REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2023 PUCV, Chile 56(111) 35-58

El análisis de autoría en Lingüística forense: Historia, concepción y revisión metodológica. Aplicación de la razón de verosimilitud a textos cortos en español

Authorship Analysis in Forensic Linguistics: History, Conception and Methodological Review. Application of the Likelihood Ratio to Short Texts in Spanish

Mario Crespo Miguel
INSTITUTO DE INVESTIGACIÓN EN LINGÜÍSTICA APLICADA
UNIVERSIDAD DE CÁDIZ
ESPAÑA
mario.crespo@uca.es

Recibido: 19-II-2021 / **Aceptado:** 04-VII-2022
DOI: 10.4067/S0718-09342023000100035

Resumen

El análisis de la autoría refiere a un conjunto de técnicas cuyo objetivo principal es determinar quién es el autor de un texto. En los últimos años se está produciendo un desarrollo considerable de este tipo de trabajos, lo que está llevando a un rápido desarrollo de la disciplina. Este estudio recorre los orígenes del análisis de autoría, identifica sus campos de actuación y define sus fundamentos metodológicos actuales. Entre ellos se destaca el uso de la razón de verosimilitud como marco que permite mostrar la fuerza de la evidencia en el ámbito forense. A partir de un corpus de textos cortos electrónicos, se muestran los principales aspectos que deben resolverse para su correcta aplicación: la selección de rasgos, la adecuada caracterización del estilo de un autor a partir de ellos, y el cálculo de probabilidades y su interpretación.

Palabras Clave: Lingüística forense, análisis de autoría, razón de verosimilitud, Lingüística de corpus, análisis de textos.

Abstract

Authorship analysis refers to all those techniques aiming at determining who has written a certain text. In recent years, there has been a big development of this type of works, leading to a quick development of the discipline. This paper defines its origin and current situation, identifies its fields of application, and outlines its current methodological foundations. In this respect, the Likelihood Ratio is an appropriate approach to show the strength of the evidence in forensic sciences. This study builds a

small corpus of short electronic texts and shows main issues to be resolved for its correct application: the selection of features, the appropriate characterization of an author's style based on them, and the calculation of probabilities and their interpretation.

Keywords: Forensic linguistics, authorship attribution, likelihood ratio, corpus linguistics, text analysis.

INTRODUCCIÓN

En los últimos años existe un creciente interés en Lingüística por las labores de análisis de autoría. Esta tarea se inserta en los estudios de Lingüística forense y tiene como objetivo determinar de la manera más exacta posible si un texto escrito ha sido producido por un sujeto en particular. Se caracteriza por el uso de conocimiento lingüístico cuando el lenguaje forma parte de las pruebas que se presentan en un juicio (Crystal, 1997; Stefanova Spassova, 2009). Para Garayzábal Heinze, Queralt Estévez y Reigosa Riveiros (2019), se trata de una disciplina que vincula aspectos formales, descriptivos y aplicados de la Lingüística con el ámbito del Derecho.

El término 'atribución de autoría' englobaría un conjunto amplio de procedimientos con la finalidad de determinar si un texto pertenece a un determinado autor o no, a partir de las características observadas en sus documentos escritos previamente. A este respecto, Ramírez Salado (2019: 480) indica que:

> "Mientras que en español observamos que 'atribución de autoría' presenta tres usos terminológicos, el inglés posee tres unidades distintas. Concretamente, atribución de autoría[1] se corresponde con *authorship attribution*, atribución de autoría[2] con *authorship verification* y atribución de autoría[3] con *authorship identification*".

De esta manera, se puede observar que la 'atribución de autoría' abarcaría un conjunto de labores tales como la 'verificación de autoría', la 'identificación de autoría', la 'elaboración de perfiles lingüísticos' o la 'detección de plagio', tareas todas de las que hablaremos más adelante. Es por ello que optaremos por la denominación 'análisis de autoría' en español para hacer referencia al conjunto de técnicas cuyo objetivo final es determinar quién es el autor de un texto o, al menos, saber cuáles son sus características (sexo, edad, procedencia, etc.). Se trataría por tanto de una acepción laxa, pero que busca así evitar posibles confusiones con otros términos que conceptualicen tareas más específicas.

Este trabajo tiene como objetivo recorrer los orígenes del análisis de autoría, identificar sus campos de actuación y definir sus fundamentos en el proceso judicial. En la última parte de este estudio, planteamos el marco metodológico de la 'razón de verosimilitud' y lo utilizaremos sobre un corpus de textos cortos provenientes de medios digitales. Queremos mostrar así los principales aspectos que deben resolverse

para su correcta aplicación: la selección de rasgos, la adecuada caracterización del estilo de un autor a partir de ellos, el cálculo de probabilidades y su interpretación.

1. Orígenes, desarrollo y campos actuales del análisis de autoría

El análisis de autoría se podría remontar al mismo inicio de la producción de textos escritos (Holmes, 1998), si bien, en la bibliografía especializada sobre el tema, apenas se documentan casos anteriores a la época moderna. Uno de los primeros antecedentes reseñado por varios autores (Stańczyk & Cyran, 2007; Goodman, Hahn, Marella, Ojar & Westcott, 2007; Frontini, 2008, citado en Picornell, 2012) sería la "Donación de Constantino", un documento de la Roma Imperial cuya autenticidad fue cuestionada durante la Edad Media. Sin embargo, las primeras controversias de autoría aparecen ya claramente en el siglo XVIII. De esta primera época, Olsson (2008) señala, por un lado, las discusiones sobre la autoría de la Biblia iniciadas por el sacerdote alemán Henning Bernhard Witter en 1711, quien afirmaba que los diferentes nombres de la divinidad en el Pentateuco indicaban que varios autores habían contribuido en su elaboración y, por otro, en las disquisiciones de James Wilmot en 1785 sobre la autoría de Bacon de ciertas obras atribuidas a Shakespeare.

Uno de los primeros estudios con una motivación científica y estadística del estilo lo encontramos con Augustus de Morgan en el siglo XIX, el primer Catedrático de Matemáticas del University College de Londres (Tweedie, Singh & Holmes, 1996; Olsson, 2004; Coulthard & Johnson, 2007; Ebrahimpour, Putniņš, Berryman, Andrew, Ng & Abott, 2015; Ramírez Salado, 2017). En concreto, Morgan trataba de determinar la autoría de la "Carta a los Hebreos" en el Nuevo Testamento por medio de la medición de la longitud de las palabras. Por su parte, Li (2010), Amuchi, Al-Nemrat, Alazab y Layto (2013), Jamak, Savatić y Can (2012) y Neal, Sundararajan, Fatima, Yan, Xiang y Woodard (2017) destacan igualmente de esta época las figuras de Thomas Mendenhall, que aplicaba técnicas de tamaño de palabras para distinguir entre las obras de Shakespeare, Marlowe y Bacon; la figura de William Benjamin Smith, que analizaba la longitud media de las frases para discriminar el estilo de autor, y la del polaco Wicenty Lutoslawski, quien acuña el término de estilometría (Fernández Trillo, 2015).

Ya en el siglo XX, aparecen trabajos con un claro interés por estudiar el estilo de una manera 'matemática' como Morozov (Morozov, 1915, citado en Jamak, Savatic & Can, 2012), Zipf (Garayzábal Heinze et al., 2019), Udney Yule (Olsson, 2004; Neal et al., 2017) o Cox y Brandwood (1959, citado en Holmes, 1998). Sin embargo, existen dos hechos fundamentales en el asentamiento del análisis de autoría forense tal como se entiende actualmente. El primero de ellos lo encontramos en 1968 cuando aparece por primera vez el término 'Lingüística forense' en el informe "*The Evans statements: A case for forensic Linguistics*" de Jan Svartvik. Este caso se considera el punto de partida de

la Lingüística forense (Coulthard & Johnson, 2007) y es un caso de análisis de autoría. A partir de este momento, se pone de manifiesto que la Lingüística puede colaborar en un proceso judicial o policial con sus conocimientos y técnicas.

