

ANÁLISIS DE TEXTOS DESDE LA ESTILOMETRÍA

Francisco Javier BLASCO PASCUAL y Cristina RUIZ URBÓN

Salamanca: Ediciones Universidad de Salamanca, 2023, 214 pp.

ISBN: 9788413117638

Análisis de textos desde la estilometría es el título de un volumen recientemente publicado dentro de la colección ‘Obras de referencia’ de la Editorial Universidad de Salamanca. Se trata de un nuevo trabajo de Javier Blasco y Cristina Ruiz Urbón, pioneros dentro de la universidad española en la aplicación de técnicas computacionales a textos literarios, una labor que llevan realizando desde hace más de diez años y que ha dado lugar a la creación de un portal web (www.estilometria.com), así como de numerosos estudios, fundamentalmente centrados en problemas de atribución de autoría y plagio, y trabajos de peritajes lingüísticos forenses. Sus investigaciones ya dieron lugar en 2010 al volumen *Hos ego versiculos feci...: estudios de atribución y plagio*, editado junto a Patricia Marín Cepeda y que presentaba una primera aproximación a estas casuísticas con las herramientas informáticas disponibles en aquel momento. El nuevo volumen amplía el foco de interés más allá de los problemas autoriales, en línea con el interés creciente en el campo de la estilometría por lo que en inglés se ha llamado *Stylometry Beyond Authorship Attribution* (esto es, la estilometría más allá de la autoría), y actualiza los métodos y herramientas empleados tomando en consideración tanto el estado de la cuestión actual de la Lingüística forense como de aquello que entendemos por estilometría en cuanto disciplina de las hoy llamadas Humanidades Digitales.

Desde la nota inicial de sus autores, *Análisis de textos desde la estilometría* se presenta no como una introducción a esta disciplina, sino como una muestra de algunos de sus enfoques, para lo cual se estructura en siete capítulos que ofrecen una descripción de la estilometría, de sus principales características y de los principales métodos de análisis utilizados, con abundantes ejemplos. Mientras que los tres primeros capítulos tienen un carácter más introductorio, los cuatro restantes recorren herramientas y metodologías de análisis a través de diferentes niveles lingüísticos: léxico (capítulo IV), morfológico (capítulo V), sintáctico (capítulo VI) y semántico (capítulo VII).

En el primer capítulo, “¿Qué es la estilometría”, se define la misma como “una disciplina dedicada al análisis estadístico de textos escritos, basada en la cuantificación de diversos aspectos del discurso, tales como la longitud de las palabras o de las frases,

la frecuencia de determinadas clases de palabras o partes de la oración (POS), el estudio de las palabras función, las palabras contenido, las colocaciones o las correlaciones..., sirviéndose para ello del concurso de otras disciplinas auxiliares, como la informática, la estadística y la filología” (p. 15). En las páginas que siguen se enfatiza el carácter interdisciplinar de esta línea de investigación y cómo la cuantificación de elementos mediante la lectura distante que ofrecen los métodos estilométricos permite aportar información objetiva complementaria al análisis filológico tradicional. A pesar de la escasez de corpus de textos disponibles y de herramientas preparadas para el español, especialmente en comparación con los recursos disponibles para el inglés, se vaticina un futuro de la investigación en literatura con gran presencia de la estilometría (sección I.1), sustentado en los casos exitosos de la misma tanto a lo largo de su historia como en la actualidad. A esta subsección sigue una dedicada a definir los objetivos del volumen (sección I.2), que se concretan en “ofrecer un ‘análisis de textos’ realizados ‘desde’ recursos y planteamientos metodológicos de la estilometría” (p. 28), y otra a definir el concepto de idiolecto (sección I.3), que de forma similar al de la “huella autorial” permite identificar a un autor en función de los rasgos lingüísticos que utiliza al hablar o escribir.

En el segundo capítulo, “El tratamiento de los textos y la obtención de datos”, se aborda el trabajo de preprocesamiento necesario antes de realizar un análisis estilométrico. Para ello se plantea cómo debe realizarse el tratamiento del corpus con el fin de garantizar su fiabilidad y para que este refleje en su totalidad el estilo del autor (sección II.1), a qué datos textuales concretos debe atenderse por su operatividad (sección II.2), y qué tareas deben llevarse a cabo para estructurar la información (sección II.3).

