EVALUACIÓN Y CUANTIFICACIÓN DE ALGUNAS TÉCNICAS DE "ATRIBUCIÓN DE AUTORÍA" EN TEXTOS ESPAÑOLES

JAVIER BLASCO
CRISTINA RUIZ URBÓN
UNIVERSIDAD DE VALLADOLID

1. INTRODUCCIÓN

La gran cantidad de textos españoles de los Siglos de Oro que han llegado a nosotros de forma anónima o bajo seudónimo ha provocado, desde hace varias centurias, infinidad de propuestas de autoría para tales textos. Sirva como ejemplo lo ocurrido con *El Lazarillo de Tormes*, obra que, desde que en 1605 José de Sigüenza se la atribuyese a fray Juan de Ortega, ha conocido –con mayor o menor éxito– muy diferentes propuestas de paternidad (bien es verdad que con diferente fortuna y con distinto rigor filológico), entre las que cabe destacar los nombres de Diego Hurtado de Mendoza, Juan de Valdés, Alfonso de Valdés, Sebastián de Horozco, Lope de Rueda, Pedro de Rúa, Francisco Cervantes de Salazar o Arce de Otalora, entre otros. Y la lista sigue abierta.

Algo parecido ha ocurrido con el *Quijote* de Avellaneda y, en fin, con muchas otras obras en prosa de interés relevante en el canon literario español de los Siglos de Oro, con numerosísimas piezas de teatro de primer nivel y con no poca poesía de los cancioneros y del romancero. Las propuestas de autoría, impulsadas por el comprensible

prurito del investigador de alzarse con el dudoso honor de llegar a unir su nombre con la resolución de un misterio, han dado aliento a estudios no siempre compensadores de los esfuerzos invertidos por el investigador ni por los que éste, al hacer público su trabajo, demanda de los lectores.

Aunque los estudios de atribución son tan viejos como la literatura misma, sólo desde hace algunas décadas se ha realizado un esfuerzo digno de elogio por incorporar las fórmulas y los procedimientos que, desde la década de los 60, se han ido consolidando en el terreno de la *Lingüística forense*, sobre todo en el ámbito de la filología anglosajona; procedimientos que, desde luego, han supuesto una apuesta interesante de cara a la construcción de una metodología fiable (por su capacidad de objetivación de los fenómenos observados) y al establecimiento de unos protocolos de actuación en los que se redujese notablemente el espacio concedido a la subjetividad o a la elección arbitraria por parte del analista de los fenómenos considerados para hacer atribución.

En esencia, los fundamentos teóricos de los estudios de atribución a los que nos estamos refiriendo resultan bastante claros y de difícil contestación: un autor, al escribir, produce un discurso que indefectiblemente —salvo manipulaciones específicamente concebidas para ocultar ciertas huellas— presenta toda una serie de marcas verbales, patrones de escritura, que lo caracterizan e identifican frente a otros discursos salidos de diferente mano. El convencimiento de que cada hablante/escritor posee un idiolecto propio y exclusivo —esto es, unos hábitos personales que se escapan a lo consciente—, es lo que ha propiciado que de un tiempo a esta parte se trabaje en el establecimiento de unos algoritmos de atribución fiables que, ante un texto anónimo o sólo respaldado por un seudónimo, permitan proponer una autoría con cierta seguridad.

Por lo general, los recientes estudios de "atribución de autoría" en el ámbito de la literatura española de los Siglos de Oro parten de la convicción anterior y son bastantes ya los trabajos que, en el mejor de los casos, pretenden aplicar alguno de los muchos protocolos metodológicos desarrollados desde la *lingüística forense* anglosajona para determinar la autoría de un texto anónimo o trasmitido bajo seudónimo. Sin embargo, todavía se observa en ellos ciertos vicios de método que, groso modo, pueden agruparse en dos:

a) la elección caprichosa de la porción de texto a analizar y de la aplicación de los parámetros de medición que más convengan, sin otro criterio que la subjetividad del analista. El responsable de la investigación tiende a seleccionar ciertos fragmentos textuales y a dar protagonismo a determinados fenómenos verbales, en detrimento de otros, únicamente en función de la candidatura de atribución que le interesa defender. Se debe desconfiar de todos aquellos análisis realizados sobre porciones "seleccionadas" de texto; y, consecuentemente, se debe poner en tela de juicio los datos resultantes cuando los fenómenos sometidos a análisis se deciden "un poco arbitrariamente".

b) El desaprovechamiento de los recursos y herramientas que nos ofrece la *Lingüística computacional*. Si bien hoy día es habitual trabajar sobre *corpus* textuales digitalizados y bases de datos informatizadas (como el CORDE, que resulta una herramienta extraordinaria para estudios atributivos de textos antiguos, a pesar de que los resultados puedan verse modificados sensiblemente por el hecho de que el *corpus* sobre el que trabaja no responda a unos criterios textuales unificados y regulados), rara vez se observa el aprovechamientos de analizadores lingüísticos para el procesamiento de los lenguajes naturales, que, sin embargo, sí han demostrado tener un gran valor precisamente por su capacidad de tratar el texto de manera totalmente objetiva y por la posibilidad que ofrecen de traducir las peculiaridades de los mismos a magnitudes cuantificables y mensurables.