El segundo de estos hitos, quizás más trascendental, es la aparición de la Informática. Desde su surgimiento en la década de los cuarenta, el uso de ordenadores empezó lentamente a incorporarse a todo tipo de tareas científicas, entre ellas la investigación lingüística. Ya en la época de los sesenta y setenta se empieza a generalizar su uso, por lo que no es casual que uno de los primeros estudios asistidos por computadora en Lingüística forense, fueran los realizados por Mosteller y Wallace en 1964 para investigar la autoría de los "*Federalist Papers*" (Mosteller & Wallace, 1964; Stamatatos, 2009; Ebrahimpour et al., 2015; Neal et al., 2017).

Desde mediados de los 80, con la universalización de la Informática, surgen nuevos métodos de análisis de corpus, lo que conlleva un incremento cuantitativo y cualitativo de los estudios de análisis de autoría. La Lingüística de corpus es una metodología empírica de trabajo basada en muestras de uso de la lengua, y facilitada por el empleo de ordenadores para procesar esos datos de una manera eficiente y rápida (Villayandre Llamazares, 2008). Dado su potencial para analizar el lenguaje, las técnicas de corpus han sido ampliamente utilizadas por los lingüistas forenses (Sousa-Silva, 2019; Nini, 2020), hasta el punto que Garayzábal Heinze et al. (2019) las consideran una de las principales metodologías de las que se nutre la Lingüística forense actual.

Durante la década de los 90 va a predominar el análisis de textos de autores de la Literatura (Holmes, 1998). Esto es debido a que los textos literarios constituían la fuente básica de creación de corpus hasta esta época. Ya en el siglo XXI, con la generalización de *Internet* y la mensajería móvil, se produce un crecimiento exponencial de las posibilidades de comunicación interpersonal (Salih, Balci & Salah, 2016), lo que lleva al análisis de autoría a diversificarse y examinar nuevas formas de interrelación como son los correos electrónicos, los blogs, los mensajes en redes sociales y los SMS (Ebrahimpour et al., 2015). Estos nuevos medios de interacción permiten la conexión entre individuos, pero pueden llegar a utilizarse para la ofensa, la amenaza y actividades delictivas (Garayzábal Heinze et al., 2019).

Cada día se envían más de 50.000 millones de mensajes (texto, *tweets*, mensajes instantáneos) a través de gran variedad de plataformas (Altamimi, Alotaibi & Alruban, 2019). Si bien el número de potenciales delitos ha aumentado, actualmente contamos con herramientas informáticas mucho más potentes, mayor disponibilidad de datos para experimentar, nuevas técnicas de aprendizaje automático, así como mejores algoritmos para el procesamiento del lenguaje natural (Neal et al., 2017). Esto hace que exista un campo de gran expansión, interés y utilidad social.

Actualmente se distingue entre varios casos de estudio en el análisis de autoría (Rocha, Shreirer, Forstall, Cavalcante, Theophilo, Shen, Carvalho & Stamatatos, 2017):

1. La verificación de la autoría. Se trata de una tarea de clasificación 1:1 con una clase positiva conocida, y una clase negativa de "todos los textos de todos los demás autores". Esta tarea trata de individualizar a un determinado autor frente a cualquier otro. Este caso también ha sido llamado análisis de autoría de 'clase abierta' (Juola, 2008).
2. Identificación de autoría. Se trata de un problema de clasificación 1:N, donde hay un autor desconocido y N posibles autores candidatos. Juola (2008) habla de análisis de 'clase cerrada' en esta situación.
3. Sobre estos dos, se añade un tercer caso genérico denominado 'perfil de autoría' (Juola, 2008; Altamimi et al., 2019; Nini, 2020) con el objetivo de extraer características del autor tales como edad, sexo, raza, cultura, educación, etc.

Esta triple clasificación es la que se encuentra en la mayoría de las referencias del campo (Elmanarelbouanani & Kassou, 2014; Ishihara, 2014; Picornell, 2014; Halvani, Winter & Pflug, 2016; Reddy, Vardhan & Reddy, 2016; Boenninghoff, Nickel, Zeiler & Kolossa, 2019; Lagutina, Lagutina, Boychuk, Vorontsova, Shliakhtina, Belyaeva, Paramonov & Demidov, 2019; Ramírez Salado, 2019). No obstante, algunos investigadores incluyen la 'detección de plagio' dentro de los problemas generales de autoría (Lalla, 2010; Gollub, Potthast, Beyer, Busse, Rangel, Rosso, Stamatatos & Stein, 2014; Halvani et al., 2016; Lagutina et al., 2019). Esto es debido a que, en el plagio, un autor puede llegar a asumir los rasgos de otro autor del que toma la información. El plagio es un problema de análisis de autoría más, desde el momento en que se intenta determinar si los rasgos estilísticos observados en un documento son consistentes con un autor determinado o si, por el contrario, existe divergencia estilística.

2. Fundamentos metodológicos de autoría forense

Todo análisis de autoría forense trata de determinar la correspondencia del texto del que se desconoce la autoría o texto 'dubitado', respecto a los rasgos de otros textos de un autor conocido o texto 'indubitado' (Olsson, 2004). Se trata de un problema de clasificación basada en clases: clase abierta y clase cerrada (cf. apartado anterior). Por clase ha de entenderse un autor, pero también una procedencia, un sexo, una edad, un tipo de personalidad, etc. Este procedimiento de análisis textual comienza con el estudio lingüístico de textos de autores conocidos (se observan determinadas características), y luego con un proceso de verificación sobre los textos de los que se desconoce la autoría aplicando los mismos descriptores anteriores (Stańczyk & Cyran, 2007). A partir de las características lingüísticas del texto en cuestión, se asocia la clase más probable o cercana (Elmanarelbouanani & Kassou, 2014).

Respecto al análisis de rasgos, por 'estilo' se entiende la manera peculiar de escribir o de hablar de un escritor o de un orador. El estilo es el resultado de las diferentes opciones que escoge el escritor de manera inconsciente de entre todas las que hubieran sido posibles (McMenamin, 2002). Cada vez que un hablante o, en este caso un escritor, emite un mensaje, producirá un texto único e idiosincrático con un determinado número de 'marcadores' y de recursos lingüísticos que lo harán irrepetible (Turell, 2008). De esta manera, para caracterizar un autor frente a otro, primeramente, se estudia la selección de unidades que se ha realizado, y, en segundo lugar, la combinación que se hace de estos elementos. Es por ello que el análisis de autoría considera los ejes lingüísticos sintagmático y paradigmático en la tarea de determinar el posible autor de un texto.

El análisis de autoría actual considera una gran variedad de técnicas de análisis y marcas de autoría (Stamatos, 2009; Stefanova Spassova, 2009; Neal et al., 2019), pero apenas existe consenso sobre el conjunto óptimo de rasgos estilísticos (Lagutina et al., 2019). Lo que funciona como buen rasgo de estilo para un autor en particular, puede no ser satisfactorio para otro (Olsson, 2004). La eficacia de la técnica empleada depende de las características del texto estudiado en cuestión y también del número de autores candidatos (Fernández Trillo, 2015). Entre los rasgos más usados destacan aquellos basados en la cuantificación de formas tales como n-gramas o secuencia de n elementos, el uso de determinadas palabras o caracteres, el tamaño de unidades y enunciados, etc. Como se puede apreciar, estos rasgos están basados en estructuras fijas, fácilmente procesables con un ordenador, lo que permite el análisis de grandes cantidades de texto muy rápidamente. Por el contrario, aquellos rasgos que introducen análisis funcionales, semánticos o pragmáticos, tales como clasificaciones de palabras de manera lógica o conceptual, relaciones léxicas, análisis sintáctico-funcionales, análisis de la cortesía, estudio de las implicaturas, etc. son los menos frecuentes.

