Revista de Humanidades Digitales

Análisis de corpus poéticos con *Litcon*

Analysis of Poetic Corpora with Litcon

Laura HERNÁNDEZ-LORENZO
Universidad de Sevilla
laurahlr@gmail.com
https://orcid.org/0000-0003-3489-2193

RESUMEN

La investigación en análisis de corpus ha experimentado un aumento y una evolución significativa en los últimos años, con un incremento de las herramientas y los recursos disponibles para análisis y exploración de corpus textuales. Tras un breve repaso de los programas disponibles de análisis de corpus, en este artículo se presenta Litcon, un nuevo software desarrollado con especial atención al estudio de textos poéticos en verso. Litcon es multiplataforma y, además de contar con herramientas habituales en estos software como la generación de concordancias, listas de frecuencias de palabras y palabras clave, incluye funcionalidades para contrastar corpus, etiquetar y analizar texto etiquetado morfológicamente, y crear muestras aleatorias. Litcon y sus herramientas se presentan a través de un estudio de caso con la poesía de Fernando de Herrera.

PALABRAS CLAVE

Análisis estilístico, análisis de contenido, literatura, software, herramientas.

ABSTRACT

Research in corpus analysis has grown and evolved significantly in recent years, with an increase in the tools and resources available for the analysis and exploration of textual corpora. After a brief review of the available corpus analysis programs, this article presents Litcon, a new software developed with special attention to the study of poetic texts in verse. Litcon is cross-platform, it includes common tools of corpus analysis software, such as the generation of concordances, word frequency lists and keywords, and even functionalities for contrasting corpora, tagging and analyzing Part-of-Speech tagged text, and creating random samples. Litcon and its tools are presented through a case study with the poetry of Fernando de Herrera.

KEYWORDS

Stylistic Analysis, Content Analysis, Literature, Software, Tools.

RHD 7 (2022)

ISSN
2531-1786

10.5944/
rhd.vol.7.2022

Recibido 20/09/2021 – Aceptado 10/11/2022

1. INTRODUCCIÓN

La investigación en análisis de corpus ha experimentado un significativo aumento y evolución desde que el padre Busa llevara a cabo sus concordancias de las obras de Tomás de Aquino, proyecto considerado como fundacional tanto para las Humanidades Digitales (HD) como para la Lingüística de corpus (McEnery & Wilson, 2001; Hockey, 2004; Rojas Castro, 2013). A este han contribuido el incremento de textos disponibles en formatos digitales y los progresos tecnológicos y metodológicos. Como resultado, en los últimos años han aumentado las herramientas y recursos disponibles para análisis y exploración de corpus textuales, muchas de las cuales cuentan con interfaces visuales y no requieren de conocimientos previos de programación. Se trata, sin embargo, de herramientas genéricas que no tienen en consideración las particularidades formales de determinados tipos de textos, como es el caso de los textos poéticos en verso.

En este artículo, se presenta el programa Litcon, que ha sido desarrollado con especial atención al análisis de textos poéticos, y permite recuperar información sobre elementos clave de este género como son los títulos de los poemas, el número de versos o las pausas de final de verso y separación de patrones métricos, ya que se tiene en cuenta el cambio de línea. Este software integra, además, tanto las clásicas herramientas de análisis de corpus como otras utilidades, que se muestran a través de un estudio de caso de la poesía de Fernando de Herrera. El trabajo se estructura del siguiente modo: tras esta introducción, se incluye una breve panorámica de las herramientas y software de análisis de corpus disponibles, clasificadas según su procedencia (sección 2); a continuación, se introducen las características generales del programa Litcon (sección 3), seguidas del estudio de caso de la poesía de Herrera a través de este programa y sus herramientas (sección 4). Por último, se presentan las conclusiones (sección 5) y se ofrece la bibliografía utilizada.

2. HERRAMIENTAS Y SOFTWARE DE ANÁLISIS DE CORPUS: BREVE ESTADO DE LA CUESTIÓN

A la hora de acometer un análisis textual, y más concretamente, un análisis de corpus, existen una serie de herramientas o software utilizados con frecuencia dentro de las HD. Algunas requieren conocimientos de programación, mientras que otras cuentan con interfaces visuales que las hacen accesibles también a usuarios e investigadores que no poseen estos saberes. Estas últimas han aumentado considerablemente en años recientes y proceden de diferentes comunidades de investigación, principalmente de la Lingüística de corpus, de comunidades de HD y de la tradición francesa de la Textometría. Se presenta seguidamente una breve exposición de estas herramientas con especial atención a las que permiten al usuario trabajar con textos literarios, y especialmente poéticos, de su elección[1].

[1] Debido a las limitaciones de espacio de este trabajo, no se incluyen en esta breve exposición programas de Procesamiento de Lenguaje Natural que realizan tareas concretas de etiquetado de textos, como etiquetado *Part-of-Speech*, lematización o etiquetado de entidades nombradas, ni programas de etiquetado y análisis métrico, pues el objetivo es hacer un repaso de los programas de exploración de corpus que tiene a su disposición el investigador en poesía sin necesidad de realizar un gran trabajo de preprocesamiento de los textos. Otros software relevantes de análisis textual, pero ya más enfocados a la Estilometría y atribución de autoría, gratuitos y que exigen conocimientos mínimos de programación, son el paquete Stylo en R

2.1. Herramientas procedentes de la Lingüística de corpus

Algunas de las herramientas actuales de análisis textual provienen de los análisis de corpus, principalmente de la Lingüística de corpus, donde son frecuentemente utilizadas. McEnery y Hardie (2012) hacen un repaso histórico de estas herramientas y las clasifican por generaciones[2], por lo cual, los programas utilizados actualmente son los considerados de tercera generación[3]. Estos se distinguen de los de la primera y la segunda generación en que soportan corpus de gran tamaño, cubren una gran cantidad de herramientas en un solo programa, incluyen procedimientos de análisis en los que intervienen medidas y procedimientos estadísticos, y soportan sistemas de codificación amplios como Unicode. Este es el caso de WordSmith tools (Scott, 2016) o Antconc (Anthony, 2022b), ambos incluyen una serie de herramientas de análisis frecuentemente utilizadas en los estudios de corpus. De este modo, además de las concordancias[4] con las que cuentan programas anteriores, incluyen listas de frecuencias de palabras (*wordlists* o *frequency lists*)[5], colocaciones (*collocations*) y palabras clave (*keywords*)[7].

(Eder et al., 2016) y PyDelta en Python (Jannidis, 2014). Otro software de atribución de autoría que no exige conocimientos de programación es el JGAAP en Java (Juola, 2005), que a diferencia de los dos anteriores es un software comercial.

[2] Otro estado de la cuestión de estas herramientas y análisis crítico desde el punto de vista de un programador que las conoce por dentro puede encontrarse en el artículo de Laurence Anthony (2013). Una introducción para el lector no especializado sobre qué aportan estos programas frente a un investigador humano en el capítulo de Mike Scott (2010) en *The Routledge Handbook of Corpus Linguistics*. En la segunda edición de este volumen, se ofrece una panorámica más avanzada de estas herramientas, su uso, fortalezas y debilidades, además de otros conceptos de interés en su uso y desarrollo (Anthony, 2022).

[3] Existen también los llamados programas de corpus de cuarta generación, cuyo uso está en aumento, y que se distinguen de los de la tercera en que los textos analizados son descargados por el propio programa de la Web (Anthony, 2013; McEnery & Hardie, 2012). Ejemplos de estos software son CQWeb, SketchEngine o Wmatrix.

[4] Las concordancias son, sin duda, el procedimiento fundacional de los estudios de corpus, desarrollado por primera vez con métodos informáticos por el padre Roberto Busa, pero ya realizado con anterioridad a mano (McEnery & Hardie, 2012, p. 37). Se trata de una herramienta que recoge una lista con todas las ocurrencias en el texto de la palabra buscada, mostrando, además, el contexto en el que aparece cada una (es decir, se muestran algunas de las palabras que la anteceden y algunas de las que la siguen en cada aparición). En *A Glossary of Corpus Linguistics*, se define *concordancia* (*concordance*) de la siguiente forma: "Also referred to as key word in context (KWIC), a concordance is a list of all the occurrences of a particular search term in a corpus, presented within the context in which they occur –usually a few words to the left and right of the search term. A search term is often a single word although many concordance programs allow users to search on multiwords phrases, words containing wildcards, tags or combinations of words and tags" (Baker et al., 2006, pp. 42-43).

[5] Las *wordlists* consisten en una lista de todas las palabras de un texto (o textos) junto con el valor de frecuencia de aparición de cada palabra: "A list of all the words that appear in a text or corpus, often useful for dictionary creation. Word lists often give the frequencies of each word (or token) in the corpus. Words are most usually ordered alphabetically, or in terms of frequency, either with a raw frequency count and / or the percentage that the word contributes towards the whole texts […] Word lists are needed when calculating key words" (Baker et al., 2006, p. 169).