La "parcialidad" del analista, tanto a la hora de seleccionar las variables que su estudio va a contemplar como a la hora de utilizar las herramientas existentes para traducir a lenguaje numérico y porcentual los resultados, va en perjuicio de la objetividad y la credibilidad de muchos de los estudios de atribución que para textos literarios hispánicos se han hecho hasta la fecha.

Llegados a este punto, se impone la necesidad de abandonar las "selecciones" arbitrarias de los fenómenos que se van a estudiar, las "porciones" de los textos sobre los que se va a trabajar hechas "sin un criterio definido" y las cuantificaciones obtenidas por la *cuenta de la vieja*, para pasar a establecer unos protocolos de actuación objetivos y acreditados. Por lo que se refiere a los textos de nuestros Siglos de Oro que tanto siguen intrigándonos por el misterio que envuelve a su

paternidad, antes de poder poner en práctica cualquier técnica de atribución se impone la necesidad de conocer con precisión el grado de seguridad y de fiabilidad que ofrecen los métodos y procedimientos de medición textual con los que abordar en ellos el tema de la autoría. Y, junto a todo lo anterior, debe seguirse reclamando un profundo conocimiento filológico del texto en cuestión, tanto de sus peculiaridades histórico-literarias como de textualidad, tanto en lo que se refiere a su gestación como a su transmisión.

Es de la constatación de la necesidad de contar con métodos fiables de la que arranca nuestro trabajo, que no tiene otra pretensión que la de someter a juicio algunos de los algoritmos atributivos más utilizados hasta la fecha, con el fin de evaluar cuantitativamente el grado de confianza de cada uno de ellos. Partiremos para ello del test realizado por Jack Grieve, de Northern Arizona University (Grieve, 2007). En este interesante trabajo, Grieve evalúa 39 tipos o procedimientos de medición textual, aplicándolos a un *corpus* formado por textos de una serie de columnistas del *Telegraphe* londinense, para determinar los mejores indicadores de la autoría para la lengua inglesa, en función del porcentaje de fiabilidad que se derivase de cada uno de ellos. Los resultados que se desprenden de su estudio son realmente interesantes(véase tabla 1): en ellos se constata que el porcentaje de acierto y fiabilidad de un determinado procedimiento de medición textual se reduce drásticamente conforme crece el número de autores potenciales, pues, a modo ilustrativo, cabe destacar que medidas tan altamente fiables para 2 candidatos, como el *perfil de grafema* (*Grapheme profile*), pueden descender enormemente cuando el número de candidatos sube a 40; de este modo, y siempre según este análisis, el "top ten" de las mediciones evaluadas alcanza o supera el 88 % de fiabilidad, cuando se trata de discriminar entre dos autores, pero baja al 80, 76, 72, 58, 46 y 34% cuando se confronta el texto anónimo con un *corpus* de 3, 4, 5, 10, 20 y 40 autores, respectivamente.

Textual measurement (Variant)	Possible authors						
	40	**20**	**10**	**5**	**4**	**3**	**2**
Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94
3-gram profile (10-limit)	61	72	78	85	88	91	94
4-gram profile (10-limit)	55	64	73	83	85	89	93
Grapheme and punctuation mark profile	50	60	70	81	84	87	93
Word profile (5-limit)	48	57	67	77	80	85	88
5-gram profile (10-limit)	47	55	66	76	79	84	90
Multiposition grapheme profile (first 6 in word)	43	53	64	76	79	84	90
Multiposition grapheme profile (last 6 in word)	42	52	63	74	79	83	90
Punctuation mark profile (by character)	34	46	58	72	76	80	89
6-gram profile (10-limit)	35	45	56	68	72	78	86
Word-internal grapheme profile	28	39	51	65	70	76	85
Single-position grapheme profile (last in word)	27	36	49	63	68	73	84
Grapheme profile	25	35	47	62	67	74	83
7-gram profile (2-limit)	34	42	45	59	64	69	81
Single-position graph profile (2^{nd} to last in word)	23	31	43	57	63	70	81
Single-position grapheme profile (1^{st} in word)	20	30	41	56	62	69	80
Multiposition word profile (first 4 in sentence)	22	31	41	55	60	67	77
Word-length profile (15 intervals of one character)	18	26	39	54	60	68	79
Single-position word profile (1^{st} word in sentence)	17	30	36	50	56	64	75
8-gram profile (2-limit)	18	24	36	50	55	62	74
2-word collocation profile	17	24	34	48	54	61	74
Tuldava's LN	11	18	31	49	55	64	77
Sentence-length profile (12 intervals of 25 characters)	12	20	31	46	53	62	74
Sentence-length profile (10 intervals of 5 words)	10	17	28	44	50	59	73
9-gram profile (2-limit)	12	18	28	41	46	55	68
Type–Token ratio	8	16	27	44	51	61	75
Herdan's C	7	14	25	42	49	59	73
Guiraud's R	7	13	24	41	48	58	73
Average word-length	7	12	22	39	46	55	70
Average sentence-length (in characters)	6	12	22	39	45	53	70
Average sentence-length (in words)	6	11	21	37	44	53	69
Yule's K and Simpson's D	6	10	18	33	38	49	65
6-gram profile (10-limit)	35	45	56	68	72	78	86
Word-internal grapheme profile	28	39	51	65	70	76	85
Single-position grapheme profile (last in word)	27	36	49	63	68	73	84