Respecto a la expresión de resultados, existen dos enfoques principales (Koehler, 2013). Un primer tipo culminaría con la opinión subjetiva por parte de un experto, y el segundo con la declaración cuantitativa sobre el grado de correspondencia observada entre un objeto desconocido y una referencia conocida. Respecto al primero, Delgado (2012) nos indica que la opinión es normalmente modelada con lo que se conoce como 'escalas de certeza'. En ellas el investigador trata de graduar, bien el nivel de confianza o convicción al que ha llegado sobre la autoría de un texto después de realizar una comparación pericial, bien el mayor o menor grado de similitud o disimilitud que se ha observado en el proceso de cotejo de muestras (Garayzábal Heinze, Jiménez Bernal & Reigosa Riveiros, 2014). La siguiente figura ilustra la escala de certeza tal como aparece en McMenamin (2002). Esta escala está basada en nueve niveles donde 0 indica eliminación, 5 inconclusión y 9 identificación en el proceso de comparación de autoría:

RESEMBLANCE: Questioned vs. Known	**CRITERIA**	**CONSISTENCY: Questioned vs. Questioned**
9 IDENTIFICATION (did write)	1. Substantial significant similarities in the range of variation 2. No limitations present	**9 DEFINITE one writer**
8 HIGHLY PROBABLE did write	1. Substantial significant similarities in the range of variation 2. Limitations are present	**8 HIGHLY PROBABLE one writer**
7 PROBABLE did write	1. Some significant similarities in the range of variation 2. Limitations are present	**7 PROBABLE one writer**

Figura 1. Extracto de la escala de certeza tal como aparece en McMenamin (2002: 143).

Este tipo de escalas tiene la finalidad de guiar las conclusiones del experto y ayudar en la interpretación a los profesionales del campo judicial y policial en el ejercicio de la presentación de evidencias. Sin embargo, se trata de un enfoque subjetivo, ya que la elección de niveles se basa en la experiencia del experto y puede variar de un lingüista a otro (Queralt Estévez, 2019). Además, cada investigador puede establecer sus propias escalas atendiendo a criterios personales. Debido a esta variabilidad, Delgado (2002) indica que lo ideal es que estas escalas sean establecidas de antemano a través de estudios experimentales o trabajos de campo.

El segundo tipo de enfoque o enfoque cuantitativo, suele expresar sus resultados bien en forma de valores booleanos (pertenece 'sí' o 'no' a una determinada clase), bien mediante la expresión de la probabilidad de pertenecer a una categoría (o autor en este caso), o bien indicando cuáles son las clases más cercanas o parecidas (Savoy, 2016). Entre las técnicas más utilizadas encontramos (Swain, Mishra & Sindhu, 2017; Altamimi et al., 2019): S*upport-vector machines* (SVM), *Naive Bayes* o la 'Distancia euclidiana'. El SVM o 'Máquinas de Soporte Vectorial' es un algoritmo de aprendizaje automático supervisado capaz de construir un modelo que separa las diferentes clases a partir de datos de entrenamiento. Este modelo construye un hiperplano que separe de forma óptima las categorías estudiadas[1]. A la hora de clasificar, SVM asigna nuevos ejemplos a una categoría u otra en función de donde se hayan situado respecto al hiperplano. *Naive Bayes* es una técnica de aprendizaje basado en el teorema de Bayes que asume que las variables predictoras son independientes entre sí. Este algoritmo asigna la clase más probable a los nuevos casos. Por último, la 'Distancia euclidiana' trata de determinar la distancia de un elemento respecto a diferentes clases a partir de las características observadas. El sistema clasifica los nuevos ejemplos en función de qué clase se encuentre más próxima. Para una panorámica de la gran variedad de rasgos y algoritmos de clasificación en el campo del análisis de autoría, las obras generales sobre el tema (Stamatos, 2009; Elmanarelbouanani & Kassou, 2014; Neal et

al., 2017; Altamimi et al., 2019; Lagutina et al., 2019) ofrecen una visión detallada del estado de la ciencia.

A este respecto, Rose (2002) indica que el investigador no solo debería indicar el grado de coincidencia del texto con una determinada clase o autor, sino también la fuerza que tiene esa evidencia. En las ciencias forenses se asiste desde los años 90 a la adopción de un enfoque más empírico caracterizado por lo que se conoce como 'razón de verosimilitud' o *likelihood ratio* (LR) (Olsson, 2004; Queralt Estévez, 2014), que trata de dar cuenta del valor que tiene las pruebas presentadas en un juicio. El siguiente apartado plantea la conceptualización que subyace a esta metodología, cómo se realiza su cálculo y cómo debe interpretarse en el ámbito del análisis de autoría.

3. El marco de la razón de verosimilitud en el análisis de autoría

Aitken y Taroni (2010) relacionan la razón de verosimilitud o LR con el concepto de 'apuesta'. De esta manera, en el ámbito del deporte, apostar 'a favor de un caballo particular 6 contra 1' o en 'en el fútbol 3 contra 2 por un equipo', supone determinar la probabilidad de que este gane una carrera o un partido respecto a que ocurra lo contrario. Para estos casos, se calcula la probabilidad de un suceso y la probabilidad del suceso opuesto. Esto se conoce como 'complementariedad' o, suceso y negación de ese suceso. Si nos remitimos a las propiedades generales de la probabilidad, sea un suceso A, si el suceso A tiene una probabilidad P(A) de ocurrir, la apuesta en contra de A será 1-P(A). Si lo aplicamos a la forma de 'apuesta':

$$\frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}$$

Si llevamos esta formulación al ámbito judicial, nos encontramos con dos hipótesis competitivas: la del fiscal y la de la defensa, que presentan sucesos complementarios ('culpable' o 'inocente', por ejemplo) (Aitken & Taroni, 2010). Dadas estas dos proposiciones, lo que realmente se está midiendo es cuántas veces es más probable que se produzca un hecho respecto a otro contrario:

$$\frac{p(Hipótesis_a|Evidencia)}{p(Hipótesis_d|Evidencia|)}$$

Siendo $p(Hipótesis_a|Evidencia)$, la probabilidad de la hipótesis de la acusación a raíz de la evidencia, y $p(Hipótesis_d|Evidencia)$ la probabilidad de la $Hipótesis_d$ de la defensa bajo la misma evidencia. Un procedimiento que se ajusta a este tipo de formulación es el teorema de Bayes. Este teorema permite calcular la probabilidad a posteriori de un suceso, teniendo información de antemano sobre el mismo, e integrándola con las nuevas evidencias de las que se dispone. De esta manera, partiendo de los sucesos A y B, este teorema se define:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)} = P(B|A)\frac{P(A)}{P(B)}$$

Siendo $p(A|B)$, la probabilidad del suceso A sabiendo que se da el suceso B, y $P(B|A)$, la probabilidad del suceso B sabiendo que se da el suceso A. Ambos son casos de probabilidad condicionada. De esta forma, el suceso contrario quedaría:

$$P(\bar{A}|B) = P(B|\bar{A})\frac{P(\bar{A})}{P(B)}$$

y si lo aplicamos sobre el cociente de probabilidades complementarias al que hacíamos referencia anteriormente daría:

$$\frac{P(A|B)}{P(\bar{A}|B)} = \frac{\frac{P(B|A)\,P(A)}{P(B)}}{\frac{P(B|\bar{A})\,P(\bar{A})}{P(B)}} = \frac{P(B|A)}{P(B|\bar{A})} \cdot \frac{P(A)}{P(\bar{A})}$$

Lo que en el ámbito judicial sería:

$$\frac{P(Hipotesis_a|Evidencia)}{P(Hipotesis_d|Evidencia)} = \frac{P(Evidencia|Hipotesis_a)}{P(Evidencia|Hipotesis_d)} \cdot \frac{P(Hipotesis_a)}{P(Hipotesis_d)}$$

Donde el valor resultante de la oposición de las hipótesis competitivas que se pueden establecer en un juicio proviene de haber evaluado la probabilidad de esa evidencia dada la hipótesis de la acusación p(Evidencia|Hipótesis$_a$), dividido por la probabilidad de esa misma evidencia dada la hipótesis de la defensa p(Evidencia|Hipótesis$_d$). A este primer cociente se lo conoce como razón de verosimilitud o LR. El valor de esta evidencia se multiplica finalmente por la división de probabilidades a priori o conocimiento previo, p(Hipótesis$_a$), probabilidad en favor de la acusación y, p(Hipótesis$_d$), probabilidad en favor de la defensa.

