ACERCA DE LA SELLECCIÓN Y CREACIÓN DE CORPUS PARA LOS FINES DE LA EVALUACIÓN DE TÉCNICAS DE ANÁLISIS LINGÜÍSTICO FORENSE

Мария Спасова,
Нов български университет

Summary

This article presents the methodological approach to corpus selection criteria for the purposes of evaluation of linguistic forensic techniques' viability of application to real case data. Factors crucial to corpus design are discussed in terms of their influence on the process of evaluation of specific authorship marker's adequateness for implementation in written text forensic linguistic analysis.

Para que una metodología sea considerada válida y fiable, sobre todo en el contexto del trabajo forense, que pone en juego no sólo la autenticidad de la técnica y la reputación del perito que la aplica, sino que puede tener consecuencias para el futuro de una persona, es indispensable someterla a una evaluación lo más completa posible. Con el fin de cumplir con este propósito, en la evaluación de una possible marca de autoría es indispensable recurrir

Годишник на департамент „Чужди езици и култури“, т. 1 (2018), стр. 262-272

a la recogida y a la explotación de dos tipos de corpus: uno de análisis y otro de control. El corpus de análisis nos sirve para establecer los límites de la capacidad discriminatoria de la marca en cuestión. Su diseño y los experimentos que se llevan a cabo han de ser ideados teniendo en cuenta los principales factores cuyo efecto se considera que puede afectar (reducir o anular) el potencial discriminatorio de una marca identificativa. Con la explotación del corpus de análisis testamos la técnica de atribución de autoría que implementa la marca a nivel de la lengua general. El corpus de control, en cambio, nos permite llevar a cabo pruebas de evaluación de la propuesta analítica que constituye dicha técnica en el contexto de los textos forenses. Este corpus viene determinado por la particularidad y la finalidad del caso de autoria concreto y las pruebas escrita que presenta. En este artículo nos detendremos a detallar las características específicas que debe poseer el corpus de análisis y las variables indeptendientes que hay que tener en consideración en su selección y creación.

1. El corpus literario como corpus de análisis en la investigación lingüística forense

El tipo de corpus de análisis que más comúnmente se emplearse en la evaluación de técnicas de análisis lingüístico forense de textos escritos por su disponibilidad, suele ser el corpus de textos literarios. A la vista de los comentarios críticos que suscita este planteamiento de trabajo metodológico en los círculos de peritos lingüistas (Grant, comunicación personal), no podemos dejar de empezar la descripción de este tipo de corpus sin prestar atención a aquellas críticas que consideramos que son fundadas y exponer los argumentos por los que cosideramos la explotación de un corpus literario para los fines de la evaluación inicial de una marca de autoría.

Lo que sobre todo provoca las críticas respecto a la elaboración y al desarrollo de nuevas técnicas de análisis lingüístico forense a partir de la experimentación con corpus literarios es la dificultad de generalizar las conclusiones acerca de su fiabilidad y de las marcas identificativas que se comprueban con ellos. En nuestra opinión, las críticas de esta índole son contundentes y difícilmente se

podrían tachar de infundadas cuando surgen a propósito de la tendencia de algunos estudios de incitar a creer que una técnica "validada" en un corpus de narrativa daría los mismos resultados en los textos de un caso real (Grant y Baker, 2001). Las pruebas de delitos lingüísticos, exceptuando los casos de plagio, no son nunca obras literarias, y tratándose de dos géneros distintos con sus propias características estilísticas, producidos para un público diferente y en circunstancias muy distintas, es posible que un método cuya eficacia ha quedado comprobada en los textos de narrativa resulte inaplicable o de rendimiento negativo en el análisis de textos forenses. De ahí que no se puede concluir que una técnica de atribución de autoría es igualmente válida para el peritaje de textos forenses sin que se haya ejecutado previamente su evaluación en un corpus de casos reales. Por este motivo, cualquier trabajo sobre el potencial discriminatorio de una marca de autoría debe incluir una serie de experimentos basados en un corpus de control, es decir, con documentos de dos casos real.

En cambio, discrepamos con las críticas que cuestionan la conveniencia de la elección de un corpus literario para la investigación en general sobre marcas lingüísticas de atribución forense de autoría. Actualmente, la metodología en autoría está en un estado de desarrollo cuyos avances dependen en mayor grado de la investigación que de la práctica en el campo. Aunque en la actualidad las opiniones y habilidades expertas de los lingüistas forenses se solicitan más a menudo que hace una década, por cuestiones de confidencialidad el perito no puede divulgar su trabajo de análisis de casos de crímenes lingüísticos o que implican pruebas lingüísticas, o usar como corpus los documentos de los casos de su "book" profesional. Por lo tanto, a la hora de elaborar métodos de análisis para fines forenses, los lingüistas han tenido que recurrir al uso de corpus más accesibles, como los corpus de textos literarios. Esta vía de progreso metodológico no desfavorece la disciplina cuando la investigación tiene las características de la que se realiza en atribución de autoría forense.