Tabla 1: Fiabilidad de las 39 mediciones estudiadas por Grieve (%)

Situados en este punto, nuestro trabajo analizará si las mediciones más fiables, según el estudio de Grieve, siguen teniendo idéntica respuesta cuando se aplican a textos en español; y, del mismo modo, si los procedimientos menos fiables siguen ofreciendo resultados poco efectivos al aplicarlos a textos escritos en la lengua de Cervantes. Para responder a esta cuestión y poder establecer el *ranking* de las variables más altamente discriminantes en nuestro idioma, someteremos a examen algunos de los algoritmos atributivos que se han ido fijando hasta la fecha.

2. *CORPUS* DE TRABAJO

A la hora de seleccionar la muestra de textos del mundo real que habría de servir de soporte a nuestro estudio, siguiendo también en esto las pautas del trabajo de Grieve, hemos contemplado la necesidad de que todos ellos formasen parte de un misma categoría y que hubiesen sido redactados en una misma época, pues debemos ser conscientes de que la variación intra-genérica y la variación intra-cronológica pueden adulterar el resultado de las mediciones. Teniendo en cuenta estos dos factores, y con el fin de provocar una situación genérico-temporal homogénea, el *corpus* de trabajo seleccionado para nuestro experimento está formado por artículos de diez autores habituales en tres de los medios de comunicación escrita más representativos de nuestro país (*El País*, *El Mundo* y *ABC*), fechados entre noviembre del 2007 y agosto del 2009.

El proceso que hemos seguido para conformar ese *corpus* de análisis ha sido el siguiente: hemos fundido distintos artículos de cada uno de los 10 columnistas estudiados, hasta configurar un conjunto textual de, aproximadamente, unas 20.000 palabras para cada autor; a continuación, hemos dividido el *corpus* de cada autor en dos partes, según una proporción 80/20, en donde el 80% (unas 16.000 palabras) constituye el texto indubitado que nos servirá para identificar las marcas verbales características de su idiolecto, en tanto que el 20% restante (unas 4.000 palabras) pasará a formar parte (junto con el 20% del resto de los autores) del conjunto textual dubitado sobre el que aplicaremos los distintos tipos de medición. De este modo, el *corpus* de análisis está formado dos conjuntos textuales:

1) Conjunto de las muestras indubitadas.- La selección de los diez periodistas se ha hecho en función de la disponibilidad de un *corpus* de artículos suficiente para cada uno de ellos en la web del periódico para el que escriben.

	AUTOR	FUENTE	FECHA	*TOKENS*
1.	Antonio Elorza	*El País*	Sept. 08 – Ag. 09	19619
2.	Benjamín Prado	*El País*	Ag. 08 – Ag. 09	20117
3.	Fernando Savater	*El Mundo*	Feb. 08 – Ag. 09	21603
4.	Casimiro García-Abadillo	*El Mundo*	Oct. 08 – Abr. 09	23079
5.	Ignacio Sotelo	*El País*	Jun. 08 – Jul. 09	19491
6.	Lucía Méndez	*El Mundo*	Dic. 08 – Jul. 09	18164
7.	Pedro J. Ramírez	*El Mundo*	Jun. 09 – Ag. 09	22570
8.	Victoria Prego	*El Mundo*	Abr. 09 – Ag. 09	20928
9.	Ignacio Camacho	*ABC*	Jul. 09 – Ag. 09	22272
10.	Juan Manuel Prada	*ABC*	Nov. 07 – Ag. 09	19732

Tabla 2: Corpus de textos indubitados

Los diez columnistas, aunque con un posicionamiento ideológico distinto, se acogen a un contexto de opinión semejante, abordan una serie de temas recurrentes (generalmente referentes al ámbito político) y escriben en un tiempo cronológico acotado. Es precisamente en un contexto igualatorio, como éste, donde los rasgos distintivos e idiosincrásicos de cada uno de ellos deberían poder verse con mayor objetividad.

2) Conjunto de las muestras dubitadas.- Las muestras dubitadas son en realidad falsos anónimos, pues, como ya hemos explicado, provienen de la anonimización consciente, pero no condicionada, de una pequeña porción de la muestra indubitada de cada candidato a autor. Los diez textos resultantes, a los que se les ha asignado un membrete alfanumérico aleatorio (d1-d10), son los que detallamos a continuación en la tabla 3:

TEXTO	*TOKENS*
d1	4261
d2	5064
d3	3318
d4	3698
d5	3970
d6	3684
d7	3878
d8	3697
d9	3995
d10	3792

Tabla 3: Corpus de textos dubitados

3. METODOLOGÍA

3.1. Establecimiento de los algoritmos de atribución:

Los algoritmos de atribución en que basaremos nuestro estudio giran en torno a cinco realidades textuales, cuya funcionalidad para la medición textual con intención atributiva cuenta ya con una bibliografía interesante (véase la relación en el apartado final): las marcas de puntuación, las palabras, los grafemas, la frecuencia de *n-gramas* y la riqueza de vocabulario.