Rose (2002) indica que es pretencioso que el experto forense trate de determinar la culpabilidad o no del sospechoso, ya que esa es precisamente la misión del Tribunal o el Jurado. El experto forense deberá concentrarse realmente en valorar la evidencia bajo la hipótesis de la acusación y bajo la hipótesis de la defensa (Aitken & Taroni, 2010), es decir, debe centrarse solamente en el cálculo del LR. La razón de verosimilitud o *likelihood ratio* (LR) en el ámbito del análisis de autoría será:

$$\frac{P(Evidencia|Autor)}{P(Evidencia|\overline{Autor})}$$

Siendo p(Evidencia| Autor), la probabilidad de una evidencia en un autor determinado y p(Evidencia| $\overline{\text{Autor}}$), la probabilidad de encontrar esa misma evidencia en cualquier otro autor. Los resultados de esta fórmula tienen propiedades

matemáticas que pueden ser útiles. Mientras que las probabilidades tienen un rango de 0 a 1, el LR tiene un rango de 0 a ∞. Rose (2002) establece un equivalente verbal a los posibles resultados del LR, lo que permite determinar la fuerza de la evidencia:

Tabla 1. Tabla de interpretación del LR en Rose (2002).

Valor del LR	**Equivalente verbal**	
>10 000	Evidencia muy fuerte a la…	Hipótesis de la acusación
1000 a 10 000	Evidencia fuerte a la…	
100 a 1000	Evidencia moderadamente fuerte a la…	
10 a 100	Evidencia moderada a la…	
1 a 10	Evidencia limitada a la…	
	Evidencia limitada a la…	Hipótesis de la defensa
1 a 0.1	Evidencia moderada a la…	
0.1 a 0.01	Evidencia moderadamente fuerte a la…	
0.01 a 0.001	Evidencia fuerte a la…	
0.001 a 0.0001	Evidencia muy fuerte a la…	
<0.0001		

Cuanto más se aproximan los resultados del LR a 1, menos información obtenemos de la evidencia que estamos analizando. Sin embargo, cuanto más supere el resultado a la unidad, mayor será el apoyo a la hipótesis de la fiscalía (> 1), es decir, ese escrito es más probable para ese autor en particular. Por el contrario, un valor muy inferior a 1 apoyaría las tesis de la defensa (< 1), es decir, ese escrito podría haber sido producido por cualquier individuo no sospechoso. Así, por ejemplo, la obtención de un LR de 10 significaría que es diez veces más probable que la prueba se produzca si el autor sospechoso y el autor dubitado sean el mismo individuo, que si fueran individuos diferentes (Ishihara, 2014). Sin embargo, como observamos en la Tabla 1, es a partir de 100 cuando realmente se empieza a considerar que existen verdaderas evidencias a favor de la autoría de un texto. Y, de manera contraria, serán los valores inferiores a 0.01 los que muestren mayores indicios para rechazar que un texto pertenezca a cierto individuo.

Finalmente, otras de las ventajas del cálculo de la razón de verosimilitud es que se pueden combinar diferentes LR mediante su producto, siempre que tales cálculos provengan de evidencias que sean independientes entre sí. Así dos LR procedentes de evidencias independientes, por ejemplo, un LR de 3 y un LR de 4.2, se podrían combinar dando lugar a un nuevo LR (3X4.3) = 12.6

Sin embargo, tal como explica Queralt Estévez (2014), este tipo de formulación aún no se ha materializado de manera clara en el análisis de autoría. Esto es debido a que este enfoque requiere de un análisis poblacional de los rasgos lingüísticos analizados, lo que llevaría a determinar la 'rareza' o 'tipicidad' de ese rasgo lingüístico. De esta manera, en 'verificación de autoría' o clase abierta (cf. apartado 1), el cálculo del LR requiere trabajar con poblaciones representativas que permitan caracterizar a un autor respecto a toda la población y la mayoría de las veces no existe esta clase de

corpus. Gran parte del problema de ausencia de textos de referencia proviene del hecho de que la Lingüística de corpus no defina de manera clara qué debe ser considerado como corpus representativo, qué variables de estudio deben considerarse de manera inequívoca al estudiarlo (Corpas & Seghiri, 2007) y cuál el mínimo de palabras o documentos necesarios para poder caracterizar el estilo con claridad (Stamatos, 2009). Todo ello va a lastrar otras disciplinas que dependen de tales consideraciones, como es el caso de la Lingüística forense.

Este trabajo plantea la aplicación del LR a la 'identificación de autoría o clase cerrada (cf. apartado 1), donde sabemos además que uno de los autores de esa población reducida es el autor del texto. Se trata de una situación simplificada donde la caracterización de la población es relativamente sencilla al ser pequeña. Para ello hemos conformado un corpus de 442 textos electrónicos cortos, procedentes de 16 autores diferentes. Con este estudio no queremos ser exhaustivos, sino mostrar algunos de los problemas existentes en análisis de autoría bajo esta forma de contrastar probabilidades.

4. Aplicación del método

4.1. Corpus, rasgos y cálculo de probabilidades

La investigación reciente en análisis de autoría se centra en textos cortos como correos electrónicos, redes sociales, mensajes de teléfonos móviles (Jiménez, 2014). Esto plantea un gran desafío, ya que muchos de ellos tienen menos de 200 palabras (Coulthard & Johnson, 2007). Los trabajos de atribución de autoría de textos cortos para el español son escasos. Entre ellos encontramos a Rico Sulayes (2012) y Crespo (2016) para textos de foros de *Internet*, y más recientemente a Castillo Velásquez, Martínez Godoy, Torres Falcón, Zavala de la Paz, Becerra Chávez y Rizzo Sierra (2020) para textos de *Twitter*. Sin embargo, estos trabajos no utilizan el LR en su metodología.

Este trabajo presenta una sencilla aplicación de la razón de verosimilitud a partir de un pequeño corpus de 442 textos aleatorios procedentes de 16 autores del foro 'Eskup' de la edición digital del periódico español El País (previa petición a este medio de prensa). Eskup es una red que permite a los lectores, periodistas y usuarios de otras redes sociales como *Twitter* o *Facebook*, enviar mensajes cortos para opinar sobre los contenidos publicados en la edición online de El País. Desafortunadamente, no siempre se tiene disponible un conjunto de documentos ideal, por lo que las peculiaridades de este medio de difusión no permiten saber las características sociodemográficas de los autores. La Tabla 2 muestra el número de mensajes por autor.

Tabla 2. Número de textos por autor.

ab0	antico	arica	ariga	bald	bend	cag	jav	Total
28	18	26	25	30	29	30	37	442
Thom	**Susil**	**Skunk**	**Portell**	**Millet**	**Meeti**	**jos**	**kbz**	
35	18	21	32	27	33	28	25	

Los textos constan de unas 30 palabras cada uno. La siguiente figura ejemplifica cómo son estos textos[2]:

ab0: Es lo que tiene vivir en un basurero, vulnera que? quiere decir que el basurero conoce la leyes Europeas, el basurero vulnera todo lo que se le ponga por delante.....
ab0: La jefa del asunto publica una carta dirigida a un medio par que la difunda, se sentían muy mal por llamarles lo que realmente son, la carta en síntesis(utilizando términos del abogado de esta insti trilera) RESPETAMOS ESCRUPULOSAMENTE LA LEGISLACIÓN VIGENTE

Figura 2. Ejemplo de autor y texto del corpus.