La investigación en el campo de la comparación forense de textos escritos es exploratoria y experimental. Nos permite calificarla de exploratoria el hecho de que la

mayoría de propuestas analíticas parten de la búsqueda, en el vasto universo de la lengua, de unidades y fenómenos lingüísticos que pueden manifestarse con usos idiosincrásicos en los idiolectos de los individuos usuarios de esta lengua. Es experimental porque su diseño está concebido de manera que mediante la formulación y evaluación de hipótesis haga posible llegar al mismo tiempo a conclusiones sobre la idiosincrasia y el potencial distintivo de la unidad candidata a marca identificativa y sobre el efecto de determinados factores que podrían influirlos.

Teniendo en cuenta todo lo mencionado anteriormente, creemos que un corpus de narrativa, a pesar de no ser la opción ideal, como sería el caso de un corpus forense, responde a las necesidades de los estudios en esta línea de investigación por varias razones. En primer lugar, compilar un corpus representativo de la lengua de análisis que permita estudiar la variable o marca en el comportamiento lingüístico de un gran número de hablantes y hacer generalizaciones sobre su carácter idiosincrásico (no sobre su aplicabilidad como marca de identificación) es más factible y, tratándose de una investigación en variación, metodológicamente más correcto. Incluso si dispusiéramos de suficientes textos forenses como para crear un corpus, nos encontraríamos con una serie de problemas de muestreo. Por un lado, los sujetos no serían caracterizados socialmente y no seríamos íaces de controlar los factores que podrían influir en la validez de los resultados del análisis. Por otro lado, muy probablemente si tuviéramos muchos sujetos de estudio no contaríamos con el mismo número de textos por autor. En consecuencia, la distribución de los textos no sería proporcional y nuestro corpus no sería homogéneo . En segundo lugar, hemos de resaltar que el corpus de narrativa brinda al investigador la oportunidad de realizar experimentos a la medida de sus objetivos e hipótesis.

Por todo lo expuesto, podemos concluir que el trabajo con textos de una tipología genérica como los narrativos constituye un buen punto de partida en la evaluación de una técnica de comparación textual forense conducente a la atribución de autoría.

2. Variables independientes que condicionan la selección del corpus de análisis

En el contexto de la lingüística forense las variables independientes son aquellas que pueden restringir el comportamiento una unidad lingüística como unidad de análisis y también, por lo tanto, su carácter discriminatorio como marca idiosincrásica.

En la investigación en atribución forense de autoría, las variables independientes que resultan relevantes y que es preciso tener en cuenta son aquellas cuya presencia conllevaría cambios en el idiolecto escrito de una persona (variables que dependen del individuo), y aquellas que podrían dificultar la ejecución correcta del análisis lingüístico forense o incluso imposibilitarlo (variables que dependen del texto).

A continuación, definimos cada tipo de variable independiente y explicamos su influencia en la práctica de comparación lingüística forense para la atribución de autoría de textos escritos.

a. Variables ligadas al autor

Las variables ligadas al autor son aquellas que están relacionadas con las características específicas del individuo, y que se considera que podrían influenciar el análisis lingüístico por suponer cambios y diferencias estilísticas. De las variables que comprende este grupo, en este trabajo exploramos solo la variedad lingüística y su efecto en el estilo escrito.

b. La variedad lingüística del autor

Variedad lingüística es el término que se ha adoptado en lingüística para denominar las diferencias lingüísticas entre los individuos usuarios de la misma lengua sin incurrir en el uso del término dialecto que a menudo puede resultar ambiguo e incluso peyorativo . En el marco de este trabajo nos centramos exclusivamente en las diferencias en el uso del lenguaje que ocurren a causa de la distancia geográfica entre los hablantes, es decir, en la variedad geográfica. Sin embargo, para referirnos a este tipo de variedad utilizaremos el término genérico variedad

lingüística. Las diferencias que abarca se contemplan en los principales niveles lingüísticos (fonético, léxico y sintáctico) y pueden ser detectadas tanto en el habla como en la escritura de una persona.

Desde el punto de vista de la lingüística forense y la atribución de autoría, podemos hablar de variedad lingüística aún cuando se trata de diferencias que se detectan en el uso de la lengua de una persona no nativa en comparación con otra nativa..