[6] Por otra parte, para McEnery y Hardie (2012), las colocaciones constituyen la abstracción estadística de las concordancias (p. 41), ya que en las colocaciones se recuperan las palabras que tienden con más probabilidad a aparecer juntas en ciertos contextos. En *A Glossary of Corpus Linguistics*, se define *collocation* de la siguiente forma: "phenomenon surrounding the fact that certain words are more likely to occur in combination with other words in certain contexts. A collocate is therefore a word which occurs within the neighbourhood of another word" (Baker et al., 2006, pp. 36-37). Para una discusión teórica del concepto de *colocación* en español, pueden consultarse los trabajos de Margarita Alonso Ramos (1994) y, especialmente, Gloria Corpas Pastor (2001).

[7] Para McEnery y Hardie, las palabras clave constituyen la abstracción estadística de las listas de frecuencias de palabras (2012, p. 41). En este caso, se trata de generar una lista con las palabras estadísticamente más significativas. En *A Glossary of Corpus Linguistics*, se define *keyword* de la siguiente forma: "A word

2.2. Herramientas procedentes de las comunidades de Humanidades Digitales

Por otra parte, otras herramientas de análisis textual, como Voyant Tools (Sinclair et al., 2016) o Corpus Explorer (Rüdiger, 2018a) han surgido dentro de las comunidades de HD.

Voyant Tools es una herramienta de explotación de corpus desarrollada por un equipo liderado por Stéfan Sinclair (McGill University) y Geoffrey Rockwell (University of Alberta). Permite subir un corpus y cuenta con una interfaz de fácil uso que ofrece diferentes opciones de visualización, como nubes de palabras, datos principales del corpus –número de palabras (*word tokens*), número de palabras concretas (*word types*) o el índice de densidad léxica–, palabras más frecuentes (y picos llamativos en la frecuencia de estas palabras), concordancias, o redes de palabras, entre otros. Además, es especialmente útil si queremos realizar un primer vistazo rápido de cómo se ve nuestro corpus a distancia, o como punto de partida sobre el que luego realicemos análisis más complejos[8]. Recientemente, esta herramienta ha sido reconocida con el premio trienal Antonio Zampolli, otorgado por la ADHO[9].

Corpus Explorer, por su parte, ha nacido dentro del seno de la comunidad alemana de HD, creado por Jan Oliver Rüdiger (Universidad de Siegen), y fue presentado en el congreso DHd de 2018 (Rüdiger, 2018b). Su autor lo describe como el resultado de crear una herramienta para estudios que aúnen, por un lado, Hermenéutica y Lingüística de corpus, y por otro, estudios lingüísticos y literarios. Cuenta con las habituales opciones de las herramientas de corpus, de forma que permite subir un corpus y hacer concordancias. Además, permite anotarlo y filtrar por categoría morfológica, entre otros. Quizás el principal inconveniente de este software es que tanto el programa como la documentación solo se encuentran disponibles actualmente en alemán, por lo que aún no es muy conocido ni usado fuera de Alemania, aunque allí esté teniendo un éxito considerable entre los romanistas.

2.3. Herramientas procedentes de la Textometría francesa

TXM (Heiden, 2018) ha sido desarrollado dentro del proyecto francés de Textometría –*Textométrie*– con sede en la Universidad de Lyon (Heiden, 2010; Heiden et al., 2010). Esta línea de investigación tiene sus comienzos en la Francia de los años 70 y puede definirse como la aplicación de una serie de cálculos lingüísticamente significativos y matemáticamente sólidos al análisis metódico de colecciones de textos (Pincemin & Heiden, s.f.). La difusión de esta metodología en dicho país ha provocado que TXM esté establecido como uno de los programas de explotación textual más usados. Frente a otros programas de análisis textual, ofrece la ventaja de que permite

which appears in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size. Usually log-likelihood or chi-squared tests are used to compare two-word lists in order to derive keywords" (Baker et al., 2006, p. 97).

[8] *The Programming Historian* ha publicado un tutorial en español sobre cómo realizar análisis textuales con Voyant Tools, preparado por Silvia Gutiérrez (2019). Además, Rockwell y Sinclair (2016) ofrecen explicaciones sobre la interpretación de textos con Voyant en su monografía *Hermeneutica. Computer-assisted Interpretation in the Humanities*.

[9] La web con los detalles del premio se encuentra accesible desde: https://adho.org/awards/antonio-zampolli-prize/.

trabajar con texto marcado y etiquetado en XML-TEI. Tiene el inconveniente de que para el etiquetado morfológico solo es compatible con Tree-Tagger (Schmid, 1994), a pesar de que existen en la actualidad herramientas más fiables y con las que se obtienen mejores resultados.

3. LITCON. LITERARY CONCORDANCES

3.1. Presentación y motivación

Como puede apreciarse, ninguna de las herramientas disponibles ha sido creada con el objetivo específico de trabajar con textos poéticos, y, en consecuencia, no tienen en cuenta cuestiones de máxima importancia en este tipo de textos como pueden ser los cortes de verso, los títulos de los poemas o el número de verso.

Litcon, *Literary Concordances*, se encuentra actualmente en su versión 1.5, ha sido programado en Java con la ayuda de un informático, y, en consecuencia, es un software multiplataforma, que funciona en Windows (véase la Figura 1), Mac (véase la Figura 2) y Linux, con el único requerimiento de que Java esté instalado y actualizado[10]. Contiene distintas opciones de análisis textuales. Para su diseño, se ha tenido en mente especialmente el análisis de textos poéticos, pero se ha procurado que pueda ser de utilidad para cualquier investigador en estudios literarios. En este sentido, se ha preparado principalmente para estudiosos de textos poéticos en español (y no exclusivamente para humanistas digitales), que son en su mayoría usuarios con pocos conocimientos informáticos. Por esta razón, se ha priorizado la sencillez de uso del software y que su interfaz fuera lo más intuitiva posible sobre la creación de una herramienta de mayor complejidad. Debido al público al que está dirigido, la herramienta está pensada para trabajar con formato de texto plano. Su aplicación a textos poéticos ha producido ya resultados de interés (Hernández-Lorenzo, 2020; Hernández-Lorenzo, 2021)[11].

Figura 1. Inicio de Litcon con la pantalla de créditos en Windows. Fuente: elaboración propia.

[10] Este puede descargarse desde su web, accesible desde: https://www.java.com/es/download/.
[11] El código y el ejecutable del programa se encuentran disponibles para la comunidad investigadora en el siguiente repositorio: https://github.com/lamusadecima/litcon.

Como puede apreciarse en las Figuras 1 y 2, Litcon cuenta con un entorno fácil de utilizar, en el que las diferentes herramientas se muestran como pestañas en la parte superior de la ventana, mientras que en la esquina derecha se ofrece la opción de cambiar la lengua del programa entre español e inglés (véase la Figura 2). El resto de la ventana muestra los contenidos de la sección en uso. Se encuentran disponibles las siguientes: 1) Visor fichero; 2) Concordancia; 3) Listado; 4) *Wordlist* (o lista de palabras); 5) Contraste; 6) Palabras clave; 7) Etiquetado; 8) POS; 9) POS II; 10) POS III; 11) Corpus. En cada una de ellas aparecen una serie de pasos u opciones, obligatorias u opcionales, que, para mayor claridad, se encuentran numeradas en todas las pestañas, menos Visor Fichero, que no lo necesita al ser más sencilla.

Figura 2. Inicio de Litcon con la pantalla de créditos en Mac e inglés como idioma. Fuente: elaboración propia.

4. CASO DE USO DE LITCON CON LA POESÍA DE FERNANDO DE HERRERA

4.1. Corpus y preparación previa

A continuación, se presenta un caso de uso del programa Litcon, en el que el corpus utilizado será la poesía del escritor sevillano Fernando de Herrera (1534-1597). Sus textos poéticos se nos han transmitido a través de distintos testimonios, entre los cuales destacan dos ediciones impresas: la edición de *Algunas obras* (1582), conocida como *H* y considerada tradicionalmente como la más autorizada; y la edición de *Versos* (1619), conocida como *P* y en torno a la cual gira buena parte de la polémica del drama textual herreriano[12]. La primera, con 91 poemas, constituye una selección de la poesía herreriana preparada por el propio Herrera y fue publicada en vida del poeta, quien incluso revisó y corrigió pruebas de imprenta. En cambio, la segunda, que contiene 365 poemas supuestamente herrerianos, apareció póstumamente, fue preparada por el pintor Francisco Pacheco (1564-1644), y, de acuerdo con la crítica, muestra importantes diferencias estilísticas con la edición de 1582, entre las cuales destacan las correcciones y modificaciones que presentan algunos de los poemas conservados a través de otros testimonios. Algunos estudiosos habían intentado recoger estas diferencias en listas realizadas manualmente, con los consiguientes errores y limitaciones de este enfoque con un corpus tan amplio de textos (Macrí, 1972). Por tanto, se trata de un caso de gran interés para realizar una exploración estilística cuantitativa. En las siguientes

[12] Para una completa exposición de esta polémica, véase el trabajo de Montero (2021).

páginas se analizarán algunos de los rasgos estilísticos de los poemas de 1619 –en ocasiones, en comparación con los poemas de 1582 y con el resto de poemas de autoría segura transmitidos mediante otras fuentes– a través de las diferentes funcionalidades que presenta Litcon.

