnado universitario según el profesorado», *Revista Iberoamericana de Educación* 50, 1, pp. 197-220.

Wijffels, Jan (2022): «Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit», https://CRAN.R-project.org/package=udpipe.

Woolls, David (2010): «Computational Forensic Linguistics searching for similarity in large specialised corpora», en Malcolm Coulthard y Alison Johnson (eds.), *The Routledge Handbook of Forensic Linguistics* (col. Routledge Handbooks in Applied Linguistics), Londres, Routledge, pp. 576-590.

Wright, David (2017): «Using word n-grams to identify authors and idiolects: a corpus approach to a forensic linguistic problem», *International Journal of Corpus Linguistics*, 22, 2, pp. 212-241, https://doi.org/10.1075/ijcl.22.2.03wri.