3.1.1. *Frecuencia de marcas de puntuación*.- En el trabajo de Grieve se contemplan cinco mediciones de frecuencia de puntuación: las tres primeras son variantes del *perfil de marca de puntuación simple (simple punctuation mark profile)*, según se calcule la frecuencia relativa de una serie de signos de puntuación en función del número total de caracteres, signos o palabras que contiene un determinado texto, y las otras dos son combinaciones del perfil de puntuación con los perfiles de grafema (*punctuation and grapheme profile*) y de palabra (*punctuation and word profile*):

a) El *perfil de marca de puntuación simple 1* se calcula dividiendo la frecuencia de una serie de signos de puntuación entre el número total de caracteres contenidos en el texto.
b) El *perfil de marca de puntuación simple 2* se obtiene dividiendo la frecuencia de cada uno de los signos de puntuación seleccionados entre el cómputo total de los que se contabilizan en el texto.
c) El *perfil de marca de puntuación simple 3* se calcula dividiendo la frecuencia de las marcas de puntuación escogidas entre el número total de palabras del texto.
d) El *perfil de puntuación y palabra* contempla y mide la frecuencia del conjunto formado por un signo de puntuación dado y la palabra que lo precede o lo sigue.
e) El *perfil de puntuación y grafema* mide la frecuencia del conjunto formado por un signo de puntuación dado y el grafema que lo precede o lo sigue.

Para nuestro test hemos optado por el *perfil de marca de puntuación simple1* al ser esta medición la que ofrece mayores garantías en la lengua de Shakespeare, pues mientras las variantes 2 y 3 ofrecen un 53 y un 57%, respectivamente, para discriminar entre diez autores, la variante 1 llega al 58%; por otro lado, hemos seleccionado la combinatoria del *perfil de puntuación y palabra* por ser la medición más discriminante de todas las analizadas por Grieve, llegando a alcanzar el 95% a la hora de discernir entre dos autores.

3.1.2. *Frecuencia de palabra*.- Grieve somete a prueba tres mediciones diferentes de frecuencia de palabra:

a) El *perfil de palabra simple* (*simple word profile*) se define como la frecuencia relativa de una serie de términos de alta frecuencia y se calcula dividiendo la frecuencia de una determinada palabra entre el número total de palabras del texto.
b) El *perfil de palabra en posición sencilla* (*single-position word profile*) mide la frecuencia relativa de una serie de palabras que aparecen en una posición particular dentro de las oraciones de un texto (por ejemplo, primera palabra, segunda palabra… o última palabra dentro de la oración), y se obtiene de dividir la frecuencia de una palabra en una posición seleccionada por el número de oraciones que contienen dicha posición.

c) El *perfil de palabra en multiposición* (*multi-position word profile*) recoge las mediciones de diversas palabras en posición individual múltiple (por ejemplo, las primeras cuatro palabras de una oración).

De las tres mediciones de frecuencia de palabra que describe Grieve en su artículo, únicamente hemos contemplado en nuestro estudio el *perfil de palabra simple* por ser la que en la lengua inglesa ofrece una mayor efectividad.

3.1.3. *Frecuencia de grafema*.- Grieve prueba cuatro tipos diferentes de frecuencias de grafemas:

a) El *perfil de grafema simple* consiste en la frecuencia relativa del total de grafemas de la lengua a analizar (26 en el caso del alfabeto inglés), que, al igual que en el caso del perfil de palabra simple, se calcula dividiendo la frecuencia de ese grafema en el texto por el número total de grafemas.

b) El *perfil de grafema en posición sencilla* hace referencia a la frecuencia relativa de los grafemas que aparecen en una posición particular dentro de las palabras de un texto (por ejemplo, primer grafema, segundo grafema o último grafema de la palabra) y se calcula dividiendo la frecuencia que tiene cada grafema en esa posición entre el número total de palabras que contienen esa posición.

c) El *perfil de grafemas en multiposición* debe entenderse como el perfil de grafemas en posición individual múltiple (por ejemplo, los tres primeros grafemas de una palabra).

d) El *perfil de grafema en interior de palabra*, que consiste en el porcentaje de palabras de un texto que contiene cada uno de los grafemas de la lengua a estudiar, y se calcula dividiendo el número de palabras en un texto que contienen al menos un caso de ese grafema por el número total de palabras.

En nuestro análisis hemos contemplado la tercera medida, dejando fuera las otras tres por el simple hecho de que el español es una lengua no fonémica que contiene algunos dígrafos —es decir, fonemas de dos grafemas (*ch*, *ll*, *qu*, *gu* y *rr*)—, lo que podría desvirtuar los resultados.

3.1.4. *Frecuencia de colocación, tanto a nivel de palabra como de grafema* (*N-gramas*).- La medición *n-gramas* recoge la frecuencia relativa de una determinada cadena de *n* grafemas o de *n* palabras, dividiendo la frecuencia de aparición de esa secuencia concreta de *n-gramas* a analizar entre el número total de secuencias existentes de esos *n-gramas*. Por ejemplo: 2-gramas de *es* = Frecuencia de la secuencia *es* / Total de secuencias de 2 caracteres.

Para nuestro estudio hemos analizado la frecuencia de un total de 6 mediciones: perfil de 2 a 5 gramas de grafema y perfil de 2 a 3 gramas de palabra.

3.1.5. *Riqueza de vocabulario*.- De los 11 tipos de mediciones que ofrece Grieve, nosotros sólo hemos contemplado la *ratio type/token* y las mediciones ofrecidas por las fórmulas de Honoré y Yule. A pesar de que los algoritmos referentes a la riqueza de vocabulario constituyen algunas de las medidas menos efectivas en el caso del inglés, principalmente cuando hay que discriminar entre 5 o más autores, en nuestro deseo por confrontar y comparar los resultados porcentuales de los algoritmos de atribución en las dos lenguas, consideramos tan interesante ver si se mantienen los altos porcentajes como si lo hacen los bajos.

En definitiva, las 15 variables contempladas en nuestro análisis, bien por ser las más operativas o las más ineficaces en la lengua inglesa, son las que señalamos a continuación:

▪ *Perfil de marca de puntuación simple 1*.- Hemos determinado, a través del programa de análisis textual *WordSmith Tools*, las frecuencias relativas de cuatro marcas de puntuación —punto (.), coma (,), punto y coma (;) y dos puntos (:)—, dividiendo el número de veces que aparece cada una de ellas entre el número total de caracteres del texto.

▪ *Perfil de puntuación y palabra (puntuación + palabra)*.- Para calcular esta medición, que combina el perfil de puntuación con el de palabra, hemos optado por unificar cualquier signo de puntuación —a pesar de que presuponemos más rentable el análisis distintivo de cada marca— por el hecho de que, en un texto del Siglo de Oro (y en última instancia este trabajo arranca de nuestra preocupación por su

literatura) la diferenciación entre los diferentes signos de puntuación muy raramente es responsabilidad del autor.

▪ *Perfil de palabra simple*.- Para computar este algoritmo, se ha contrastado la frecuencia relativa de ocho de las palabras más frecuentes en los textos dubitados (*de, la, que, el, en, y, a, los*) con las frecuencias que en cada uno de los autores tienen esas palabras.

▪ *Perfil de grafema en multiposición*.- Se ha seleccionado el análisis de los 6 grafemas iniciales de palabra y se ha descartado el de los seis últimos porque, aunque para la lengua inglesa resulta una medición altamente operativa, consideramos que la flexión del español podría alterar los resultados. Para este cálculo nos hemos ayudado del programa *KfNgram* (versión 2002-2007) de William H. Fletcher.

▪ *Perfil de n-grams*.- A través de *KfNgram*, hemos calculado la frecuencia relativa de 2 y 3 gramas de palabra y de 2, 3, 4 y 5 gramas de grafema. La frecuencia relativa de 2-gramas de grafema se calcula dividiendo la frecuencia de una determinada secuencia formada por dos grafemas entre el número total de combinaciones de dos grafemas; y así, con el resto de caso.

▪ *Ratio type/token*.- La *ratio type/token* se calcula dividiendo el número de *types* (formas o palabras diferentes) entre el número de tokens (palabras totales). Si tenemos en cuenta que la lista de palabras funcionales vacías de significado (*closed set*) es mucho más reducida que la que constituye la lista de palabras significativas de una lengua (*open set*), advertiremos que la *ratio type/token* será desproporcionada en función de la longitud de los textos, pues, a partir de una determinada extensión, las palabras funcionales tenderán a repetirse pero no necesariamente las palabras lexicales. Por ello, para calcular la *ratio type/token*, tanto de los dubitados como de los indubitados, hemos tomado muestras con las primeras 500 palabras de cada texto, pues de otro modo el análisis se vería condicionado a la longitud de los textos[1].

[1] Grieve aplica esta medida a las 139 primeras palabras de los textos, pero nosotros hemos preferido seleccionado una porción algo más elevada.

Hemos establecido 3 niveles de análisis distintos: la *ratio type/token* global, la *ratio type/token* de las palabras lexicales y la *ratio type/token* de las funcionales.

▪ *Riqueza de Honoré* y *Yule*.- A través del programa *Vocalyse Toolkit* (JVocalyse v 2.05), diseñado por David Woolls, hemos podido obtener de manera inmediata los resultados de las dos mediciones de *richness* que hemos cotejado: Honoré y Yule.

La fórmula de Honoré (H) calcula la probabilidad de que un determinado autor vuelva a usar en un texto una forma ya dada antes que una nueva:

$$\text{Honoré (H)} = \frac{100 \log N}{1 - (V_1 / V)}$$ donde N (tokens) = Nº total de palabras del texto; V (types) = Nº total de formas del texto; V_1 = Nº types usados una sola vez

La fórmula de Yule (K) mide la riqueza de vocabulario de un texto en función de la ratio de repetición léxica:

$$\text{Yule (K)} = \frac{10^4 (\sum_i^2 V_i - N)}{N^2}$$ donde N (tokens) = Nº total de palabras del texto; V_i = Nº de types que ocurren *i* veces

La *riqueza de Honoré* se ha calculado sobre la totalidad del *corpus* de análisis mientras que la de *Yule* se ha aplicado sobre las muestras reducidas de 500 palabras, pues, a pesar de que está pensada para no verse afectada por la longitud del texto, lo cierto es que el conjunto de los dubitados (de menor longitud) daba resultados mucho más elevados que el conjunto de los indubitados (de mayor extensión).

3. 2. Comparación de los resultados

A la hora de contrastar los resultados obtenidos, y con la intención de ser rigurosamente objetivos, hemos establecido dos criterios de confrontación:

1) Distancia entre dos valores (χ) → Para contrastar las variables con una única medida, calcularemos la distancia existente entre el valor observado en un texto dubitado y el valor esperado en cada una de las indubitadas; el resultado de menor diferencia, es decir, el que más se aproxime a cero, nos ofrecerá el nombre del candidato que en esa medición más se asemeja al autor del texto anónimo.

$$\chi = \text{Valores observados} - \text{Valores esperados}$$

2) *Chi cuadrado* (χ^2) → Para contrastar las variables que combinan varios valores, utilizaremos el test del *chi squared* de Pearson. El χ^2 es una prueba estadística no paramétrica que permite determinar de modo totalmente objetivo si un determinado fenómeno, representado por una serie de frecuencias observadas, podría haber sido producido por alguno de los candidatos seleccionados, representados a su vez por una serie de frecuencias esperadas:

$$\chi^2 = \sum ((\text{Valores observados} - \text{esperados})^2 / \text{Valores observados})$$

De nuevo, el resultado del autor que más tienda a cero será el que más se parezca al del anónimo analizado.

3.3. Resultados obtenidos / Interpretación de los resultados

Los resultados obtenidos de las 15 mediciones analizadas pueden visualizarse en la tabla adjunta, en donde se señala el número del autor (o los números de los autores) que cada medición ha discriminado en primera posición (*X* cuando no discrimina a ninguno de ellos) y se marca en sombreado los casos de acierto; en la última columna se recoge el porcentaje de fiabilidad de cada variable y en la última fila la proporción de acierto para cada autor:

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	
Marca de puntuación simple1	9	1 3	5	6	8	1	9	2	4	10	**85%**
Puntuación + palabra	7	3	5	6	4	6	6	5	4	10	**60%**
Palabra simple	7	10	5	8	8	1	9	6	4	10	**70%**
Grafema multiposición (6 inic.)	7	3	5	6	6 8	8	9	2	6	3	**65%**
2-gramas de palabra	7	3 5	5	5	2 8	1	1	2 8	4 5	10 4	**55%**
3-gramas de palabra	7	8	X	6 4	8	1	5	2	2	X	**45%**
2-gramas de grafema	1	2	1	6	2	1	1 5 9	7	1	9	**23%**
3-gramas de grafema	5	3	5	2	2	1	5	1	6	1	**30%**
4-gramas de grafema	5 7 10	4	5	7	8	7	9	2 8 9	6	4	**36%**
5-gramas de grafema	7	3	6	5	7	1	9	2	6	4 5	**50%**
Type/token (limit. 500)	2	8	2	2	5	10	2	10	2	6 8	**0%**
Type/token léxica (limit. 500)	7 9	4 8	4 8	5	5	7 9	7 9	1	1	4 6 8	**10%**
Type/token funcional (limit. 500)	7	7 10	8	1	7	5	9	10	8	8	**20%**
Honoré	9	7	2	2	1 5	9	9	7	6	10	**20%**
Yule's K (limit. 500)	53	5	5	9	9	6	2	3	9	7	**10%**
	8/15	**5/15**	**8/15**	**4,5/15**	**5/15**	**7/15**	**7,8/15**	**4,8/15**	**3,5/15**	**4,5/15**	

Tabla 4: Resultados de las mediciones analizadas

[Recordemos el número de identificación de cada autor: 1. Antonio Elorza; 2. Benjamín Prado; 3. Fernando Savater; 4. Casimiro García-Abadillo; 5. Ignacio Sotelo; 6. Lucía Méndez; 7. Pedro J. Ramírez; 8. Victoria Prego; 9. Ignacio Camacho; y 10. José Manuel Prada]

En la siguiente tabla se recoge el *ranking* con los seis algoritmos de atribución que, según nuestro experimento, resultan más fiables en lengua española, señalando el grado de fiabilidad que ofrecen (tanto para discriminar entre *corpus* de dos como de diez autores) y su confrontación con el porcentaje de acierto que tenían en la lengua inglesa:

		LENGUA ESPAÑOLA		LENGUA INGLESA	
	Algoritmos de atribución	**%**		**%**	
		10	**2**	**10**	**2**
1	Marca de puntuación simple1	85	90	58	89
2	Palabra simple	70	90	67	88
3	Grafema multiposición (6 inic.)	65	70	64	90
4	Puntuación + palabra	60	70	80	95
5	2-gramas de palabra	55	80	34	74
6	5-gramas de grafema	50	60	66	90

Tabla 5: Cotejo de la fiabilidad del ranking español con el inglés (%)