Para poder procesar los poemas de Herrera en Litcon, estos han sido digitalizados a partir de la edición más autorizada y completa de su poesía (Herrera, 1975) mediante OCR y una posterior revisión manual para corregir los errores de este. El texto resultante se encuentra en formato de texto plano con codificación Unicode UTF-8, que es el aceptado por la mayoría de los programas de análisis textual, entre ellos Litcon. Además, los títulos de los poemas se han marcado entre almohadillas (#), porque de esta forma podrán ser recuperados por el programa. Del mismo modo, se incluyen los números de verso de cinco en cinco al comienzo de la línea.

4.2. Análisis de concordancias

En primer lugar, se ha realizado un análisis de concordancias de los poemas de *P* mediante la funcionalidad Concordancia de Litcon. Más específicamente, dada la importancia de los términos relacionados con la *osadía* en la poesía de Herrera, se han generado las concordancias de todas las formas que comparten la raíz léxica *osad-*. Esto ha podido hacerse gracias a la implementación de comodines de búsqueda en esta herramienta y que pueden consultarse haciendo clic en la ayuda de buscar (véase la Figura 3). Se ha decidido implementar comodines de búsqueda y no expresiones regulares, puesto que el uso de estas requiere que el usuario se encuentre familiarizado con las mismas. En cambio, los comodines de búsqueda desarrollados, que se complementan con unas sencillas instrucciones en la ayuda de buscar, pueden ser empleados fácilmente por estudiosos en poesía con pocos conocimientos informáticos.

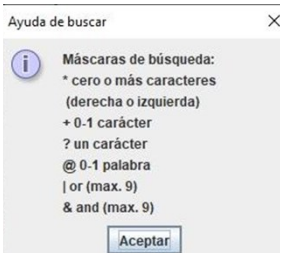

Figura 3. Máscaras de búsqueda que ofrece la herramienta Concordancia de Litcon, a través de la ventana *Ayuda de buscar*. Aquí se indican todos los comodines o máscaras de búsqueda que permite la generación de concordancias. Fuente: elaboración propia.

Tras subir el fichero textual que contiene los poemas de *P* y establecer las concordancias deseadas en el cuadro de buscar escribiendo *osad** (de forma que se recuperen todas las palabras que contengan la raíz léxica *osad-*), se ha marcado la opción *Datos de poema*, de forma que se mostrará también en la lista de concordancias el título del poema en el que se halla cada una y el número de verso. A continuación, se ha hecho clic en el botón *Ejecutar* y se han obtenido las concordancias (véase la Figura 4). El total de palabras encontradas (en este caso, 110) se indica en rojo en la esquina superior derecha de la tabla de concordancias. Los resultados se ordenan por orden de aparición en el texto y se presentan teniendo en cuenta los cambios de líneas, por lo que se preservan las pausas de final de verso y la separación entre distintos patrones métricos. Entre las concordancias obtenidas se encuentran *osado*, *osadía* y *osadamente*, y se observa que estas for-

mas pueden encontrarse tanto al comienzo, como en mitad y al final de verso. Además de consultarse en la ventana del programa, la lista de concordancias puede guardarse como lista en un archivo TXT (mediante la opción *Guardar Txt*) o como tabla en un archivo CSV para abrir en programas como Microsoft Excel (*Guardar Excel*). Para facilitar todo este proceso, los pasos están numerados por el orden en que deben realizarse.

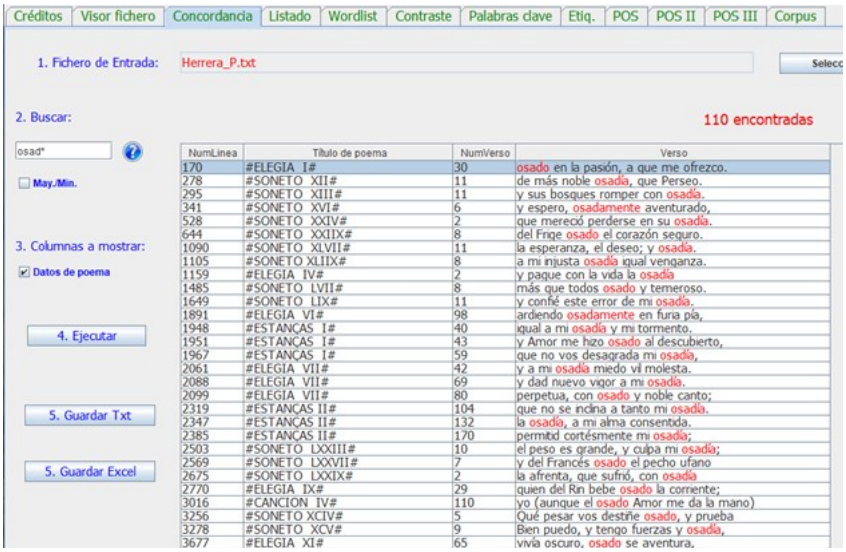

Figura 4. Concordancias de *osad-* en *Versos*, generadas a través de Litcon en Windows (herramienta Concordancia). Se ofrecen 110 resultados. Fuente: elaboración propia.

Asimismo, para consultar un mayor contexto de una concordancia en concreto, como los números de verso anteriores o posteriores, basta con hacer clic en esa concordancia y el programa cambiará a la herramienta Visor fichero, mostrando el lugar del texto concreto en el que aparece. Por ejemplo, al clicar la primera concordancia, *osado en la pasión, a que me ofrezco*, se comprueba que pertenece al último verso de uno de los tercetos encadenados que forman la elegía (véase la Figura 5). En la esquina inferior derecha se indica el formato de codificación que posee el texto y si el fichero tiene o no BOM[13].

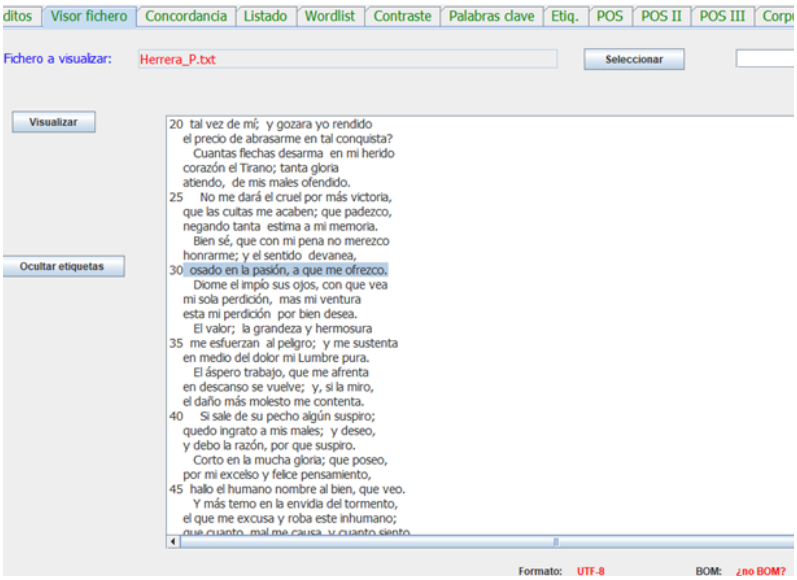

Figura 5. Contextualización de la primera ocurrencia de *osad-* en *Versos*, a través de la herramienta Visor de Litcon en Windows. El verso en el que aparece la ocurrencia se encuentra subrayado en color azul. En la parte inferior de la ventana se indica que el formato de codificación del texto es UTF-8 y que el fichero no tiene BOM. Fuente: Elaboración propia.

[13] Un BOM (*Byte Order Mark*) es una marca que se encuentra en ocasiones al inicio de un fichero de texto

También puede generarse automáticamente una lista de las concordancias de todas las palabras del texto a través de la herramienta Listado de Litcon. Se trata de una herramienta muy similar a la anterior de concordancias. Se diferencia en que, en vez de extraer las concordancias de una palabra concreta, se realiza un listado de todas las concordancias de todas las palabras del texto. Como puede observarse en la Figura 6, tras subir el texto de Versos, se indica el nombre con el que se guardará este listado. Opcionalmente, se pueden excluir palabras del análisis, o bien escribiéndolas en el cuadro de la opción sombreada, o bien subiendo un fichero que las contenga. En este caso, se han excluido las palabras agramaticales como determinantes, artículos, conjunciones y preposiciones, cuyas concordancias podrían tener menor interés, y se han ocultado las etiquetas del mismo (todo lo incluido entre almohadillas #), de forma que no se generen concordancias de las palabras incluidas en los títulos de los poemas. Se decide no marcar la opción May./Min., para no hacer separación entre mayúsculas y minúsculas (todas las palabras se convertirán a minúsculas). Al hacer clic en el botón Generar palabras se comprueba que, tras las exclusiones anteriores, las palabras de las que sí se generarán las concordancias son 6.348. Por último, se hace clic en el botón Generar concordancias, que pondrá en marcha la barra de progreso hasta llegar al 100 %, lo cual indica que ya se ha creado el archivo TXT con el listado de concordancias.