Todos estos mensajes fueron analizados observando tres tipos de rasgos: bigramas[3] de caracteres, lemas[4] y clases de palabras. En el primer caso, se creó una lista de todos los bigramas del corpus y se analizó el número de veces que aparecían por texto. Para obtener los lemas y las clases de palabras se recurrió al programa *TreeTagger* (Schmid, 1994). Göhring (2009) indica que su precisión media llega a alcanzar el 93,53% para el español. Igualmente, se analizó el número de veces que aparecían en cada texto. La Figura 2 muestra esta caracterización:

Rasgo	*Autores*			
Ejemplo de rasgos	ab001	ab001	antico	antico
Clase de palabra				
cque	2	2	3	1
csubx	1	2	0	1
Lema				
ser	4	2	1	0
un	0	3	3	3
Bigrama				
Pr	3	4	2	7
Ge	8	5	7	7

Figura 3. Ejemplo de aparición de clases de palabras en los textos de nuestro corpus.

Observamos una serie de problemas iniciales. Existen rasgos que solo aparecen en uno de los textos del corpus, por lo que su valor como evidencia es nulo. De esta manera y como criterio metodológico, solo mantuvimos aquellos rasgos que podían observarse en, al menos, 10 de los 442 textos y en más de un autor. Para los casos en los que un rasgo no era observado en un determinado autor, que haría que el producto final fuese cero, aplicamos el suavizado de *Laplace* o *Laplace Smoothing*, una técnica que ayuda a abordar este problema. Para estos casos, añadimos uno al recuento de cada clase, lo que elimina todos los posibles valores cero.

Igualmente, como ya indicamos en el apartado anterior, solo se pueden combinar diferentes LR si las evidencias son independientes entre sí. Para determinar si existía la suficiente independencia, se hizo un 'análisis de correlación' entre todos los rasgos. La correlación es una técnica matemática de amplio uso que mide la fuerza de la relación entre valores cuantitativos pareados *x* e *y* en una muestra (Triola, 2013). Cuanto más fuerte sea la asociación entre las dos variables, obtendremos valores más cercanos a 1 o -1. Si dos variables son independientes, estarán incorrelacionadas, por lo que el valor del coeficiente más se acercará a 0. Concretamente hemos usado el 'Coeficiente de correlación de Spearman' que no exige que los datos analizados vengan de una distribución normal. Finalmente, Evans (1996) sugiere los valores que abarcan el rango <0,39 y >-0,39 muestran una correlación débil. A partir de este punto se tratarán los rasgos con baja correlación como independientes. Tras correlacionar todas las características entre sí y descartar todos aquellas que superaban este umbral, obtuvimos una lista reducida de rasgos:

Tabla 3. Reducción de rasgos tras aplicación de *Spearman*.

Bigramas		**Clases de palabras**		**Lemas**		**Bigramas + Clases de palabras + Lemas**	
Inicial	Tras filtrado	Inicial	Tras filtrado	Inicial	Tras filtrado	Inicial	Tras filtrado
1143	322	108	42	2447	211	3698	453

Finalmente, debido a que los textos son extremadamente cortos (sobre 30 palabras cada uno), solo se observó si el rasgo aparecía o no en el texto, independientemente del número de veces que aparecía. A partir de ahí se calcularon probabilidades de manera sencilla. Se trata de un caso de probabilidad condicionada, que se resuelve de dividir el número de casos donde se observa el rasgo entre el total del autor o del resto de autores:

$$P(Evidencia|Autor) = \frac{P(Evidencia \cap Autor)}{P(Autor)}$$

$$P(Evidencia|Autor) = \frac{P(Evidencia \cap \overline{Autor})}{P(\overline{Autor})}$$

Para ello se suele recurrir a un diagrama de árbol:

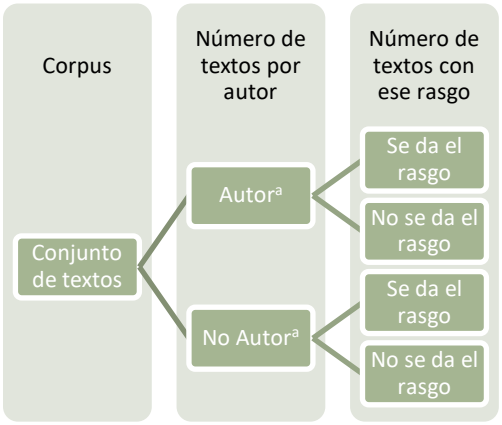

Figura 4. Diagrama de árbol aplicado para computar probabilidades.

Finalmente, como los rasgos son independientes podemos hacer el producto de cada uno de ellos:

$$P(R_1|A) \times \ldots \times P(R_n|A) = \prod_{i=1}^{n} P(R_i|A)$$

Lo que llevado al ámbito de la razón de verosimilitud obtendríamos:

$$\frac{P(Evidencia|Autor)}{P(Evidencia|\overline{Autor})} = \frac{\prod_{i=1}^{n} P(Evidencia_i|Autor)}{\prod_{i=1}^{n} P(Evidencia_i|\overline{Autor})}$$

4.2. Resultados

Cada uno de los 442 textos fue separado del corpus sucesivamente y considerado como texto 'dubitado'. A partir de los rasgos observados en este texto en cuestión, se computaba con el resto de ellos la probabilidad de encontrar tales características en un autor y de encontrar tales elementos en cualquier otro. De la división de tales probabilidades obteníamos el LR de la evidencia. Se probó cada grupo de rasgos por separado (bigramas, lemas y clases de palabras) y su combinación (Com).

Este método se probó con cada texto y para con todos los autores del corpus, ya que queríamos determinar hasta qué punto el sistema atribuía y rechazaba autoría de manera correcta e incorrecta. Esto hace un total de 7.072 comparaciones, de las que en 442 casos se comparaba con su autor real (AR), y en 6.630 con otro autor diferente (AD). Como sabíamos a priori a quién pertenece cada texto, se podía contabilizar si el sistema atribuía autoría de manera correcta y el grado de fuerza de esa predicción basada en el LR. A partir de los resultados se calculó la tasa de 'verdaderos positivos' (TP), 'falsos positivos' (FP), 'verdaderos negativos' (TN) y 'falsos negativos' (FN), así como el número de 'indeterminados' (IND) procedentes de diferentes umbrales de

aceptación y rechazo. En este estudio hemos asumido tres escalas diferentes: una primera que acepta autoría cuando es superior a 10 y rechaza cuando es inferior a 0,1; una segunda que asume autoría cuando es mayor que 100 y rechazo cuando es inferior a 0,01, y una tercera que acepta atribución cuando es mayor que 1000 y rechazo al ser inferior a 0,001. Los valores intermedios se consideran indeterminados.

Con los valores de TP, FP, TN, FN e IND, se calculó el 'valor predictivo positivo' o *Positive Predictive Value* (PPV), calculado como $\frac{TP}{TP+FP}$, que indica la precisión del algoritmo para aceptar autoría, y el 'valor predictivo negativo' o *Negative Predictive Value* (NPV), $\frac{TN}{TN+FN}$, que estima la exactitud para rechazar que un texto pertenezca a un determinado autor. Para analizar el índice de error, mostramos la 'tasa de falsos negativos' o *False Negative Rate* (FNR) que establece el grado de falsos negativos de aquellos que deberían haber sido positivos y se calcula como el cociente de falsos negativos entre el total de los que deberían haber sido positivos $\frac{FN}{AR}$. Finalmente, mostramos la 'tasa de falsos positivos' o *False Positive Rate* (FPR) que muestra la proporción de falsos positivos de aquellos que deberían haber sido considerados negativos, y se calcula mediante el cociente de falsos positivos del total de aquellos que deberían haber sido considerados negativos $\frac{FP}{AD}$. La siguiente tabla muestra los resultados:

Tabla 4. Tabla de resultados.