Analizar estas diferencias lingüísticas en lingüística forense, y en atribución de autoría en concreto, nos puede servir para crear un modelo para el trazado de perfiles lingüísticos basado en las marcas idiolectales que son propias de cada variedad.

c. Variables ligadas al texto

Las variables ligadas al texto tienen que ver con las características textuales de las pruebas lingüísticas. Las que tienen mayor peso a la hora de tomar una decisión sobre la viabilidad del análisis lingüístico forense son la extensión, el género textual y el tiempo de mediación en la producción de los escritos.

d. La extensión del texto

La extensión de los documentos que componen el corpus de análisis en un caso real (textos dubitados e indubitados) es la variable independiente decisiva para el peritaje. Es así, porque, pese a la experiencia y a la alta competencia en la lengua y sus variedades y particularidades, el lingüista forense es incapaz de llevar a cabo la pericia si solo tiene en su poder un único texto cuyo contenido se limita a unas pocas palabras o líneas (por ejemplo, “Págame o te mato”). Este es el ejemplo de un caso extremo, pero no inverosímil, pues también es cierto que la mayoría de los textos dubitados que llegan a las manos de los expertos forenses raras veces exceden una centena de palabras, y por lo general suelen ser más cortos. Por ello, la buena práctica en comparación lingüística forense para la atribución de autoría exige que los textos que se peritan tengan una extensión que permita que el lenguaje

escrito de los sujetos implicados pueda someterse al análisis cuantitativo y cualitativo . Esta exigencia se refiere tanto a los textos dubitados como a los indubitados. Una excepción a esta regla son los casos en los que se dispone de un número considerable de los dos tipos de texto y sujeto de forma que el volumen compensa las carencias del corpus en longitud.

En atribución de autoría otra cuestión en relación a la extensión de las muestras del corpus concierne las marcas identificativas. Esta cuestión tiene que ver con el grado de dependencia que existe entre la aplicabilidad de un rasgo idiosincrásico como marca y el tamaño de las muestras de texto. Es decir, la probabilidad de que una marca de cualquier tipo ocurra en el texto de análisis y su potencial discriminatorio como tal disminuyen a la par que la extensión del texto.

Para estimar el nivel en el que la variable extensión del texto incide en el potencial discriminatorio de los n-gramas, realizamos un estudio en el que aplicamos el análisis lingüístico forense basado en los n-gramas a textos cortos (300 palabras) y a textos largos (600 palabras) y contrastamos los resultados obtenidos.

e. El tiempo de mediación

Cuando hablamos de tiempo de mediación en atribución de autoría nos solemos referir al tiempo que ha transcurrido entre la producción de una prueba lingüística oral o escrita y otra de la misma tipología. A diferencia de la variación relacionada con la variable edad, cuya influencia también implica el transcurso del tiempo, la variación ligada a la variable tiempo de mediación es más fácil de medir ya que podemos obtener observaciones de los eventuales cambios idiolectales y estilísticos intra autor a causa del efecto del tiempo de mediación, con la recogida de muestras de cada punto intermedio en el período de estudio. Con la variable edad sería muy difícil, si no imposible, hacer lo mismo porque la evolución intelectual y cultural del ser humano no son procesos que se desarrollan de manera uniforme en las fases de su ciclo vital. Esta es la razón por la cual los escasos estudios

en estilometría sobre la variación intra autor (Can y Patton, 2004), prefieren centrarse en el análisis del efecto del tiempo de mediación. En particular, nuestro estudio se ocupa del análisis de la variación estilística intra autor en dos contextos: uno en el que la distancia del tiempo de mediación entre los escritos de un autor es mayor, y otro en el que es menor.

f. El género textual

Concluimos la descripción de las variables independientes y el comentario introductorio sobre su repercusión en la metodología en atribución de autoría con la variable más problemática en relación a la evaluación experimental y la aplicación de las técnicas de comparación de pruebas lingüísticas de casos reales: el género textual.

El género textual al que pertenece un escrito se determina en función de sus características específicas y conforme criterios socio-culturales y funcionales. Las pruebas lingüísticas, es decir, los textos dubitados, son difíciles de clasificar dentro de la tipología textual convencional por su carácter “camaleónico”.

> Many texts, [...], which are analysed as part of forensic casework, are not inherently criminal; they may be more mundane including for instance, personal letters and diaries. […] Given the variety of texts subject to forensic analysis there is real danger in attempting to make generalizations about their character. (Grant, 2007: 216)

Al tiempo que comparten muchas de las propiedades de un género textual, los textos forenses poseen sus propias características peculiares y únicas. Por ejemplo, un mensaje circular de empresa que contiene información infamatoria y ofensiva sobre uno de los empleados puede tener la mayoría o todas las características típicas de una carta formal. Es así porque en realidad un texto, cualquiera que sea su género, se convierte en un documento dubitado cuando se le considere de carácter delictivo y surja la necesidad de su análisis lingüístico.