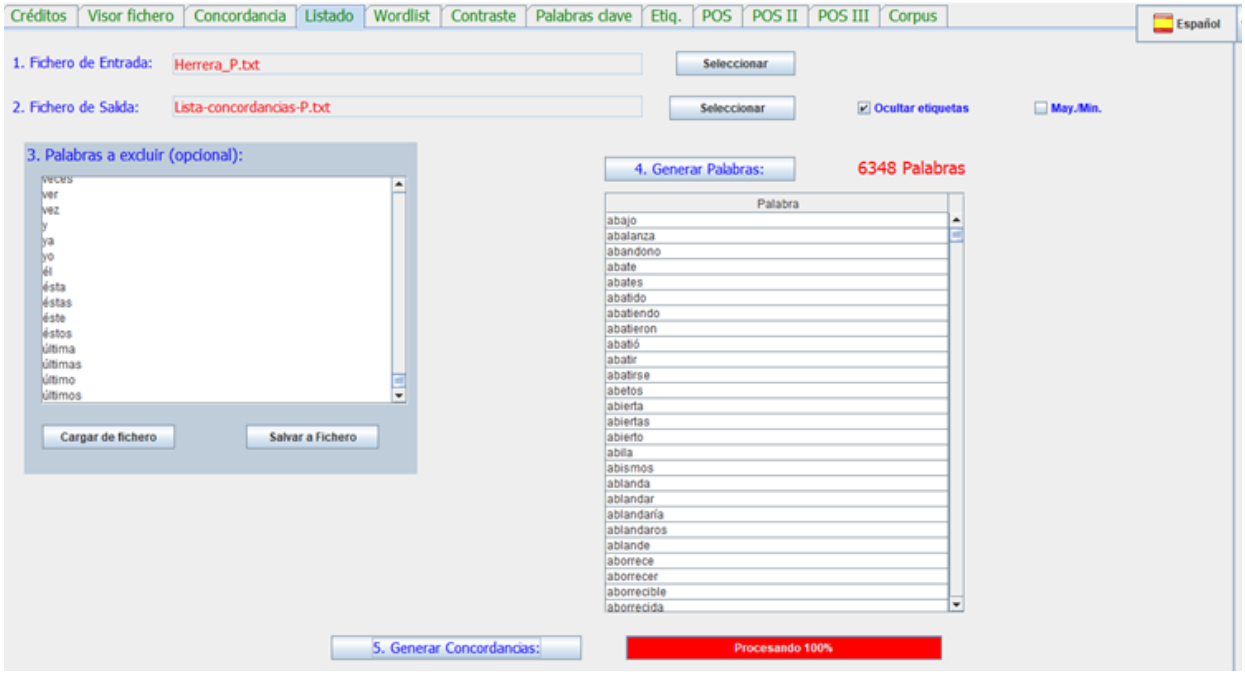

Figura 6. Ventana para generar el listado de todas las concordancias de *Versos*, a través de la herramienta Listado de Litcon. Al generar las palabras del texto, se obtienen 6.348, como aparece marcado en rojo encima de la esquina superior derecha del cuadro de palabras. Tras pinchar en *Generar concordancias*, se ha finalizado la creación de la lista, como indica la barra de progreso roja. Fuente: elaboración propia.

4.3. Análisis de contraste

Sin embargo, las herramientas anteriores no ofrecen información suficiente para comparar la aparición de una palabra en diferentes textos. Con este fin se puede recurrir a la herramienta Wordlist, o lista de palabras, que genera la lista de palabras concretas (word type) que aparecen

(frecuentemente con textos en Unicode) y recoge información sobre la codificación concreta del texto, por lo que resulta muy útil a la hora de detectar el formato.

en el texto, junto a su frecuencia de aparición (véase la Figura 7). Se ha introducido de nuevo el texto de Versos. Del mismo modo que en la pestaña anterior, se ofrece la posibilidad de excluir determinadas palabras del análisis (Palabras a excluir), así como de ocultar etiquetas y distinguir entre mayúsculas y minúsculas (casilla May./Min., que no se ha marcado). En este caso, no se ha excluido ninguna palabra con el objetivo de generar la lista completa de palabras de esta obra. Sí se ha marcado nuevamente la opción de ocultar etiquetas para que no se tengan en cuenta las palabras incluidas en los títulos de los poemas. También se puede marcar la opción de convertir el texto a caracteres ASCII (Carac. Ascii), pero no es recomendable realizarlo para textos en español, por lo que no se ha marcado. Además, si se marca la casilla Frec./1000, se ofrecerá junto al número de ocurrencias de cada palabra la frecuencia por mil, lo cual permitirá comparar estas frecuencias con las de otro texto de distinto tamaño. Tras generar la wordlist, se indica en rojo tanto el número de palabras concretas (word type) como el número total de palabras del texto (word tokens). Se comprueba que el número de palabras concretas de Versos es 6.636, mientras que el número total de palabras asciende a 71.834. La wordlist o lista de palabras puede ordenarse tanto por orden descendente de frecuencia (por defecto) como alfabéticamente, y puede guardarse a un fichero externo tanto TXT (Guardar Txt) como CSV (Guardar Excel).

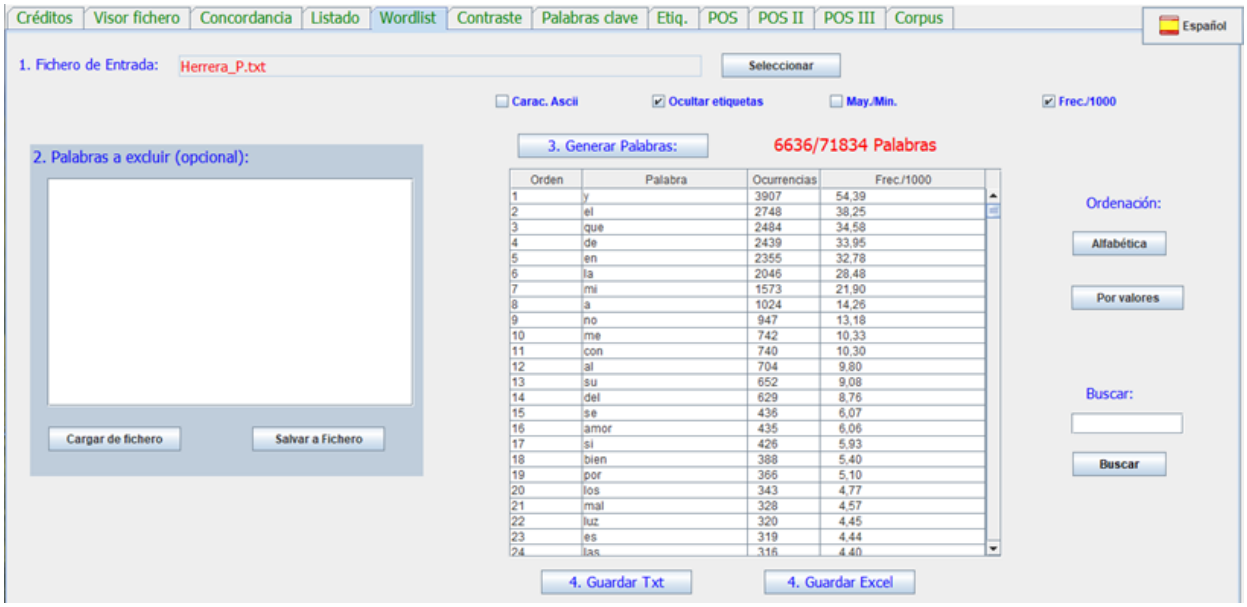

Figura 7. Lista de frecuencias de palabras de *Versos*. Estos datos han sido extraídos a través de la herramienta Wordlist de Litcon en Windows. Tras generar las palabras en el siguiente paso, se indica en color rojo, encima de la esquina superior derecha del cuadro de palabras, el número de palabras concretas o *word types* (6.636) seguido del número total de palabras o *word tokens* (71.834). Fuente: elaboración propia.

La herramienta Wordlist también dispone de un cuadro de búsqueda para buscar una palabra concreta dentro de la lista de palabras del texto. Tras buscar la palabra osadía, se ha podido comprobar que esta se encuentra en la posición 166 del orden de frecuencias, y consta de 54 ocurrencias, que tomando en cuenta la extensión de la obra completa, se traduce en una frecuencia por mil de 0,75 (véase la Figura 8). Tras repetir este proceso con los poemas de 1582 se comprueba que en estos la palabra osadía tiene 12 ocurrencias y una frecuencia por mil de 0,64. Por tanto, en los textos de 1619 se produce un ligero incremento en el uso de este término.