Resultados												
	<0,1 rechazo >10 aceptación				**<0,01 rechazo >100 aceptación**				**<0,001 rechazo >1000 aceptación**			
	Big	**CP**	**Lem**	**Com**	**Big.**	**CP**	**Lem**	**Com**	**Big.**	**CP**	**Lem**	**Com**
TP	**223**	**62**	**266**	**260**	**147**	**17**	**148**	**196**	**96**	**8**	**56**	**116**
FP	**528**	**114**	**1273**	**717**	**169**	**12**	**286**	**264**	**28**	**0**	**37**	**87**
TN	**4399**	**807**	**1604**	**4081**	**3143**	**165**	**517**	**2905**	**1964**	**0**	**96**	**1850**
FN	**83**	**5**	**10**	**61**	**39**	**0**	**0**	**37**	**14**	**0**	**0**	**15**
IND	**1839**	**6084**	**3919**	**1953**	**3574**	**6878**	**6121**	**3670**	**4970**	**7064**	**6883**	**5004**
PPV	**29%**	**35%**	**17%**	**26%**	**46%**	**58%**	**34%**	**42%**	**77%**	**100%**	**60%**	**57%**
NPV	**97%**	**99%**	**99%**	**98%**	**98%**	**100%**	**100%**	**98%**	**99%**	**IND**	**100%**	**99%**
FNR	**30%**	**1%**	**3%**	**22%**	**14%**	**0%**	**0%**	**13%**	**5%**	**0%**	**0%**	**5%**
FPR	**21%**	**4%**	**51%**	**28%**	**6%**	**0%**	**11%**	**10%**	**1%**	**0%**	**1%**	**3%**

En primer lugar, podemos observar cómo a medida que aumentamos el umbral del LR, mejora la precisión tanto a la hora de determinar si un texto pertenece a un determinado autor (PPV), como para rechazar autoría (NPV). Igualmente, las tasas de falsos positivos y falsos negativos (FNR y FPR) también mejoran en las mismas situaciones. Sin embargo, el número de indeterminados se acrecienta a medida que restringimos tales umbrales. En estos casos no existe la suficiente evidencia para poder asignar el texto a una clase determinada.

De todos los rasgos, las clases de palabras (CP) llegan a una tasa del 100% en la atribución positiva (TP), pero su tasa de rechazo (TN) es nula y posee un alto grado de indeterminados. Comparativamente serán los bigramas y lemas los que obtienen mejores tasas de acierto y rechazo, mejorando incluso a la combinación de rasgos (Com). También es destacable que la tasa de precisión en el rechazo de autoría (TN) se mantiene estable en general, debido a que obtiene valores de LR altos, lo que hace que no se vea afectada por la adopción de valores más restrictivos.

4.3. Discusión

Es evidente que lo que funciona como buen rasgo de estilo para un autor, no tiene por qué ser satisfactorio para otro (Olsson, 2004). Es por ello que el análisis de autoría debe preguntarse qué elementos, del conjunto de rasgos posibles, caracterizan más a un autor en particular. Los rasgos de estilo no deben entenderse de manera privativa, sino más bien como un uso preponderante o no frente a la generalidad. Es posible analizar las diferencias entre variables nominales o cualitativas en términos de su aparición en los textos. Una de las pruebas más usadas con este fin es el 'Test Chi-cuadrado' (χ^2), que permite determinar si las diferencias observadas entre variables de estudio son lo suficientemente grandes para determinar que se comportan diferente (Triola, 2013). Aplicado al análisis de autoría llevaría a mostrar si el uso de determinado rasgo en un autor difiere de lo normal. Sin embargo, solo es aplicable a datos de una tabla de contingencia si las frecuencias esperadas son lo suficientemente grandes. Una prueba similar es el 'test exacto de Fisher', que permite trabajar con muestras pequeñas que no cumplen las condiciones de aplicación del test χ^2. Este test se basa en una distribución hipergeométrica, que calcula la probabilidad exacta de obtener una determinada distribución de eventos. Supóngase la siguiente tabla de contingencia:

Tabla 5. Tabla de contingencia 2x2.

		Característica A		
		Presente	**Ausente**	**Total**
Característica B	**Presente**	a	b	a + b
	Ausente	c	d	c + d
	Total	a + c	b + d	n

$$p - valor = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

Figura 5. Fórmula para el test exacto de Fisher.

Se asume la independencia con un 'p-valor' inferior a 0,05, lo que en la práctica significa que, para un autor determinado, solo se toman aquellos rasgos que obtengan

un p-valor inferior a este número. Las Tabla 6 y Tabla 7 muestran el número total de rasgos usados finalmente y los usados finalmente por autor:

Tabla 6. Rasgos usados por autor tras el filtrado.

		Rasgos por autor tras filtrado							
	Rasgos iniciales	ab0	antico	arica	ariga	bald	bend	cag	jav
Big	**346**	**54**	**22**	**28**	**52**	**41**	**38**	**36**	**23**
CP	**42**	**4**	**1**	**1**	**9**	**3**	**7**	**7**	**4**
Lemas	**211**	**25**	**4**	**16**	**18**	**18**	**21**	**21**	**19**
Todos	**599**	**62**	**24**	**33**	**57**	**49**	**44**	**44**	**31**
	Rasgos iniciales	jos	kbz	meet	mill	porte	skunk	susil	thom
Big	**346**	**21**	**25**	**42**	**29**	**27**	**23**	**24**	**48**
CP	**42**	**0**	**5**	**1**	**4**	**4**	**1**	**4**	**10**
Lemas	**211**	**11**	**11**	**23**	**7**	**11**	**14**	**8**	**35**
Todos	**599**	**25**	**30**	**52**	**32**	**31**	**28**	**26**	**60**

Como se aprecia en las dos tablas anteriores, los autores no comparten la misma cantidad de rasgos específicos. De manera general, autores como 'thom' poseen una gran cantidad de rasgos que lo individualizan frente a la generalidad y otros, como 'antico', con casi tres veces menos. Una vez determinados los rasgos específicos de cada autor, se aplicaba la misma metodología descrita anteriormente (cf. apartado 4.2). Se computaba la probabilidad de encontrar tales características en un autor y de encontrar tales elementos en cualquier otro usando el corpus de referencia. En el caso de que no se hallaran rasgos específicos para un autor, no se calculaba la evidencia basada en ese rasgo. Los resultados tras la aplicación de este filtrado son los siguientes:

Tabla 7. Tabla de resultados.

	<0,1 rechazo >10 aceptación				**<0,01 rechazo >100 aceptación**				**<0,001 rechazo >1000 aceptación**			
	Big	**CP**	**Lem**	**Com**	**Big.**	**CP**	**Lem**	**Com**	**Big.**	**CP**	**Lem**	**Com**
TP	**139**	**56**	**94**	**142**	**67**	**11**	**30**	**60**	**20**	**8**	**9**	**29**
FP	**196**	**70**	**100**	**188**	**22**	**5**	**1**	**24**	**4**	**0**	**0**	**2**
TN	**3961**	**561**	**1891**	**4295**	**2162**	**0**	**552**	**2521**	**878**	**0**	**72**	**1095**
FN	**81**	**3**	**45**	**98**	**23**	**0**	**11**	**20**	**2**	**0**	**0**	**3**
IND	**2695**	**6382**	**4942**	**2349**	**4798**	**7056**	**6478**	**4447**	**6168**	**7064**	**6991**	**5943**
PPV	**41%**	**44%**	**48%**	**43%**	**75%**	**68%**	**96%**	**71%**	**83%**	**100%**	**100%**	**93%**
NPV	**97%**	**99%**	**87%**	**97%**	**98%**	**IND**	**98%**	**99%**	**99%**	**IND**	**100%**	**99%**
FNR	**29%**	**1%**	**16%**	**35%**	**8%**	**0%**	**3%**	**7%**	**0%**	**0%**	**0%**	**1%**
FPR	**7%**	**2%**	**4%**	**7%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**

Podemos observar una mejora generalizada de todos los valores de precisión y error. Como ya ocurría anteriormente, la adopción de umbrales más restrictivos mejora los resultados de PPV y de NPV. Sin embargo, la reducción de rasgos a los más relevantes ha traído consigo un aumento de los textos indeterminados. Los mejores resultados los obtiene la combinación de rasgos al obtener la mayor cantidad de TP y TN, la menor cantidad de indeterminados y las tasas de error más bajas. La

siguiente tabla muestra el porcentaje de mejora de estos resultados respecto a los que presentábamos en la sección anterior:

Tabla 8. Variación porcentual respecto a los anteriores (Tabla 12 respecto a la Tabla 11).

	<0,1 rechazo >10 aceptación				<0,01 rechazo >100 aceptación				<0,001 rechazo >1000 aceptación			
	Big	**POS**	**Lem**	**Com**	**Big.**	**POS**	**Lem**	**Com**	**Big.**	**POS**	**Lem**	**Com**
PPV	**+12**	**+9**	**+31**	**+17**	**+29**	**+10**	**+62**	**+29**	**+6**	**0**	**+40**	**+36**
NPV	**0**	**0**	**-12**	**-1**	**0**	**IND**	**-2**	**+1**	**0**	**IND**	**0**	**0**
FNR	**-1**	**0**	**+13**	**+13**	**-6**	**0**	**+3**	**-6**	**-5**	**0**	**0**	**-4**
FPR	**-14**	**-2**	**-47**	**-21**	**-6**	**0**	**-11**	**-10**	**-1**	**0**	**-1**	**-3**

Como se puede apreciar, los rasgos con mayor mejora porcentual son los lemas y, los que menos, las clases de palabra. De esta manera, los lemas llegan a conseguir hasta una mejora en PPV del 62% con un umbral de <0,01 rechazo >100 aceptación y una reducción del error de hasta el 47% en los falsos positivos (FPR). También se destaca que la mejora de resultados es mucho más pronunciada para TP que para TN. Esto se explica porque las tasas de NPV de la sección anterior ya mostraban buenos resultados.

CONCLUSIONES

En este artículo hemos llevado a cabo un recorrido por la historia y concepción actual del análisis de autoría en el ámbito de la Lingüística forense. Se trata actualmente de un campo de gran investigación, donde la aplicación de técnicas propias de la Lingüística computacional y la Lingüística de corpus está contribuyendo al desarrollo de esta disciplina. En general estos métodos permiten el tratamiento de grandes cantidades de texto muy rápidamente. Sin embargo, desde el ámbito del Derecho, las técnicas deben cumplir con ciertos estándares científicos que permitan su aplicación en el ámbito judicial y policial con una serie de garantías. Es por ello que cobra especial importancia cómo se va a realizar el análisis de los datos y la posterior expresión de resultados.

Como hemos podido ver, la tendencia actual en el marco forense es determinar el grado de fuerza de la evidencia considerando igualmente tanto la hipótesis de la acusación como la de la defensa. Para ello se recurre a lo que se conoce como razón de verosimilitud. La aplicación de esta técnica en Lingüística forense supone un gran esfuerzo para el experto ya que la comparación no deber hacerse solo considerando a un determinado autor, sino también observando el valor que puedan tener las evidencias en la población. Es por ello que no puede haber un análisis de autoría forense real si no se cuenta con un corpus de referencia sobre los usos estilísticos de la población a la que pertenezca el sospechoso, y poder saber qué hablantes utilizan qué estructuras, en qué tipo de circunstancias (destinatarios, mensajes) y con qué fines (actos de habla transmitidos en cada comunicación). Esto nos permitirá determinar si

los rasgos lingüísticos que se están considerando son realmente discriminantes. Si el análisis de autoría forense quiere consolidarse como disciplina deberá abordar este tipo de aspectos metodológicos.

Hemos planteado un pequeño estudio que muestra el uso de la razón de verosimilitud como método de expresión de resultados. En primer lugar, hemos hallado problemas a la hora de seleccionar los rasgos adecuados. Muchos de estos no se encuentran en todos los textos, además el cálculo del LR exige determinar que estos sean independientes entre sí. Una vez solucionado este aspecto, el cálculo de la probabilidad se ha realizado mediante conteo. Sin embargo, existen otros métodos de mayor complejidad y mejores resultados que deberán ser explorados en el futuro. Entre ellos destacamos la 'Regresión logística binaria', de amplio uso también en el ámbito forense.

Comprobamos cómo la adopción de diferentes umbrales de aceptación y rechazo más altos en el LR permite hacer que el sistema mejore en sus valores de precisión y en el número de falsos positivos y negativos, un problema serio en el ámbito judicial. Sin embargo, estos umbrales tan restrictivos llevan parejo que crezca el número de indeterminados. En este sentido, consideramos que el trabajo forense debe estar destinado a mostrar el valor de la evidencia, así como a intentar preservar la presunción de inocencia, por lo que no consideramos un problema serio esta pérdida de eficacia, ya que apenas se cometen errores cuando se ha podido llegar a realizar la atribución.

Finalmente hemos demostrado cómo la selección de rasgos por autor es una herramienta eficaz para mejorar los resultados. El 'test exacto de Fisher' permite cumplir con los requerimientos estadísticos e individualizar el estilo del autor frente al resto. Sin embargo, no todos los escritores se caracterizan de la misma manera. Es por ello que habría que contar con una gama de rasgos lo suficientemente rica, variada y representativa de lo que puede ser el estilo de un autor, así como disponer de un corpus lo suficientemente amplio para poder hacer las adecuadas observaciones y extraer los datos necesarios para el cómputo de estadísticas. Ambos van a ser aspectos cruciales y determinarán los buenos resultados de este tipo de trabajos y, por ende, del desarrollo de esta disciplina. Nuestro trabajo futuro está encaminado de esta manera.

REFERENCIAS BIBLIOGRÁFICAS

Altamimi, A., Alotaibi, A. & Alruban, A. (2019). Surveying the Development of Authorship Identification of Text Messages. *International Journal of Intelligent Computing Research (IJICR), 10*(1), 953-966.

Amuchi, F., Al-Nemrat, A., Alazab, M. & Layto, R. (2013). Identifying Cyber Predators through Forensic Authorship Analysis of Chat Logs. *Proceedings of Third Cybercrime and Trustworthy Computing Workshop* (pp. 27-38). Los Alamitos, CA: IEEE.

Aitken, C. & Taroni, F. (2010). *Estadística y evaluación de la evidencia para expertos forenses*. Madrid: Dykinson, SL.

Boenninghoff, B., Nickel, R., Zeiler, D. & Kolossa, D. (2019). Similarity Learning for Authorship Verification in Social Media. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 2457-2461). Brighton, United Kingdom: IEEE.

Castillo Velásquez, F. A, Martínez Godoy, J. L., Torres Falcón, M. C. P., Zavala de Paz, J. P., Becerra Chávez, A. & Rizzo Sierra, J. A. (2020). Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático. *Research in Computing Science*, 149, 91-101.

Corpas, G. & Seghiri, M. (2007). Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor. *Procesamiento del Lenguaje Natural*, 39, 165-172.

Coulthard, M. & Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. Londres & Nueva York: Routledge.

Crespo, M. (2016). Analysis of Parameters on Author Attribution of Spanish Electronic Short Texts. *RiCL*, 4, 25-32.

Crystal, D. (1997). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Delgado, C. (2002). *La identificación de locutores en el ámbito forense*. Tesis doctoral, Universidad Complutense de Madrid, Madrid, España.

Ebrahimpour, M., Putniņš, T. J., Berryman, M. J., Andrew, A., Ng, B. W.-H. & Abbott, D. (2015). Automated Authorship Attribution Using Advanced Signal Classification Techniques. *PLOS ONE*, *8*(2), 1-12.

Elmanarelbouanani, S. & Kassou, I. (2014). Authorship Analysis Studies: A Survey. *International Journal of Computer Applications*, *86*(12), 22-29.

Evans, J. D. (1996). *Straightforward Statistics for the Behavioral Sciences*. Calif.: Brooks/Cole Publishing: Pacific Grove.

Fernández Trillo, J. J. (2015). *La problemática autoría de Carlos Castaneda. Un estudio estilométrico de candidato único*. Tesis Doctoral, Universidad Autónoma de Madrid, Madrid, España.

Garayzábal Heinze, E., Jiménez Bernal, M. & Reigosa Riveiros, M. (2014). Glosario básico para entender la Lingüística Forense. *Lingüística forense: La lingüística en el ámbito legal y policial* (pp. 411-417). Madrid: Euphonia Ediciones.