Los resultados que se derivan de nuestro experimento (tablas 5 y 6) y su confrontación con los resultados de Grieve (tabla 1) permiten establecer una serie de conclusiones en torno a dos cuestiones:

1. Conclusiones relativas a los tipos de medición textual de cara a establecer una propuesta de atribución.- Al igual que ocurría en el caso del inglés, el porcentaje de fiabilidad de las variables sometidas a análisis asciende conforme descienden los posibles candidatos a autor. De la observación exhaustiva de los porcentajes de acierto de las mediciones realizadas para discriminar entre diez autores, se desprende:

a) que el porcentaje de acierto del *perfil de marca de puntuación simple1* se sitúa en un 85% mientras que el del *perfil de puntuación + palabra* desciende al 60%, tal vez por haber calculado la frecuencia (para esta medición) unificado todos los signos de puntuación y no haber hecho la distinción por separado de cada uno de ellos;

b) que el grado de confianza del *perfil de palabra simple* se sitúa en el 70% y que el del perfil de *grafema en multiposición (6*

iniciales) se sitúa en un digno 65%, acordes con el 67 y 64% del inglés;

c) que la fiabilidad de los *perfiles de n-gramas de grafemas* crece conforme aumenta el número de grafemas analizados (23% para 2, 30% para 3, 36% para 4 y 50% para 5), algo que contrasta con lo que ocurría en lengua inglesa, en donde el perfil más fiable era *2-gram profile* (79%) y el menos *9-gram profile* (18%);

d) que los *perfiles de n-gramas de palabras* tienen un porcentaje de credibilidad medio, en torno al 50 % (concretamente, 55% para los *2-gramas* y 45% para los *3-gramas*);

e) y que las medidas de riqueza de vocabulario no son de modo alguno algoritmos de atribución fiables a la hora de discriminar entre 10 autores, ni en lengua española ni en lengua inglesa, y que tal vez debieran probarse nuevas fórmulas que jugasen con la distinción entre palabras de contenido y palabras de función[2].

2. Estimaciones concernientes a los posibles candidatos de autor.- La proporción de algoritmos que discriminan correctamente a cada autor nos permite establecer dos reflexiones:

a) se puede concluir que los autores en los que el mayor número de mediciones textuales ofrece una respuesta positiva, Fernando Savater, Antonio Elorza y Pedro J. Ramírez son autores con un modo de expresión mucho más idiosincrásico y personal que aquellos otros para los que los métodos de medición textual resultan menos eficaces, Casimiro García-Abadillo, Ignacio Camacho y Juan Manuel Prada;

b) y que los columnistas con un idiolecto menos marcado escriben en el *ABC*, lo que puede hacernos reflexionar sobre si es posible que el criterio unificador de los correctores y estilistas de dicho periódico haya producido cierta "des-personalización" en los artículos de estos dos autores, borrando ciertas huellas de su idiolecto.

[2] Woolls y Coulthard evidencian que "the method of calculation gives a higher score when the proportion of once-only use increases" (1998: 52), y proponen, en un intento de contrarrestar el efecto de los diferentes índices de crecimiento de las palabras de contenido y las palabras de función en los textos de desigual extensión, sustituir en la fórmula de Honoré los *hapax legomena* (V_1) totales del texto por los *hapax legomena* de contenido. Sea como fuere, lo cierto es que los algoritmos de riqueza de un texto pueden verse afectados por la mayor o menor presencia de topónimos, antropónimos, patronímicos, abreviaturas, siglas, etc., y que tal vez sea necesario eliminar todas esas palabras forzadas antes hacer el cálculo de estas mediciones

4. CONCLUSIONES FINALES

Nuestro estudio permite establecer un *ranking* de los algoritmos de atribución más eficaces para el idioma del español, al tiempo que permite poner en tela de juicio algunos de los que se han puesto en práctica hasta el momento, pues su bajo grado de discriminación niega su utilidad, al menos en las condiciones en las que nuestro análisis se ha realizado, y plantea la necesidad de buscar una medición más fiable para calcular la riqueza de vocabulario (que, posiblemente, esté relacionada con la distinción entre las palabras de contenido y las de función).

Sin embargo, a pesar de que este experimento permite establecer una serie de algoritmos fiables para los estudios de atribución de autoría de textos del español actual, encontramos problemas graves (pero en modo alguno insalvables) a la hora de poder adaptarlos al caso de textos anónimos auriseculares, y ello es así por dos razones, una relacionada con la fase de creación y la otra con la fase de edición y de las obras:

a) Los hábitos de escritura en los Siglos de Oro son poco conocidos, pero pensamos que la existencia de varias manos tras la conformación de una obra literaria era más frecuente de lo esperado (bien por verdadera colaboración entre autores, bien por actuar uno como corrector de otro).

b) Los correctores, componedores y cajistas de imprenta podrían ser en alta medida los responsables de muchas de las marcas que nuestros procedimientos de medición actuales contemplan. Las lecciones accidentales (ortografía y puntuación) conservadas en los diversos testimonios en que nos han llegado no son las originarias de los escritores, y, por tanto, imposibilitan que las variables basadas en mediciones relacionadas con los grafemas o las marcas de puntuación puedan ser aplicadas —o, al menos, que deba concedérseles el mismo grado de confianza. Desgraciadamente del "top ten" obtenido para el español sólo resulta operativo el *perfil de palabra simple*.

Por estas dos razones, una conclusión final se impone: aunque, sin duda, hay que seguir investigando (sobre todo para superar los dos problemas arriba comentados), el camino correcto pasa por el establecimiento de unos sistemas de atribución basados en la estilometría y en la cuantificación objetivadora de los fenómenos

suceptibles de ser medidos. Queda por delante un camino sembrado de retos, pero también de promesas. Las vías de análisis que se abren y que nosotros aquí hemos querido evaluar, si atentan (es cierto) contra el "misterio" y el "prestigio" que siempre rodean al anonimato, contribuirán sin duda a devolver el foco de atención a los aspectos puramente filológicos de la obra.

BIBLIOGRAFÍA

* FUENTES PRIMARIAS

ABC, disponible en <www.abc.es>
El mundo, disponible en <www.elmundo.es>
El país, disponible en <www.elpais.com>

* FUENTES SECUNDARIAS

General

Grieve, Jack (2007), "Quantitative Authorship Attribution: An Evaluation of Techniques", *Literary and Linguistic Computing*, Vol. 22, nº 3, pp. 251-70.

Marcas de puntuación

Chaski, C. E. (2001), "Empirical evaluation of languagebased author identification techniques", *Forensic Linguistics*, 8, 1–65.
O'Donnell, B. (1966), "Stephen Crane's The O'Ruddy: A Problem In Authorship Discrimination", en Leed (ed.), *The Computer and Literary Style*, Kent, Kent State University Press, pp. 107-15.

Frecuencia de palabras

Burrows, J. F. y H. Craig (2001), "Lucy Hutchinson and the authorship of two seventeenth-century poems: a computational approach", *The Seventeenth Century*, 16, 259-82.
Ellegard, A. (1962), *A Statistical Method for Determining Authorship: 1769–72*, Gothenburg, Acta Universitatis Gothoburgensis.
Holmes, D., I. Gordon y C. Wilson (2001), "A widow and her soldier: stylometry and the American civil war", *Literary and Linguistic Computing*, 16, pp. 403-20.

Morton, A. Q. (1978), *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*, New York, Scribners.

—— y M. Levison (1966), "Some indicators of authorship in Greek prose", en Leed (ed.), *The Computer and Literary Style*, Kent, Kent State University Press, pp. 141-79.

Smith, M. W. A. (1983), "Recent experience and new developments of methods for the determination of authorship", *Association for Literary and Linguistic Computing Bulletin*, 11, pp. 73-82.

Tweedie, F. y H. Baayen (1998), "How variable may a constant be? Measures of lexical richness in perspective", *Computers and the Humanities*, 32, pp. 323-53.

Grafemas (frecuencia y colocación)

Herdan, G. (1966), *The Advanced Theory of Language as choice and Chance*, New York, Springer-Verlag.

Ledger, G. (1998), "An exploration of differences in the pauline epistles using multivariate statistical analysis", *Literary and Linguistic Computing*, 10, pp. 85-97.

Merriam, T. (1998), "Heterogeneous authorship in early Shakespeare and the problem of Henry V", *Literary and Linguistic Computing*, 13, pp. 15-28.

N-gram

Clement, R. y D. Sharp (2003), "Ngram and Bayesian classification of documents", *Literary and Linguistic Computing*, 18, pp. 423-447.

Keselj, V., F. Peng, N. Cercone y C. Thomas (2003), "N-gram based author profiles for authorship attribution", en *Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING)*, Dalhousie University, Halifax, Nova Scotia, Canada, pp. 255-264.

Riqueza léxica

Holmes, D. y R. Forsyth (1995), "The Federalist revisited: new directions in authorship attribution", *Literary and Linguistic Computing*, 10, pp. 111-127.

Pollatschek y Y. T. Radday (1985), "Vocabulary Richness and Concentration", en Y. T. Radday y H. Shore (eds), *Genesis – an Authorship Study*, Rome, Biblical Institute.

Woolls, D. y M. Coulthard (1998), "Tools for the trade", *Forensic Linguistics. The International Journal of Speech, Language and Law*, 5 (1), pp. 33-57.

—— (2003), "Better tools for the trade and how to use them", *Forensic Linguistics. The International Journal of Speech*, Language and Law, 10 (1), pp. 102-112.

—— (2005), "La equivalencia y la diferenciación en la determinación forense de autoría de textos", en M. T. Turell (ed.*), Lingüística forense, lengua y derecho. Conceptos, métodos y aplicaciones*, Barcelona, IULA / Universitat Pompeu Fabra, pp. 299-310.

*** PROGRAMAS INFORMÁTICOS:**

Fletcher, W. H. (2007), *KfNgram*, versión 1.3.1., KwiCFinder.

Scott, M. (2008), *WordSmith Tools*, versión 5.0, Liverpool, Lexical Analysis Software.

Woolls, D (2005), *Vocalyse Toolkit*, versión JVocalyse v 2.05, CFL Software Development.