Como ya se ha comentado, a partir del cuarto capítulo, se desarrollan una serie de herramientas y técnicas de análisis clasificadas por niveles lingüísticos. El primero de ellos, “Análisis léxico”, se centra en la importancia del vocabulario para distinguir a una persona de otra, dedicando dos subapartados a las herramientas *stylo* (sección IV.1) y LIWC (sección IV.2). En el caso de *stylo* se destacan tanto técnicas no supervisadas implementadas en este, fundamentalmente los clásicos análisis de grupos (a los que los autores se refieren como *clustering* o agrupaciones), como supervisadas, centradas en funciones de clasificación. Los métodos de verificación de autoría también tienen cabida a través de la técnica de los impostores. A continuación, aparecen una sección dedicada a las palabras función (sección IV.3), de gran relevancia en atribución de autoría por escapar al control consciente del autor, y otra a la función tf-idf (sección IV.4), cuyo poder para extraer palabras distintivas de un texto se ejemplifica con el análisis de varios entremeses cervantinos. El capítulo se completa con dos secciones: una dedicada a herramientas *online* que trabajan con corpus de respaldo (sección IV.5), y otra centrada en herramientas a las que los autores se refieren como “de análisis mixtos” (p. 90), pues permitirían aunar el enfoque cuantitativo con el cualitativo (sección IV.6), y donde se incluyen técnicas como el análisis de sentimientos.

El foco de atención del quinto capítulo, “Morfología”, se traslada a la utilidad de los análisis de clases de palabras o categorías morfológicas (sección V.1), obtenidas mediante

Part-of-Speech tagging (esto es, etiquetado morfológico automático) a través de diferentes herramientas, con ejemplos que muestran los resultados de etiquetado fundamentalmente con el paquete *udpipe* de lenguaje de programación R y con el programa LIWC. Otros rasgos morfológicos cuya utilidad se discute para atribución de autoría son la prefijación y la derivación (sección V.2), con ejemplos de cómo pueden utilizarse para diferenciar dos obras de Cervantes y Lope de Vega.

El siguiente capítulo, “Sintaxis: la frase y sus componentes”, presenta diferentes aproximaciones al análisis sintáctico estilométrico (orientado especialmente a atribución de autoría). Se aborda el interés del análisis de frase medida en palabras (sección VI.1), los ngramas de palabras (sección VI.2), las colocaciones y coocurrencias (sección VI.3), los marcadores del discurso (sección VI.4), clases de palabras dentro de la frase (sección VI.5), nombres de persona y dependencias (sección VI.6) tipos de oración (sección VI.7), y, por último, de análisis sintácticos cualitativos (sección VI.8), con especial atención a las conjunciones y a la codificación de argumentos.

El último capítulo, “Semántica”, explora el comportamiento de un conjunto de textos ante campos semánticos obtenidos automáticamente con el programa LIWC (sección VII.1), el uso de la función tf-idf para obtener las palabras clave de un texto (sección VII.2), la obtención automática de los temas de un texto a través de nubes de frecuencias de palabras y de la aplicación de métodos de Topic Modelling (sección VII.3), y el interés de las asociaciones de palabras como aproximación al análisis del contenido semántico de un texto (sección VII.4).

En las páginas de conclusión que cierran el volumen, destaca la defensa del carácter complementario de este tipo de análisis al tiempo que se pone en valor la importancia del filólogo como especialista en el lenguaje. Resulta también especialmente reseñable de entre los puntos planteados la necesidad de la creación de diccionarios y recursos nativos en español que resulten altamente fiables, así como de corpus representativos con los que la investigación estilométrica pueda seguir avanzando. El volumen se completa con un listado de las herramientas, el corpus de referencia y los textos empleados en los análisis presentados, un útil glosario y un índice onomástico.

En definitiva, *Análisis de textos desde la estilometría* constituye una obra valiosa que plantea varios puntos de interés para el análisis estilométrico, y combina, además, la explicación de estos métodos y conceptos con nociones de Lingüística forense, en un esfuerzo por integrar y relacionar la estilometría con otras disciplinas. Supone, por tanto, una aportación de relevancia a la bibliografía ya existente aplicada a textos hispánicos, que, si bien es aún muy inferior a la de otras tradiciones literarias, va en notable aumento.

Laura Hernández-Lorenzo
Universidad Nacional de Educación a Distancia



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND).