ÐL ESPAÑOL
REVISTA DE LENGUA

Revisión de ALTXA 1.2

FUNCIONES Y MANEJO DEL *SOFTWARE* PARA ANÁLISIS DE TEXTOS

Review of ALTXA 1.2

FUNCITONALITIES AND USE OF THE SOFTWARE FOR TEXT ANALYSIS

Juan Antonio Latorre
Universidad Complutense de Madrid
jualator@ucm.es
https://orcid.org/0000-0002-0818-636X

Resumen: El presente artículo es una revisión de la versión 1.2 del *software* ALTXA para el análisis forense de textos. Esta herramienta computacional aúna en una interfaz accesible procedimientos básicos dentro de los estudios forenses de atribución de autoría, tales como el cálculo del número medio de palabras por frase de una muestra y su riqueza léxica, con otros de mayor complejidad, como es el caso de la identificación de los n-gramas compartidos entre dos muestras y la realización del *Zeta test*. El desarrollo de este programa forma parte de una iniciativa que pretende facilitar la implementation de la lingüística forense en contextos educativos. El artículo también expondrá brevemente el objetivo del canal de YouTube *Project ALTXA*, donde se subirán videotutoriales sobre el uso del programa informático y vídeos divulgativos sobre las distintas ramas de la disciplina.

Palabras clave: lingüística forense, lingüística computacional, atribución de autoría, ALTXA.

Abstract: The present article is a review of the version 1.2 of the software ALTXA for forensic text analysis. This computational tool combines in an accessible interface basic procedures within the field of forensic authorship studies, such as the calculation of the average number of words per sentence of a sample and its lexical richness, with more complex ones, as it is the case of the identification of the n-grams shared by two samples and the conduction of the Zeta test. The development of this program is part of an iniciative that seeks to facilitate the implementation of Forensic Linguistics in educational contexts. In addition, the article will briefly address the objective of the YouTube channel *Project ALTXA*, where videotutorials on how to use the software and informative lectures about the distinct branches of the discipline will be uploaded.

Keywords: Forensic Linguistics, Computational Linguistics, authorship attribution, ALTXA.

Recibido: xx/xx/xxxx.
Aceptado: xx/xx/xxxx.

1. Introducción

Hace ya más de una década, Tim Grant (2007) identificó una tendencia creciente hacia el empleo de métodos cuantitativos en el ámbito disciplinario de la lingüística forense, ya que estos permiten una presentación científica de los resultados que repercute positivamente en la credibilidad del investigador.

El desarrollo de las herramientas computacionales que ha tenido lugar durante las últimas décadas ha contribuido en gran medida al perfeccionamiento de estos métodos cuantitativos, pues su uso facilita la consideración de variables estadísticas que en otros tiempos no eran accesibles (Kinney, 2009).

Dentro del contexto de los estudios forenses de atribución de autoría, los investigadores cuentan con un catálogo de herramientas computacionales cada vez más amplio, lo cual está directamente relacionado con el auge que está experimentando la disciplina en el mundo académico, donde el número de publicaciones relacionadas con la materia crece exponencialmente.

Dichas herramientas están disponibles para el usuario en forma de programas descargables, como es el caso de WordSmith Tools y AntConc (véase Smith [2021] para una revisión exhaustiva de una de las últimas actualizaciones de ambos), y plataformas online como Voyant Tools (véase Alhudithi [2021] para una revisión detallada de sus funcionalidades), SAUTEE (véase López-Escobedo, Sierra y Solórzano [2019] para una demostración práctica de sus posibilidades) y el archiconocido Sketch Engine (véase Arias Rodríguez y Fernández-Pampillón Cesteros [2020] para un videotutorial dirigido a aquellos que aún no están familiarizados con su uso). Asimismo, si bien las herramientas previamente mencionadas poseen funcionalidades limitadas, los lenguajes de programación como Java, Python y C++ ofrecen infinitud de posibilidades, aunque su manejo suele estar restringido a aquellos usuarios con conocimientos informáticos sólidos. En este último apartado destaca también el lenguaje de programación R, que cuenta con un paquete orientado al cálculo estilométrico. No obstante, cabe reiterar que el dominio de los lenguajes de programación suele estar fuera del rango de conocimientos de un lingüista.

El presente artículo pretende realizar una revisión de la versión 1.2 del programa informático ALTXA, desarrollado en Java por el Doctor en Lingüística Forense Juan Antonio Latorre García y el programador informático Carlos Antón Castaño con el propósito de combinar un catálogo amplio de

métodos forenses de atribución de autoría con una interfaz intuitiva y, por consiguiente, accesible para todo tipo de usuarios. El objetivo a largo plazo tras la creación de esta herramienta es facilitar su inclusión en contextos educativos, donde no solo hay una escasez de expertos forenses, sino de herramientas orientadas a la enseñanza. La validez de ALTXA ha sido testada en un estudio sobre la autoría de la obra isabelina *Arden of Faversham* (1592), en el que William Shakespeare y Christopher Marlowe fueron considerados los autores potenciales (Latorre García [2022]).

2. Funcionalidades y manejo de ALTXA

A continuación presentaremos las distintas funcionalidades de ALTXA y el modo de acceder a ellas a través de su interfaz, así como las principales similitudes y diferencias que esta herramienta presenta con respecto a otras también empleadas en el ámbito de la lingüística forense. Cada subsección se centrará en una pestaña de la interfaz y las posibilidades que esta ofrece.

Previo a dichas subsecciones, cabe mencionar que ALTXA puede descargarse tanto en el perfil de GitHub *JuanAntonioLatorre* como en la cuenta de Twitter *@projectaltxa*, donde también se publicarán futuras actualizaciones. ALTXA es compatible con todos los sistemas operativos y, para su correcto funcionamiento, es imprescindible que el usuario tenga Java descargado en el dispositivo donde lo vaya a utilizar.

2.1 La pestaña «Text Analysis»

Nada más abrir el programa, el usuario se encontrará por defecto en la pestaña *Text Analysis* (**Figura 1**). El primer fichero, llamado *Text file*, está diseñado para almacenar la muestra que el usuario desee analizar, la cual debe subirse en formato *.txt*. Al clicar *Execute*, el usuario tendrá acceso inmediato a una serie de variables básicas sobre la muestra, siendo la primera de estas su número de oraciones. Para calcular este parámetro, ALTXA considerará como final de oración el punto, los dos puntos, el signo de cierre de exclamación y el de cierre de interrogación. Asimismo, el programa también mostrará el número total de palabras o *tokens* de la muestra, su número medio de letras por palabra, su número medio de palabras por frase, su número de palabras únicas o *types* y, por último, el porcentaje de su riqueza léxica. Este último parámetro con-

siste en la división del número de palabras únicas de la muestra o *types* entre su número total de palabras o *tokens* multiplicada por cien. Así pues, la riqueza léxica representa el porcentaje de palabras distintas que posee la muestra.

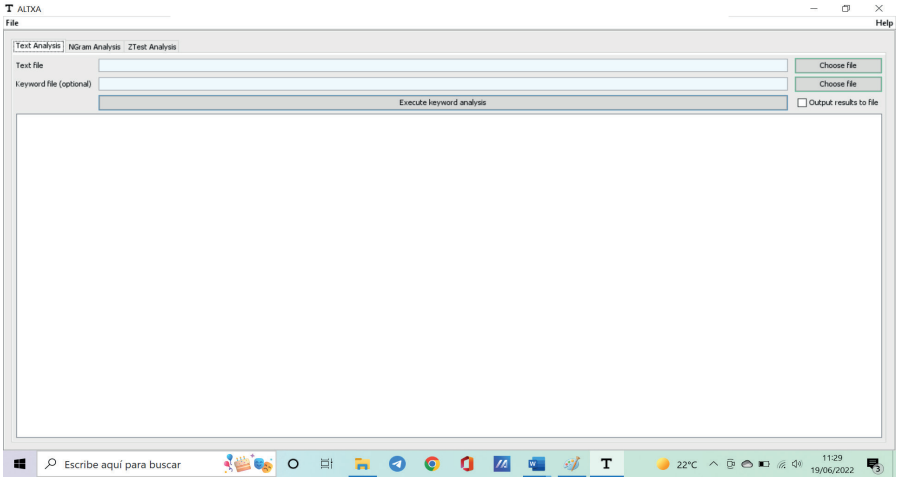

Figura 1. Pestaña *Text analysis*

El usuario también tiene la posibilidad de calcular la frecuencia relativa de una lista de palabras clave de su elección en la muestra si, antes de clicar *Execute*, sube al fichero *Keyword file* un documento *txt.* con dichas palabras escritas en minúscula y separadas por un espacio. La frecuencia relativa de una palabra clave en una muestra se calcula dividiendo su número de apariciones entre el número total de palabras o *tokens* de la muestra y multiplicando el resultado por cien. En la **Figura 2** tenemos una muestra del aspecto de la pestaña *Text Analysis* con salida de datos para todos los parámetros descritos hasta el momento.

Las funcionalidades descritas en este apartado se corresponden con el cálculo de parámetros básicos. Puesto que dichos parámetros también son accesibles desde otras herramientas computacionales mencionadas en el apartado anterior, como es el caso de Voyant Tools y WordSmith Tools, podría decirse que ALTXA no ofrece nada novedoso en este apartado. La inclusión de estas funcionalidades en el programa está más bien relacionada con la idea de cubrir las necesidades básicas del usuario promedio de este tipo de herramientas. La próxima subsección del artículo mostrará un nicho de acción más específico de ALTXA.

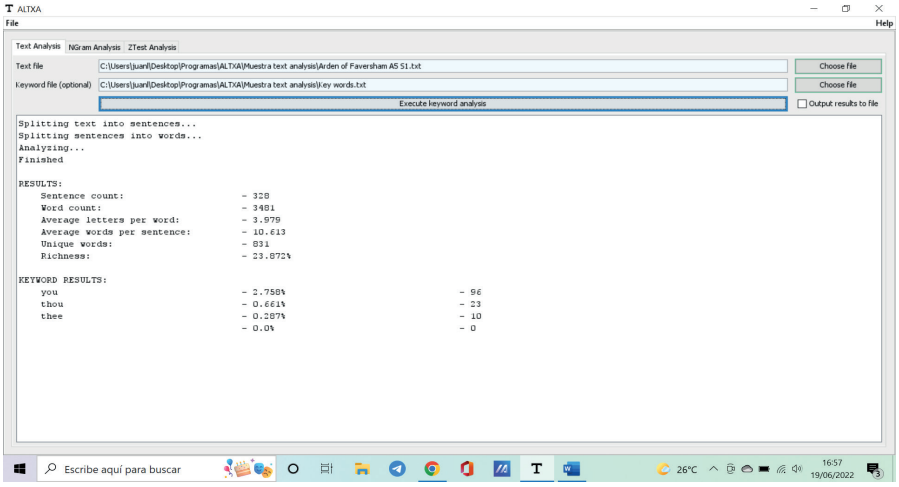

Figura 2. Muestra de salida de datos en la pestaña *Text analysis*

2.2 La pestaña «N-gram Analysis»

Los n-gramas suelen definirse como combinaciones de una o más formas lingüísticas (normalmente caracteres o palabras, aunque existen de otros tipos) que tienen lugar de forma consecutiva dentro de una misma oración (Cheng, Greaves y Warren [2006]; Ishihara, 2014; Grieve, Clarke, Chiang, Gideon, Heini, Nini y Waibel [2018]).

Tal y como podemos observar en la **Figura 3**, en la pestaña *N-gram Analysis* de ALTXA el usuario tiene la posibilidad subir dos muestras independientes en formato *.txt*. Al clicar *Execute*, el programa identificará los n-gramas de palabras que estas dos muestras comparten. Para optimizar el tiempo, es recomendable subir la muestra de menor tamaño al primer fichero, pues ALTXA elaborará una lista de todos los n-gramas de palabras que esta contiene para después buscar coincidencias en la segunda muestra. Esto significa que, si se sube la muestra de mayor tamaño al primer fichero, el proceso puede tardar algunos segundos más.