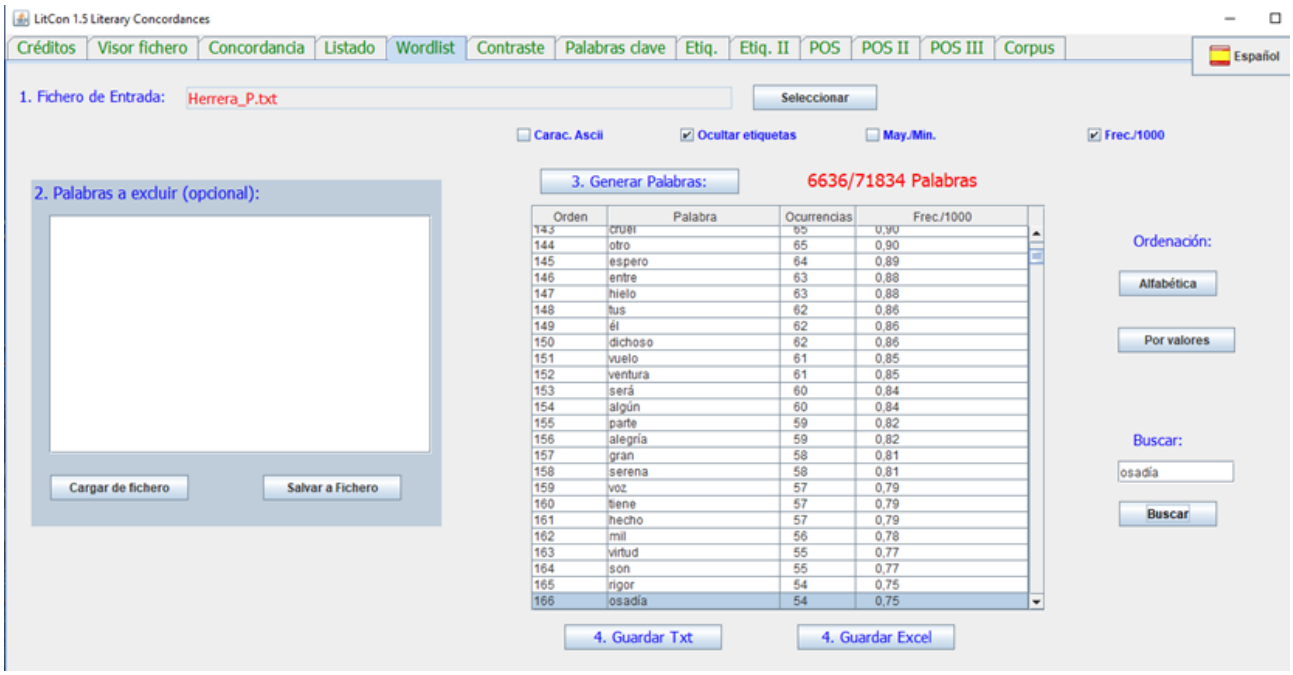

Figura 8. Posición de la palabra *osadía* en la lista de palabras de *Versos*, junto a su número de ocurrencias (54) y su valor de frecuencia por mil (0,75). Estos datos han sido extraídos a través de la herramienta Wordlist de Litcon en Windows. Se ha buscado la palabra *osadía*, que queda seleccionada en azul en el cuadro. Fuente: Elaboración propia.

A continuación, se analizan las palabras coincidentes y únicas de los poemas de *Algunas obras* y *Versos* a través de la herramienta Contraste. Esta permite comparar dos textos (o conjuntos de textos, como se indica al pinchar en el botón de ayuda) y señalar sus similitudes y diferencias en términos léxicos, mostrando las palabras comunes a ambos y las palabras que solo aparecen en cada uno de ellos. Del mismo modo que en las pestañas anteriores, se puede subir un fichero con palabras que se deseen excluir del análisis, y se pueden marcar las casillas de ocultar etiquetas, distinguir entre mayúsculas y minúsculas y convertir a caracteres ASCII. En este caso únicamente se ha marcado la primera de estas casillas para excluir los títulos de los poemas del análisis. Tras generar el contraste, se obtienen tres listas: una primera a la izquierda con las palabras únicas del primer fichero (es decir, *Algunas obras*, *H*), una segunda en el centro con las palabras comunes a ambos ficheros (las que aparecen tanto en *H* como en *P*), y una última con las palabras únicas del segundo fichero (*Versos*, *P*) a la derecha (véase la Figura 9). Encima de cada lista, en color rojo, se indica el número de palabras, y cada una de las listas puede guardarse tanto en formato TXT como CSV. *H* y *P* tienen en común 3.225 palabras, hay 166 formas que son únicas de *H*, y 3.411 palabras son únicas de *P*.

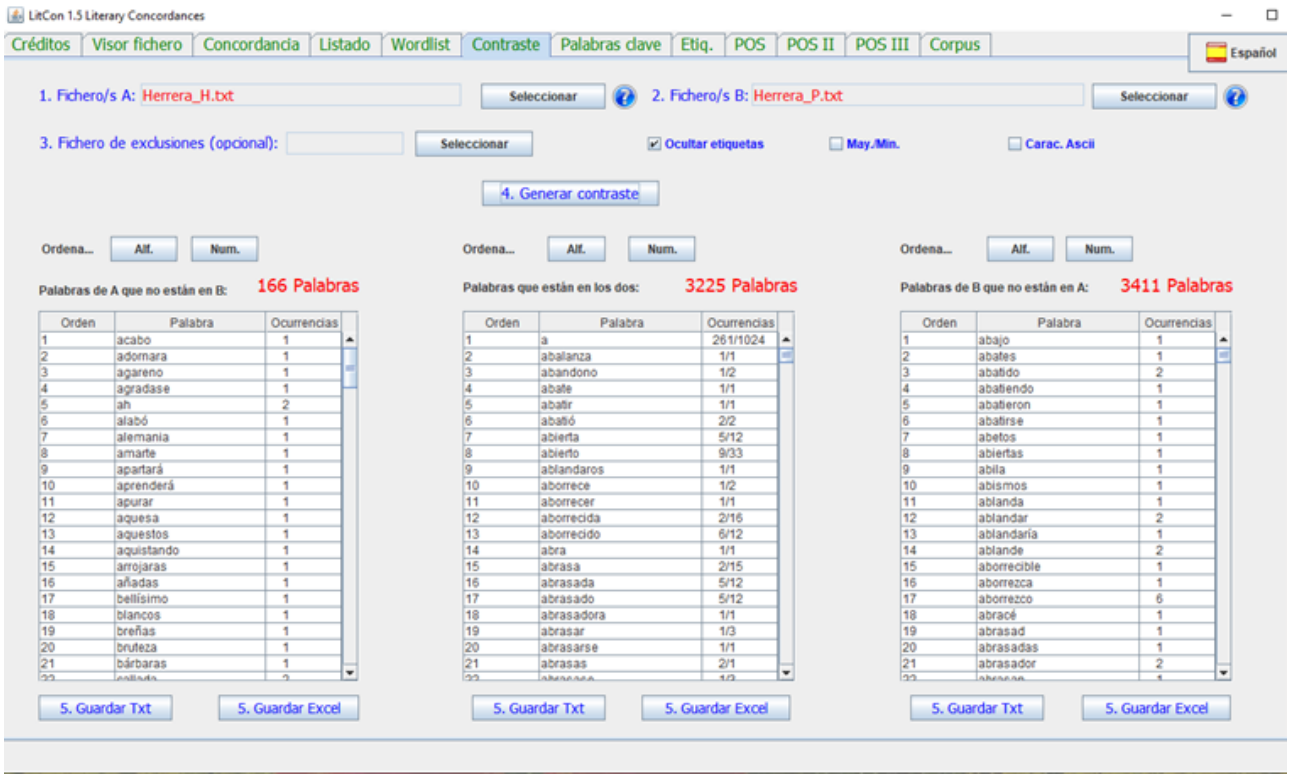

Figura 9. Resultados del análisis de contraste entre *H* y *P* a través de la herramienta Contraste de Litcon en Windows. Se ha seleccionado como Fichero A *Algunas obras*, y como Fichero B, *Versos*. No se ha seleccionado ningún fichero de exclusiones de palabras. Los resultados muestran que hay 3.225 palabras comunes a *H* y *P*, mientras que en *H* hay 166 palabras que no aparecen en *P*, y *P* tiene 3.411 palabras que no aparecen en *H*. Fuente: elaboración propia.

4.4. Análisis de palabras clave

Otro análisis comparativo de dos corpus textuales que puede ofrecer gran interés es el análisis de palabras clave. Este se ha realizado a través de la herramienta Palabras clave de Litcon, que permite generar una lista de palabras clave o *keywords*, que incluye las palabras estadísticamente más significativas de un corpus frente a otro. Para ello, Litcon utiliza por defecto la medida estadística *log-likelihood*, que es más apropiada para corpus relativamente pequeños, como suelen ser los literarios y, especialmente, los poéticos[14] (Rayson & Garside, 2000). Se ha utilizado como fichero de entrada el texto completo de *H*, del cual se extraerán las palabras clave, y como corpus de referencia, los poemas completos de *P*. Al igual que en las opciones anteriores, es posible excluir determinadas palabras del análisis mediante un fichero de exclusiones. Se encuentran disponibles de nuevo para marcar las casillas para ocultar etiquetas, distinguir entre mayúsculas y minúsculas y convertir a codificación ASCII. Como puede apreciarse en la Figura 10, tras generar las (palabras) claves, se obtienen dos listas: la de la izquierda, que muestra las palabras clave con valor positivo (aquellas con una relevancia estadística excepcionalmente alta), y la de la derecha, que muestra las palabras clave con valor negativo (aquellas con una relevancia estadística excepcionalmente baja). Las claves positivas corresponden a las palabras más destacadas estadística-

[14] En Lingüística de corpus es habitual manejar corpus de millones o billones de palabras. Por ello, los corpus literarios suelen ser excesivamente pequeños en comparación. La fórmula de *log-likelihood* implementada en Litcon puede encontrarse en el trabajo citado de Rayson y Garside (2000).

mente de *H* frente a *P* y las claves negativas, a las menos destacadas de *H* frente a *P* y más representativas de *P* frente a *H*. El número total de palabras en cada una de las listas se indica en color rojo encima de cada una: 2.559 palabras en las claves positivas y 832 palabras en las claves negativas. Cada una de las listas puede guardarse en formato TXT y CSV. Gracias a estos resultados, puede comprobarse que muchas de las palabras estadísticamente más representativas de los poemas de 1582 están relacionados con la égloga venatoria que contiene (*jabalí*, *ciervo*, *cazadora*, *prado*, *garza*, *halcón*…), subgénero al cual no pertenece ninguno de los poemas de 1619, aunque la más relevante es el pronombre *os*. En cambio, las claves negativas muestran que el pronombre *vos* es más representativo de *P*, así como palabras como *esplendor*, *ardor*, *muerte* o *ausencia*.