Garayzábal Heinze, E., Queralt Estévez, S. & Reigosa Riveiros, M. (2019). *Fundamentos de la lingüística forense*. Madrid: Síntesis.

Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E. & Stein, B. (2013). Recent Trends in Digital Text Forensics and Its Evaluation. En P. Forner, H. Müller, R. Paredes, P. Rosso & B. Stein (Eds.), *Information Access Evaluation. Multilinguality, Multimodality, and Visualization. CLEF 2013. Lecture Notes in Computer Science* (pp. 282-302). Berlin, Heidelberg: Springer.

Goodman, R., Hahn, M., Marella, M., Ojar, C. & Westcott, S. (2007). The Use of Stylometry for Email Author Identification: A Feasibility Study. *Proceedings of Student/Faculty Research Day, CSIS* (pp. 1-7). Nueva York: Pace University.

Göhring, A. (2009). *Spanish Expansion of a Parallel Treebank* (Lizentiatsarbeit). Zurich: University of Zurich.

Halvani, O., Winter, C. & Pflug, A. (2016). Authorship Verification for Different Languages, Genres and Topics. *Digital Investigation*, 16, 33-43.

Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, *13*(3), 111-117.

Ishihara, S. (2014). A Likelihood Ratio-Based Evaluation of Strength of Authorship Attribution Evidence in SMS Messages using N-grams. *International Journal of Speech, Language & the Law*, *21*(1), 23-50.

Jamak, A., Savatić, A. & Can, M. (2012). Principal component analysis for authorship attribution. *Southeast Europe Journal of Soft Computing, 1*(1), 49-56.

Jiménez, M. (2014). La Lingüística forense: Licencia para investigar la lengua. *Lingüística forense: La lingüística en el ámbito legal y policial* (pp. 79-94). Madrid: Euphonia Ediciones.

Juola, P. (2008). Authorship Attribution. *Foundations and Trends in Information Retrieval*, *1*(3), 233-334.

Koehler, J. J. (2013). Linguistic Confusion in Court: Evidence from the Forensic Sciences. *Journal of Law & Policy*, *21*(2), 515-539.

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I. & Demidov, P.G. (2019). A Survey on Stylometric Text Features. *25th Conference of Open Innovations Association*. Helsinki: IEEE.

Lalla, H. (2010). *E-mail Forensic Authorship Attribution*. Tesis de Magíster, Universidad de Fort Hare, Bhisho, Sudáfrica.

Li, C. (2010). *Handbook of Research on Computational Forensics, Digital Crime, and Investigation: Methods and Solutions. Information Science Reference*. Hershey PA: IGI Global.

McMenamin, G. R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics*. Washington: CRC Press.

Mosteller, F. & Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Londres: Addison-Wesley.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y. & Woodard, D. (2017). Surveying Stylometry Techniques and Applications. *ACM Computing Surveys*, *50*(6), 1-36.

Nini, A. (2020). Corpus Analysis in Forensic Linguistics. En C. A. Chapelle (Ed.), *The Concise Encyclopedia of Applied Linguistics* (pp. 313-320). Hoboken: Wiley-Blackwell.

Olsson, J. (2004). *Forensic Linguistics. An Introduction to Language, Crime and the Law*. Londres: Continuum.

Picornell, I. (2014). La aplicación de la atribución de autoría en la investigación e inteligencia: La aplicación práctica (y su problemática). *Lingüística forense: La lingüística en el ámbito legal y policial* (pp. 79-94). Madrid: Euphonia Ediciones.

Queralt Estévez, S. (2014). Acerca de la prueba lingüística en atribución de autoría hoy. *Revista de Llengua i Dret*, 62, 35-48.

Queralt Estévez, S. (2019). The Creation of Base Rate Knowledge of Linguistic Variables and the Implementation of Likelihood Ratios to Authorship Attribution in Forensic Text Comparison. *Language and Law= Linguagem e Direito*, *5*(2), 59-76.

Ramírez Salado, M. (2017). Antecedentes de la lingüística forense: ¿Desde cuándo se estudia el lenguaje como evidencia? *Pragmalingüística*, 25, 525-539.

Ramírez Salado, M. (2019). *Terminología y lingüística forense: Usos terminológicos relacionados con los ámbitos de actuación de la lingüística forense y su interfaz con otras disciplinas*. Tesis doctoral, Universidad de Cádiz, Cádiz, España.

Reddy, T. R., Vardhan, B. V. & Reddy, P. V. (2016). A Survey on Authorship Profiling Techniques. *International Journal of Applied Engineering Research*, *11*(5), 3092-3102.

Rico Sulayes, A. (2012). *Quantitative Authorship Attribution of Users of Mexican Drug Dealing Related Online Forums*. Tesis doctoral, Universidad de Georgetown, Georgetown, Estados Unidos.

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A. & Stamatatos, E. (2017). Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*, *12*(1), 5-33.

Rose, P. (2002). *Forensic Speaker Identification*. Nueva York: Taylor & Francis.

Salih, R., Balci, K. & Salah, A. (2016). Authorship Recognition in a Multiparty Chat Scenario. *Proceedings of 4th International Conference on Biometrics and Forensics* (IWBF) (pp. 1-6). Cyprus: IEEE.

Sapkota, U., Bethard, S., Montes, M. & Solorio, T. (2015). Not all Character N-grams are Created Equal: A Study in Authorship Attribution. En *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies,* 93-102. Berkeley, CA: ACL.

Savoy, J. (2016). Estimating the Probability of an Authorship Attribution. *Journal of the Association for Information Science and Technology*, *67*(6), 1462-1472.

Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Vol. 12, Manchester, United Kingdom.

Sousa-Silva, R. (2019). Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts. *Language and Law= Linguagem e Direito, 5*(2), 118-143.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556.

Stańczyk, U. & Cyran, K. A. (2007). Machine Learning Approach to Authorship Attribution of Literary Texts. *International Journal of Applied Mathematics and Informatics, 1*(4), 151-158.

Stefanova Spassova, M. (2009). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español* Tesis doctoral, Universitat Pompeu Fabra, Barcelona, España.

Swain, S., Mishra, G. & Sindhu, C. (2017). Recent Approaches on Authorship Attribution Techniques —*An Overview. International Conference on Electronics, Communication and Aerospace Technology*, 557-566.

Triola, M. (2013). *Estadística* (11ª ed.). Naucalpan de Juárez: Pearson Educación de México.

Turell, M. T. (2008). Plagiarism. En J. Gibbons & M. T. Turell (Eds.), *Dimensions of Forensic Linguistics* (pp. 265-299). Ámsterdam & Filadelfia: John Benjamins.

Tweedie, F. J., Singh, S. & Holmes, D. I. (1996). Neural Network Applications in Stylometry: The Federalist Papers. *Computers and the Humanities, 30*(1), 1-10.

Villayandre Llamazares, M. (2008). Lingüística con corpus (I). *Estudios Humanísticos. Filología*, 30, 329-349.

NOTAS

[1] En la geometría un hiperplano es un subespacio de una dimensión menor que su espacio ambiente. De ese modo si un espacio es tridimensional, entonces, sus hiperplanos son los planos de 2 dimensiones, mientras que si el espacio es bidimensional, sus hiperplanos serán en una dimensión.

[2] El corpus puede consultarse libremente en https://ila.uca.es/laboratorio-de-linguistica-computacional/

[3] Un n-grama es una subsecuencia de n elementos de una cadena dada, los bigramas incluirían dos elementos (en nuestro estudio, caracteres). Sapkota, Bethard, Montes y Solorio (2015) indican que los n-gramas de caracteres son uno de los rasgos con más éxito en análisis de autoría. La razón hay que verla en el hecho de que permiten capturar elementos de diferentes niveles lingüísticos: morfológico, léxico, sintáctico e incluso el estilo.

[4] Un lema es un término que representa y unifica todos los elementos de un conjunto de palabras morfológicamente similares. Representa una entrada de diccionario sin flexión. El lema de las unidades 'cantaba', 'canto', 'hemos cantado' sería 'cantar'.