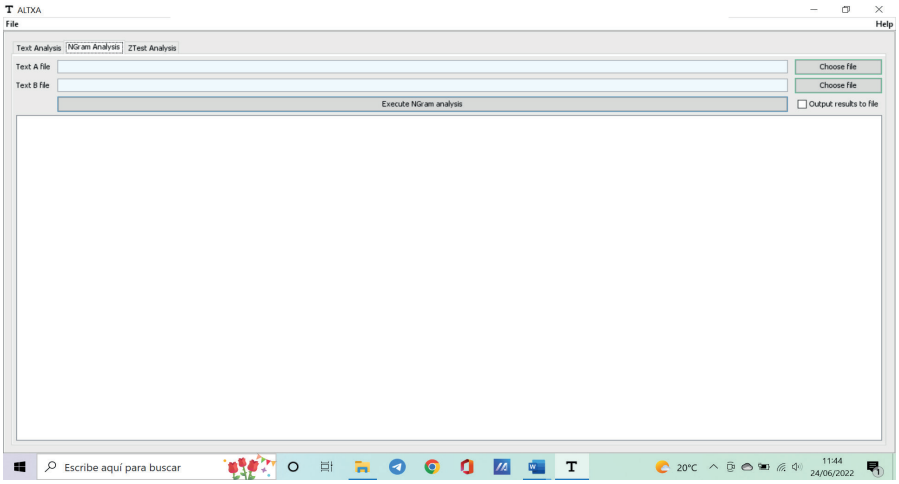

Figura 3. Pestaña *N-gram tracing*

La **Figura 4** constituye un ejemplo de salida de resultados. Como se puede observar en la imagen, ALTXA no solo muestra el número de n-gramas de cada tipo que comparten las muestras, sino que también revela cuáles son esos n-gramas y cuántas veces aparecen en cada una de las dos muestras, en caso de que el usuario desee realizar un análisis cualitativo de los mismos. El orden de aparición de los n-gramas en la interfaz de ALTXA estará determinado por su tamaño y, por ejemplo, los 6-gramas precederán a los 5-gramas, y así sucesivamente.

Si bien la identificación de n-gramas es una funcionalidad presente en otras herramientas computacionales como AntConc y Sketch Engine, cabe destacar que, mientras que en estas está orientada principalmente a la realización de una búsqueda customizada de n-gramas dentro de un determinado corpus, ALTXA se centra en la identificación de aquellos compartidos por dos muestras o corpus independientes, lo cual puede ser de mayor utilidad en estudios de atribución de autoría.

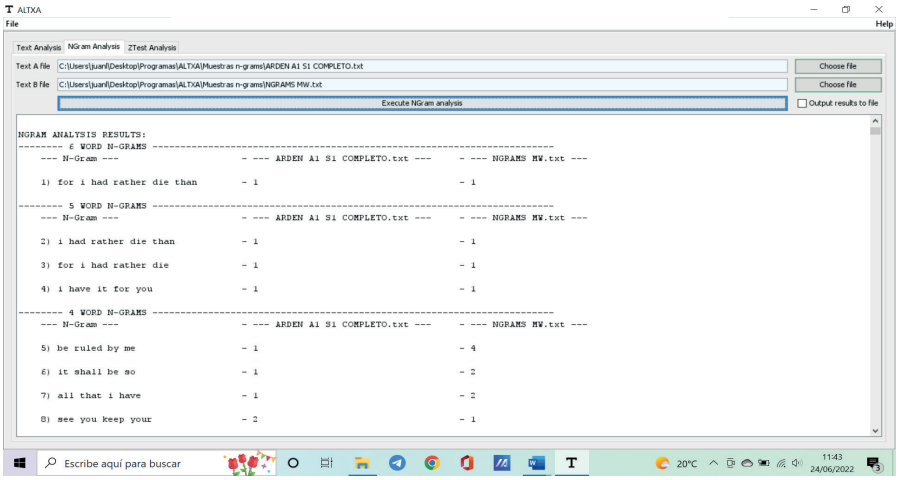

Figura 4. Pestaña *N-gram tracing*

2.3 La pestaña «ZTest Analysis»

Por último, explicaremos de forma sintética los distintos pasos que conforman la realización de la variante del *Zeta test* propuesta por Craig y Kinney (2009) y la forma en que este puede llevarse a cabo a través de ALTXA. Este método se utiliza en estudios de atribución de autoría para comparar una muestra dubitada con dos corpus de referencia para ver con cuál de ellos presenta mayor similitud léxica.

En primer lugar, los dos corpus de referencia deben dividirse en fragmentos de 2000 palabras y aquellas residuales que se encuentren al final de un texto dentro del corpus deben anexionarse al último fragmento de 2000 palabras de dicho texto. Si procede, la muestra dubitada debe dividirse siguiendo los mismos criterios.

El siguiente paso consiste en la obtención de 500 marcadores para cada uno de los dos corpus de referencia. Estos marcadores deben ser, según los autores, palabras léxicas, por lo que las palabras gramaticales y los nombres propios deben ser ignorados durante el proceso. Para determinar si una palabra léxica puede convertirse en uno de los 500 marcadores de uno de los corpus de referencia, es necesario calcular el porcentaje de fragmentos de 2000 palabras o más de este en los que aparece, independientemente del número de apariciones, y el porcentaje de fragmentos de 2000 palabras o más del otro corpus en los que no aparece. Si estos porcentajes de aparición

y no aparición de cada palabra léxica se suman y el resultado es superior a 100, esta pasa a ser un marcador potencial del primero de los corpus. Las 500 palabras que obtengan mediante este proceso un resultado superior a 100 se convertirán, por tanto, en los marcadores del primer corpus, y este proceso debe realizarse a la inversa para obtener los 500 marcadores del segundo corpus.

Una vez que se han obtenido los 500 marcadores para cada uno de los dos corpus de referencia, es el momento de realizar una representación gráfica de los fragmentos de 2000 palabras o más en los que estos y la muestra dubitada se dividieron. Para ello, estos deben colocarse sobre un plano de coordenadas en el que el valor del eje *x* para cada fragmento esté determinado por la división del número de marcadores del primer corpus de referencia que contenga entre su número de palabras distintas, mientras que el valor del eje *y* resulte de la división entre el número de marcadores del segundo corpus de referencia que contenga el fragmento y su número de palabras distintas.

Los fragmentos de los dos corpus de referencia se agruparán en posiciones opuestas del eje de coordenadas, por lo que la autoría de los fragmentos de la muestra dubitada se determinará por su proximidad hacia una u otra. En caso de que no sea posible discernir a simple vista hacia cuál de las dos agrupaciones de fragmentos indubitados se aproximan más aquellos de la muestra dubitada, se debe calcular el centroide de ambas agrupaciones y calcular la distancia de estos con respecto a la posición de cada fragmento dubitado.

Si bien el *Zeta test* es un método complejo de atribución de autoría, el diseño de la interfaz de ALTXA permite que su realización sea accesible. En la pestaña *Ztest*, el usuario encontrará cuatro ficheros (**Figura 5**). Los dos primeros permiten almacenar los corpus de referencia en formato *.txt*. Si estos corpus están integrados por más de un texto, el usuario debe escribir la combinación de símbolos #@# al final de cada texto dentro del corpus para que ALTXA pueda dividirlos correctamente, es decir, en fragmentos de 2000 palabras, pero añadiendo aquellas palabras residuales al final de cada texto al último fragmento de este. La muestra dubitada debe subirse al tercer fichero en formato *.txt* y siguiendo las mismas indicaciones descritas con anterioridad en caso de que contenga más de un texto.

El cuarto fichero permite almacenar lo que se conoce como *stop list*, es decir, una lista de palabras ignoradas para la obtención de los 500 marcadores de ambos corpus. En ella, el usuario debe incluir las palabras gramaticales de la lengua en la que esté realizando su estudio, a las cuales se puede acceder a

ISSN: 3020-2434, https://doi.org/10.33776/dlesp.v1.7921

través de numerosas páginas web donde aparecen recogidas en forma de lista, los nombres propios de los corpus de referencia y cualquier otro ítem léxico que desee ignorar. La posibilidad de elaborar una *stop list* para adaptar la conducción de cada *Zeta test* a las necesidades específicas del investigador es una de las ventajas que ofrece ALTXA para la conducción de este procedimiento.

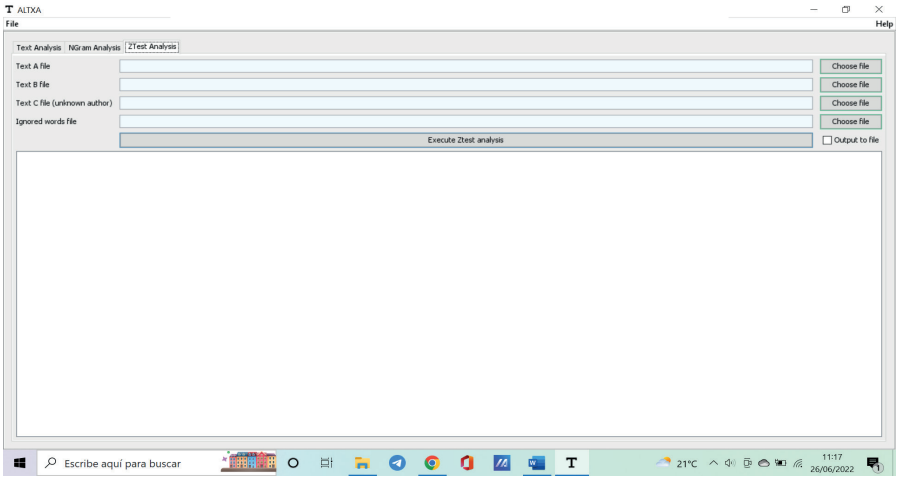

Figura 5. Pestaña *Zeta test*

Al clicar *Execute*, ALTXA dividirá las muestras siguiendo los criterios descritos anteriormente, calculará los 500 marcadores para los dos corpus de referencia y colocará cada uno de los fragmentos de 2000 o más palabras en un plano de coordenadas que el usuario recibirá en forma de archivo *.png* como el que podemos observar en la Figura 6 (véase Latorre García [2022] para ver la gráfica en su contexto). El programa representará los fragmentos del primer corpus de referencia con cuadrados de color rojo, los fragmentos del segundo corpus de referencia con círculos de color azul y, por último, los fragmentos de la muestra dubitada con triángulos de color negro. Asimismo, ALTXA generará un documento de Excel donde se detallen las coordenadas de cada fragmento sobre el plano de coordenadas, en caso de que el usuario desee exportarlas a otra base de datos con facilidad, y otro con los 500 marcadores de cada corpus de referencia para que el usuario pueda revisarlos en busca de errores.

En la actualidad, ALTXA es la única herramienta computacional que permite la conducción del *Zeta test* de una forma tan intuitiva y adaptada a las necesidades del usuario.