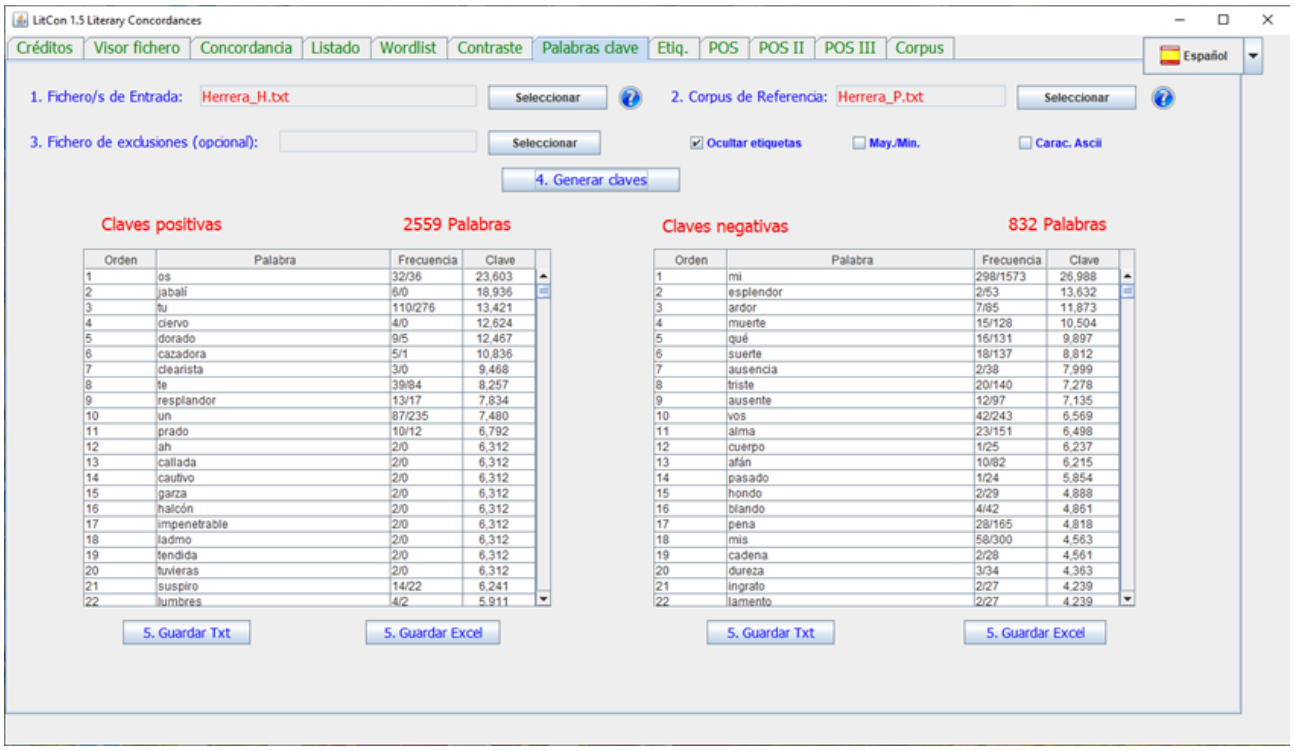

Figura 10. Resultados del análisis de palabras clave de *H* (utilizando *P* como corpus de referencia) a través de la herramienta Palabras clave de Litcon en Windows. Tras generar las palabras clave, se obtienen 2.559 palabras clave positivas y 832 palabras clave negativas. Fuente: elaboración propia.

4.5. Etiquetado morfológico (*Part-of-Speech*)

Aunque los análisis con corpus no etiquetados pueden ser de gran interés, también es habitual en Estilística de corpus y computacional trabajar con corpus en los que se han etiquetado determinados rasgos lingüísticos. Entre estos, destaca el etiquetado morfológico a través de las etiquetas *Part-of-Speech*. A continuación, se presenta cómo se ha realizado el etiquetado morfológico automático de los poemas de *H* a través de la funcionalidad habilitada en Litcon.

Los programas de Procesamiento de Lenguaje Natural y etiquetado morfológico automático suelen incluir algoritmos que contribuyen a mejorar los resultados de forma significativa. Estos algoritmos son el resultado de testar el programa con grandes cantidades de textos, de los que recogen información sobre el contexto determinado en el que una palabra tiene una función concreta, así como su probabilidad de aparición, por lo que resultan de gran utilidad para ayudar a decidir

al programa de etiquetado cuál es la etiqueta más correcta en el caso de que una palabra pueda tener dos o más etiquetas (Jurafsky & Martin, 2021)[15]. Por estas razones, se desarrolló una función de etiquetado compatible con el software de Procesamiento del Lenguaje Natural FreeLing (Padró, 2011). Se ha elegido que Litcon sea compatible con FreeLing, ya que, además de funcionar con una variedad de lenguas europeas, es la herramienta de etiquetado morfológico automático que ofrece en la actualidad los mejores resultados para textos en español.

El proceso de etiquetado automático de *H* con esta herramienta se ha llevado a cabo de la siguiente forma: en primer lugar, se eliminaron los números de verso, ya que podían dar problemas en el etiquetado. Después, tras realizar el etiquetado con FreeLing[16] y obtener el archivo de salida (véase la Figura 11), en la herramienta Etiquetado de Litcon se ha subido como fichero de entrada el texto sin etiquetar de H, como fichero Freeling el archivo de salida etiquetado de FreeLing, y se ha indicado la ubicación y el nombre con el que se guardará el archivo final etiquetado por Litcon (véase la Figura 12). Por último, se ha hecho clic en el botón *Etiquetar fichero*.

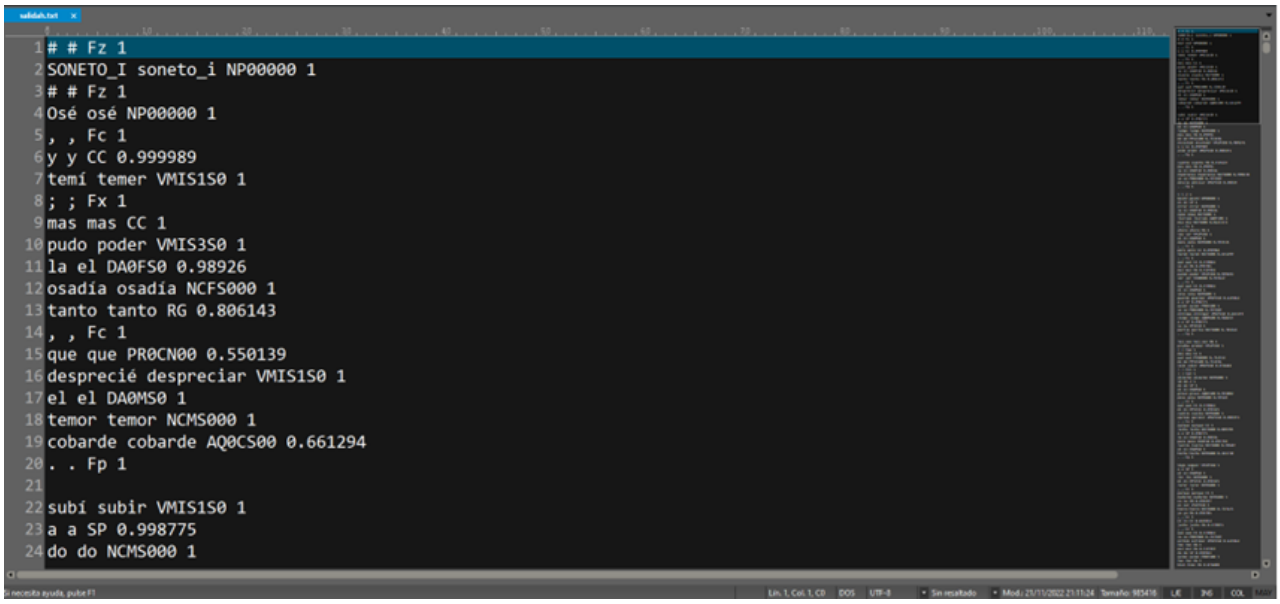

Figura 11. Fichero de salida en UltraEdit tras etiquetar H con el analizador de POS de FreeLing. Fuente: elaboración propia.