En la investigación en el campo de la atribución de autoría para el fin de la evaluación de los métodos y las técnicas de análisis lingüístico forense, desde los inicios de la disciplina se ha recurrido casi siempre al uso de cor-

pus constituidos por textos del género literario (Ellegard, 1962; Mosteller y Wallace, 1964; Holmes, 1994, 1998, 2001; Baayen et al., 1996; Hoover, 2001, 2002, 2003, entre otros) .

Esta práctica, sin embargo, ha suscitado críticas y serios debates en los congresos europeos e internacionales de lingüistas forenses sobre su adecuación para la evaluación de las técnicas de atribución de autoría. La discusión proviene del hecho de que el género literario no es el género prototípico de los textos dubitados (ni de los indubitados) y es muy probable de que los resultados de los experimentos de evaluación basados en ellos no se puedan generalizar a los contextos de trabajo con textos de casos reales.

> If there are some differences in the character of texts in more literary authorship analysis when compared to those of forensic case work, this raises the interesting question of whether methods and assumptions from the more academic field can be transferred to the applied setting […].(Grant, 2007: 217)

A pesar de que compartimos estas ideas, creemos que el trabajo experimental con textos literarios es un paso previo substancial por el que empezar en el proceso de evaluación de técnicas de atribución de autoría.

Además de esta última que acabamos de destacar, existen otras razones por las que las investigaciones en esta rama de la lingüística forense se han basado habitualmente en textos literarios. Una es que el acceso a textos de casos reales suele estar restringido a los órganos de la ley y a los abogados de las partes implicadas, mientras que los textos literarios son de dominio público y se encuentran al alcance del investigador científico. Y otra razón es que la experimentación en atribución de autoría basada en este género permite seleccionar muestras de una extensión significativa, mientras que los textos de casos reales suelen ser en general bastante cortos.

En la actualidad es más importante tratar de resolver otro problema trascendental que tiene que ver con la variable género textual, la disparidad en la tipología textual de los documentos dubitados e indubitados habitual en una pericia. Esto ocurre porque muchas veces no se

pueden aportar para el análisis lingüístico textos indubitados del sospechoso o los sospechosos que pertenezcan al mismo género que los dubitados. Chaski (2001) pone en duda que en estos casos se pueda realizar el peritaje argumentando que cada género se caracteriza por sus propios esquemas textuales, construcciones argumentativas, estructuras y léxico, por lo que resulta más fácil discriminar entre textos de diferentes géneros que entre textos del mismo género. Debido a estas diferencias, la comparación de los textos puede llevar a conclusiones erróneas respecto a su autoría.

En conclusión queremos enfatizar el hecho de que las que acabamos de describir no agotan la diversidad de posibles variables que pueden influir en la comparación lingüística forense de textos escritos. En el contexto de una pericia, como resultado de las características propias del corpus de análisis y de las personas implicadas, como también de otra índole, pueden surgir muchas más variables que resultan difíciles de prever y de considerar en una sola investigación. Sin embargo, las que hemos son las variables que, basándonos en nuestra experiencia en el trabajo con textos forenses, podemos decir que son comunes a la mayoría de casos y por ello debe prevalecer su estudio.

BIBLIOGRAFÍA

Baayen, R.H., van Halteren, H. et al. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing,* vol.11(3), págs.121-131

Can, F., Patton, J.M. (2004). Change of writing style with time. *Computers and the Humanities*, 38. págs. 61-82

Chaski, C.E. (2001). Empirical evaluation of language-based author identification techniques. *Forensic Linguistics, 8(2).* Págs. 1-65

Grant, T. y Baker, K. (2001) Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics*, 8(1), págs.66-79

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language and the Law,* 14(1), págs. 1-25

Holmes, D.I. (1998). The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing, 13(3). págs.111-117

Holmes, D. I.(1994). Authorship Attribution. *Computers and the Humanities,* 28(2), págs.87-106

Holmes, D.I. et al. (2001). Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution. Computers and the Humanities, 37. págs. 315-331

Hoover, D.L. (2001). Statistical analysis and authorship attribution: an empirical investigation. *Literary and Linguistic Computing*, 16 (4). págs. 421-444

Hoover, D.L. (2002). Frequent words frecuencies and statistical stylistics. Literary and Linguistic Computing, 17(2). págs. 157-180

Hoover, D.L. (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing, 18(3)*. págs. 261-286

Mosteller, F. y Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. New York: Springer-Verlag.