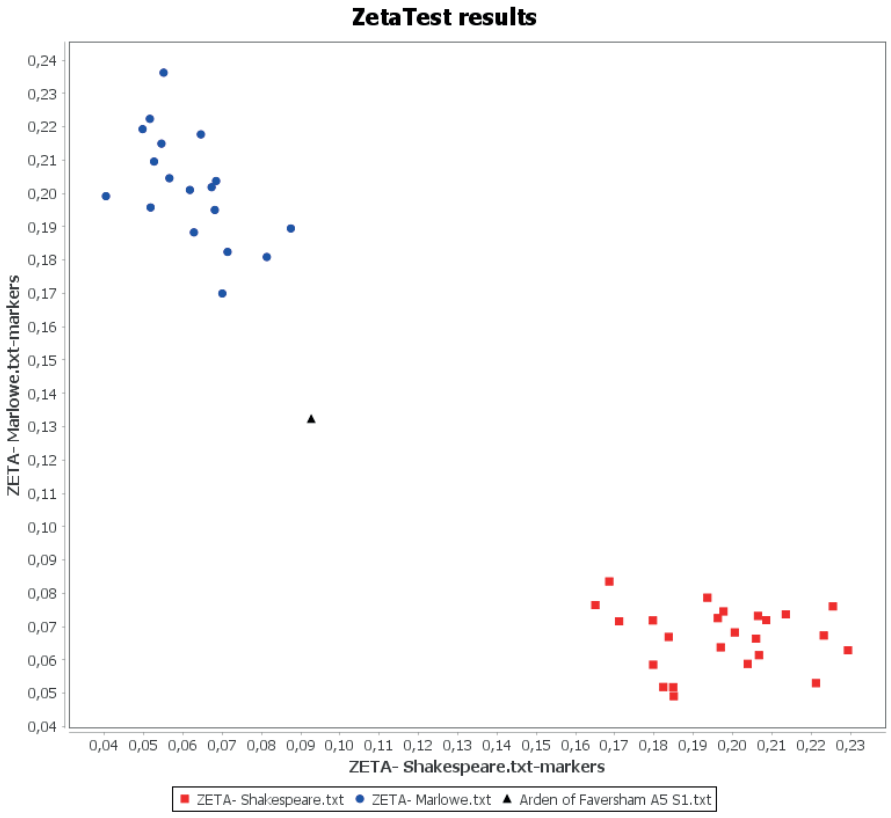

Figura 6. Muestra de salida de datos en la pestaña *Zeta test*

2.4 Opciones avanzadas

La versión 1.2 de ALTXA ofrece una serie de opciones avanzadas para el tratamiento de las muestras. Si el usuario clica el botón *File*, que se encuentra en la esquina superior izquierda de la interfaz (véanse **Figuras 1**, **2**, **3**, **4** y **5**), tendrá la posibilidad de activar la opción de que el programa ignore todas las secuencias de texto incluidas entre corchetes. Asimismo, también tendrá la posibilidad de comprobar el modo en el que ALTXA divide las palabras y las oraciones de las muestras activando el *Debug mode*, lo cual puede ser de utilidad a la hora de corregir errores o entender algunos de los resultados. Por último, aquellos usuarios con conocimientos básicos de programación tienen la posibilidad de eliminar secuencias de texto usando *REGEX* (*Regular*

Expressions). La **Figura 7** muestra el aspecto de este panel de opciones avanzadas sobre la interfaz del programa.

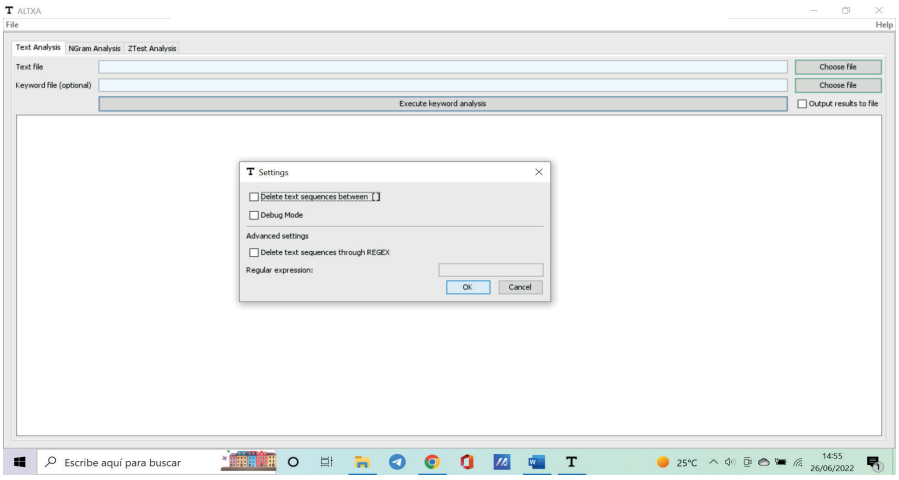

Figura 7. Opciones avanzadas en ALTXA

3. Futuras actualizaciones y el canal de YouTube *Project ALTXA*

A comienzos de cada año se publicará una actualización de ALTXA que ofrecerá al menos una nueva funcionalidad orientada a los estudios forenses de atribución de autoría. La versión 2.0 de ALTXA, que saldrá en enero de 2023, añadirá la posibilidad de realizar un *Principal Component Analysis* (véase Craig y Kinney [2009] para una explicación exhaustiva de este método). Así pues, ALTXA continuará expandiendo de forma gradual su catálogo de funcionalidades, las cuales se presentarán siempre de forma intuitiva al usuario para facilitar el asentamiento de la disciplina tanto en contexto profesionales como docentes.

Con este propósito en mente, en octubre de 2022 se estrenó el canal de YouTube *Project ALTXA*, el cual no solo ofrecerá videotutoriales acerca del uso de la herramienta, sino también vídeos divulgativos de diversa índole relacionados con la lingüística forense. Estos podrán tratar aspectos más teóricos de la disciplina, como el *Plain English Movement* (Felsenfeld [1981]), o explorar casos de carácter práctico, tales como el asesinato de Dulcéliz Díaz (Fitzgerald [2014]), los cuales pueden ser de gran utilidad para estudiantes e

investigadores en ciernes a la hora de orientar sus propios proyectos. De esta forma, el canal de YouTube servirá para reforzar una iniciativa que pretende la democratización de la disciplina en todas sus formas.

ISSN: 3020-2434, https://doi.org/10.33776/dlesp.v1.7921

Bibliografía

Alhudithi, Ella (2021): «Review of Voyant Tools: *See through your Text*», *Language Learning & Technology*, 25, 3, pp. 43-50, https://doi.org/10125/73446.

Anthony, Laurence (2022). *AntConc* (versión 4.0.3) [programa informáfico], https://www.laurenceanthony.net/softEware/antconc/.

Arias Rodríguez, Iván (2020). *Taller de Sketch Engine* [vídeo], https://www.youtube.com/watch?v=rLNs2UUVHB8.

Cheng, Winnie; Greaves, Chris, y Warren, Martin (2006). «From N-gram to Skipgram to Concgram», *International Journal of Corpus Linguistics*, 11, 4, pp. 411-433, https://doi.org/10.1075/ijcl.11.4.04che.

Craig, Hugh, y Kinney, Arthur F. (2009). «Methods», en Hugh Craig y Arthur F. Kinney (eds.), *Shakespeare, Computers and the Mystery of Authorship*, Cambridge: Cambridge University Press, pp. 15-39.

Felsenfeld, Carl (1981). «The Plain English Movement in the United States», *FLASH: The Fordham Law Archive of Scholarship and History*, 6, pp. 408-421.

Fitzgerald, James R. (2014). «Atribución de autoría y supuestas notas de suicidio: Análisis lingüístico forense y su papel en los tribunales penales estadounidenses en dos crímenes violentos ocurridos en 2007», en Elena Garayzábal, Miriam Jiménez Bernal y Mercedes Reigosa (coords.), *Lingüística forense: la Lingüística en el ámbito Legal y Policial*, Madrid, Euphonía Ediciones, pp. 49-77.

Grant, Tim (2007). «Quantifying evidence in forensic authorship analysis», *The International Journal of Speech, Language and the Law*, 14, 1, pp. 1-25, https://doi.org/10.1558/ijsll.v14i1.1.

Grieve, Jack; Clarke, Isobelle; Chiang, Emily; Gideon, Hannah; Heini, Annina; Nini, Andrea, y Waibel, Emily (2018). «Attributing the Bixby Letter Using N-gram Tracing», *Digital Scholarship in the Humanities*, 34, 3, pp. 493-512, https://doi.org/10.1093/llc/fqy042.

Ishihara, Shunichi (2014): «A Likelihood Ratio Based Evaluation of Strength of Authorship Attribution Evidence in SMS Messages Using N-grams», *The International Journal of Speech, Language and the Law*, 21, 1, pp. 23-49, https://doi.org/10.1558/ijsll.v21i1.23.

Kilgarriff, Adam, y Rychlý, Pavel (2003). *Sketch Engine* [herramienta en línea], https://www.sketchengine.eu/.

Kinney, Arthur F. (2009). «Authoring Arden of Faversham», en Hugh Craig y Arthur F. Kinney (eds.), *Shakespeare, Computers and the Mystery of Authorship*, Cambridge, Cambridge University Press, pp. 78-99.

Latorre García, Juan Antonio (2022). *Attribution of Authorship of «Arden of Faversham»: A Forensic Linguistic Study of William Shakespeare and Christopher Marlowe*, tesis doctoral dirigida por María Goicoechea y Elena Martínez Caro, Madrid, Universidad Complutense de Madrid.

Latorre García, Juan Antonio; y Antón Castaño, Carlos (2021). *ALTXA* [programa informático], https://github.com/JuanAntonioLatorre/ALTXA.

López-Escobedo, Fernanda; Sierra, Gerardo; y Solórzano, Julián (2019). «SAUTEE: un recurso en línea para análisis estilométricos», *Linguamática*, 11, 1, pp. 69-81, https://doi.org/10.21814/lm.11.1.270,

Scott, M. (2021). *WordSmith Tools* (Versión 8) [programa informático], https://www.lexically.net/wordsmith/.

Sierra, Gerardo; López, Fernanda; Solórzano Soto, Julián; Torres, Juan Manuel y Montes, Asucena (2019). *SAUTEE* [herramienta en línea], http://www.corpus.unam.mx/sautee#.

Sinclair, Stéfan, y Rockwell, Geoffrey (2022). *Voyant Tools* (Versión 2.5.3) [herramienta en línea], https://voyant-tools.org/.

Smith, Emily Louisa (2021). «A Review of the Computational Linguistics Tools WordSmith Tools (Version 8) and AntConc (version 3.5.8)», *Renaissance and Reformation*, 44, 1, pp. 200-214, https://doi.org/10.33137/rr.v44i1.37062.