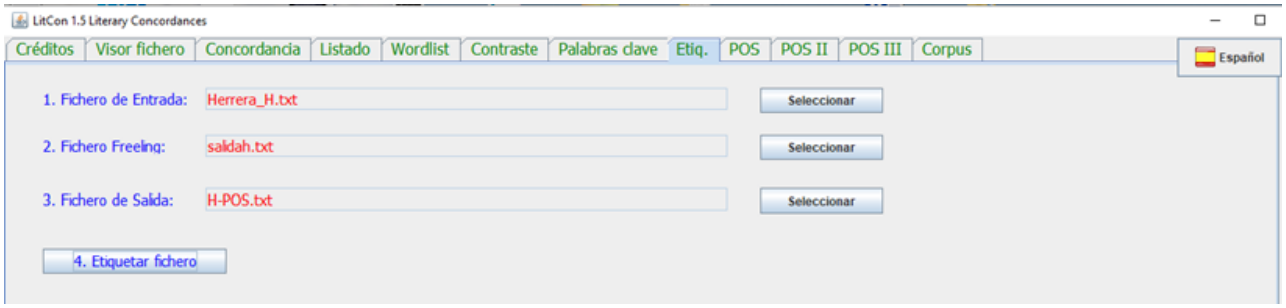

Figura 112. Ventana para generar el etiquetado morfológico de H a través de la herramienta Etiquetado de Litcon. Se ha seleccionado como fichero de entrada H, a continuación, se incluye el fichero de salida de FreeLing (*Fichero FreeLing*). Por último, se indica la ubicación y el nombre donde se guardará el archivo final etiquetado por Litcon. Fuente: elaboración propia.

[15] Valga como ejemplo la palabra *amo*, que dependiendo del contexto será un sustantivo o una forma verbal.

[16] Web de FreeLing, accessible desde: http://nlp.lsi.upc.edu/freeling/index.php/node/1. Las etiquetas de FreeLing se describen dentro de la documentación de la herramienta, en el siguiente enlace: https://freeling-user-manual.readthedocs.io/en/latest/tagsets/tagset-es/.

De esta forma, el etiquetado morfológico de *H* se ha realizado con FreeLing, y Litcon se ha encargado de reconvertir el resultado en un nuevo fichero de salida en el que se conserva la información sobre la separación versal y se obvian las etiquetas asignadas a los títulos de los poemas y a la puntuación (véase la Figura 13).

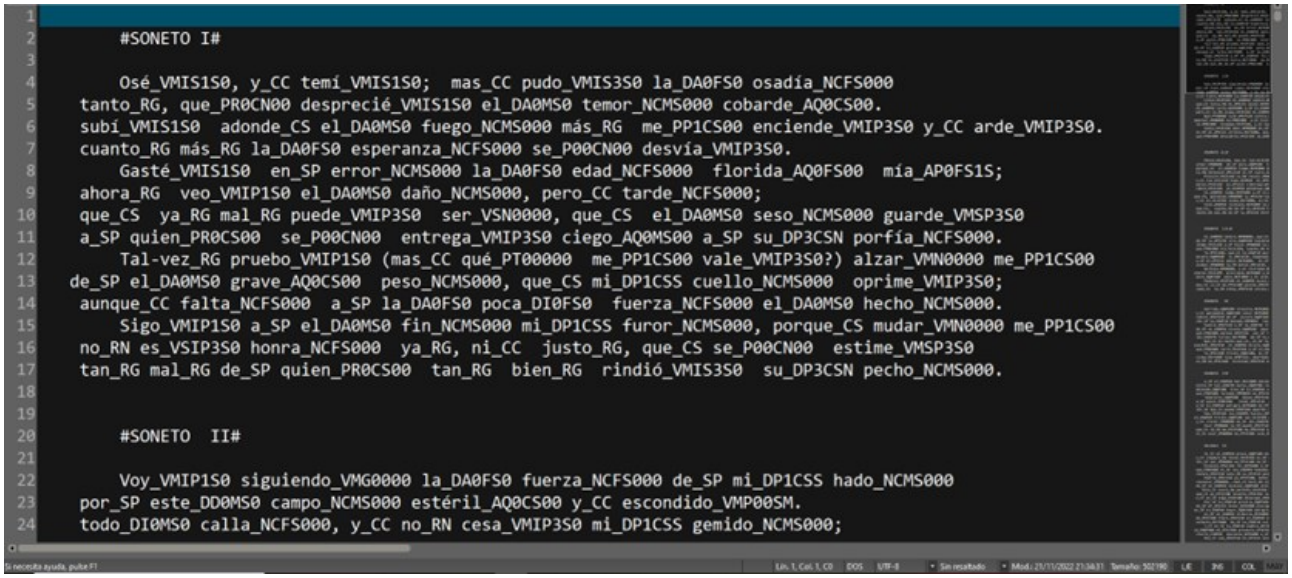

Figura 13. Fichero final etiquetado con Litcon usando el fichero original sin etiquetar y el fichero de salida de FreeLing. Vista del fichero en UltraEdit. Fuente: elaboración propia.

También puede visualizarse el texto etiquetado morfológicamente de *H* seleccionando el archivo en la herramienta Visor fichero (véase la Figura 14). Esta cuenta con una opción para ocultar las etiquetas morfológicas (botón *Ocultar etiquetas*), que deben ir precedidas de un guion bajo (véase la Figura 15).

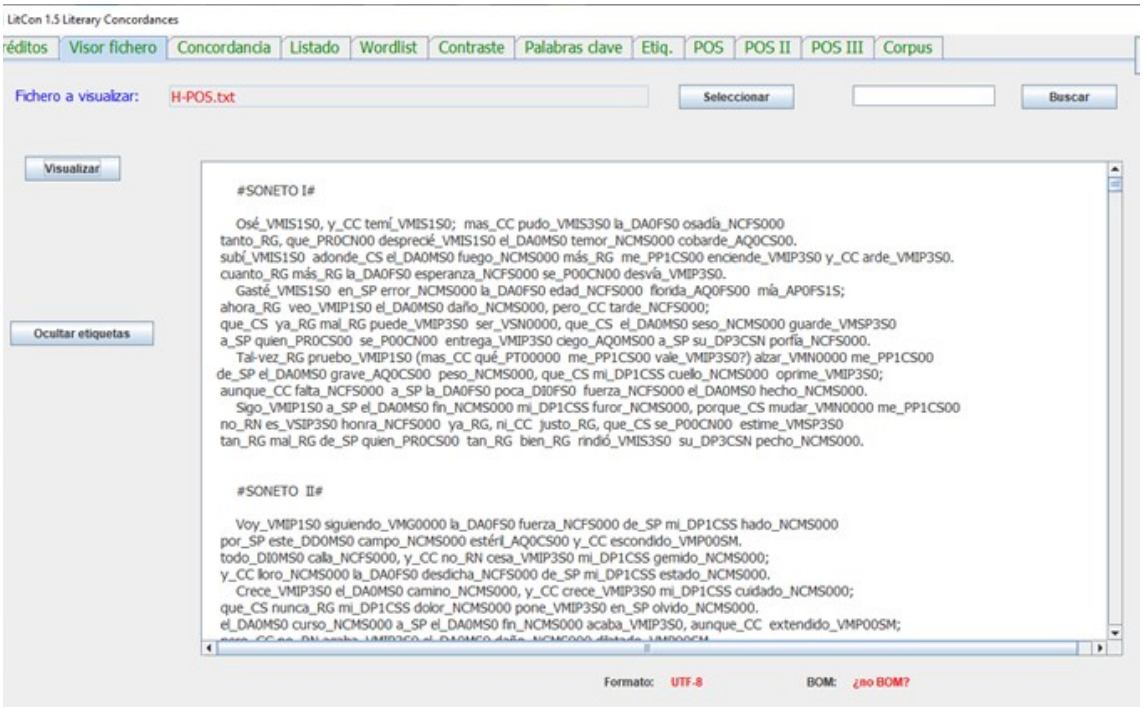

Figura 14. Visualización del texto etiquetado morfológicamente a través de la herramienta Visor fichero de Litcon en Windows. Se indica que el formato es UTF-8 y que no posee BOM. Fuente: elaboración propia.

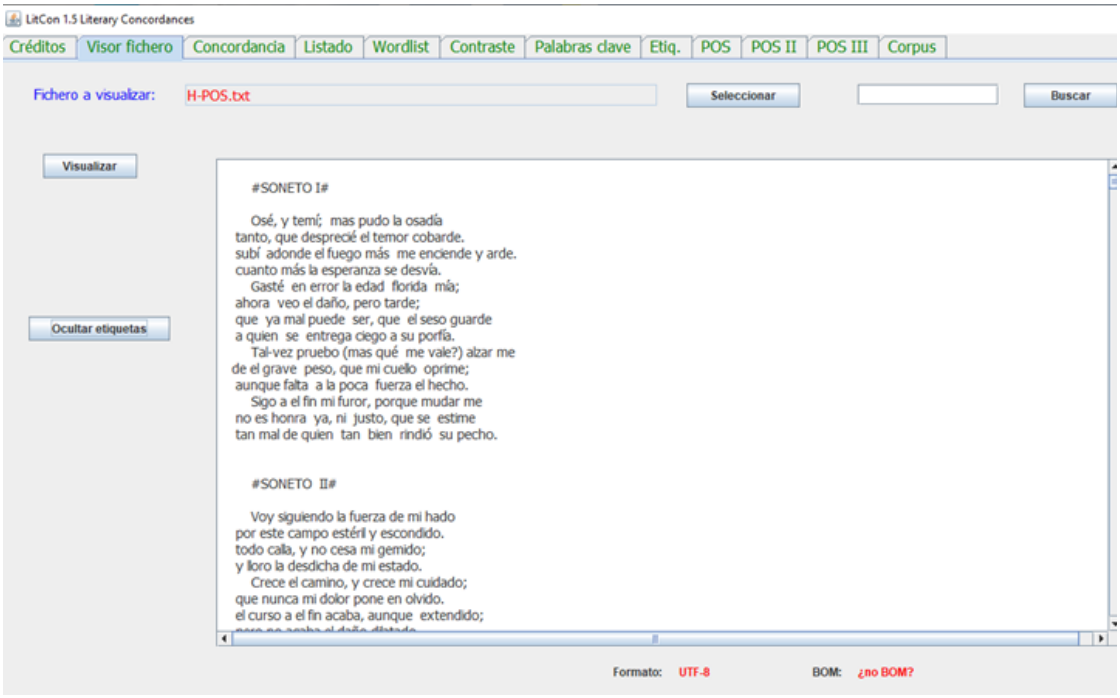

Figura 15. Visualización del texto etiquetado morfológicamente a través de la herramienta Visor fichero de Litcon en Windows. Se trata de la misma visualización de la figura 14, con la diferencia de que se ha pinchado en el botón *Ocultar etiquetas*, y estas han sido escondidas de la visualización, mostrándose solo las palabras. Fuente: elaboración propia.

4.6. Análisis morfológico

Una vez etiquetado el texto de *H* con la herramienta Etiquetado, se repitió el proceso con *P* y los poemas sueltos para tener etiquetado todo el corpus poético de Herrera. A continuación, este ha sido analizado con las herramientas de análisis morfológico que ofrece Litcon, que ayudan a explorar el texto ya etiquetado morfológicamente. Se ha atendido especialmente al análisis de la morfológico de *P*.

En primer lugar, se ha examinado el sintagma adjetivo calificativo seguido de nombre común en los textos de 1619. Con este fin, se ha utilizado la herramienta POS de Litcon, que permite extraer información sobre patrones morfológicos de dos categorías morfológicas (véase la Figura 16). Tras subir el fichero de entrada, el siguiente paso es indicar qué par de etiquetas se desea recuperar. Para mejorar los resultados obtenidos, la herramienta cuenta con dos pasos opcionales. El primero (*Descartar relación*) sirve para delimitar los signos de puntuación que no se permiten entre las dos categorías o etiquetas (por defecto aparece el punto), y, además, para textos en español se pueden marcar las casillas de ruptura por género y número, las cuales provocan que solo se nos muestren los resultados en los que las palabras concuerden en este sentido. El siguiente paso (*Relación con palabras en medio*), también opcional, permite que entre las dos categorías haya un número concreto de palabras, que puede fijar el usuario. En este estudio de caso, se ha mantenido el punto como signo de puntuación que rompe el sintagma que se desea analizar y se han marcado las casillas de ruptura por género y número, para que concuerden los adjetivos y nombres recuperados. Puesto que no se ha indicado ningún valor en *Relación con palabras en medio*, solo se recuperarán los casos en los que ambas categorías morfológicas aparezcan seguidas. Además, se ha marcado la casilla de datos de poema para que en la lista de resultados se nos muestre el título

del poema y el número de verso donde aparece el patrón/relación.

Al generar las relaciones, se nos presentan los resultados acompañados de estos datos y se nos proporciona también información general en rojo. En la primera línea, se nos indica cuántos ejemplos de esa relación concreta se han encontrado en el corpus, su frecuencia por mil y el número total de palabras del corpus (*word tokens*): se han recuperado un total de 2.866 sintagmas de adjetivo calificativo seguido de nombre común, lo cual se traduce en un 39,13 en frecuencia por mil. En la segunda línea, a continuación de *Estadística*, se ofrecen datos más generales sobre el texto, como la cantidad de elementos que incluye de cada categoría morfológica: se observa que los verbos son los más abundantes (hay un total de 12.200 en el texto). Entre los resultados se han obtenido sintagmas como *vano error*, *triste corazón*, *alegre semblante*, *libres almas* o *alto coro*. Estos pueden guardarse en formato TXT con el botón de la esquina inferior.

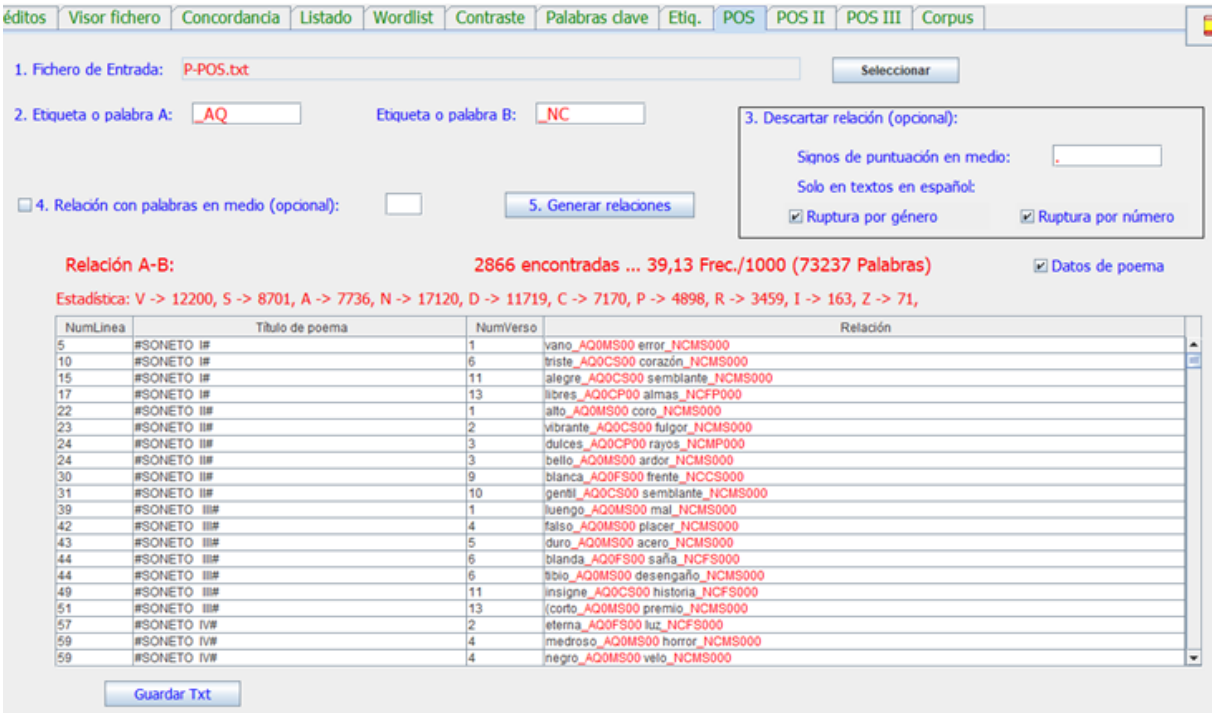

Figura 16. Resultados de adjetivo calificativo seguido de nombre común en *Versos*, extraídos mediante la herramienta POS de Litcon. Se detectan 2.866 resultados de esta relación – lo cual supone una frecuencia por mil de 39,13– teniendo en cuenta que el texto contiene 73.237 palabras. La información estadística de categorías indica que en el texto se cuentan 12.200 formas verbales, 8.701 preposiciones, 7.736 adjetivos, 17.120 sustantivos, determinantes, 7.170 conjunciones, 4.898 pronombres, 3.459 adverbios, 163 interjecciones y 71 numerales. Fuente: elaboración propia.

Seguidamente, se ha realizado una nueva búsqueda de patrones morfológicos en *P* a través de la herramienta POS II de Litcon. Frente a la pestaña POS, que únicamente permite buscar parejas de etiquetas / categorías, con POS II se puede buscar cualquier patrón y ofrece contexto tanto del texto anterior como del posterior a la relación buscada (véase la Figura 17). Como patrón de etiquetado, se ha definido el de artículo seguido de adjetivo calificativo, seguido a su vez de nombre común, y se han marcado las casillas de ruptura por género y número. Al pinchar en *Generar ítems*, se han generado los resultados: se han localizado 973 casos de patrones de artículo + adjetivo calificativo + nombre común, entre los cuales se encuentran *el triste corazón*, *el alto coro* y *el duro acero*. En la parte inferior de la ventana, estos pueden guardarse en formato TXT o CSV.

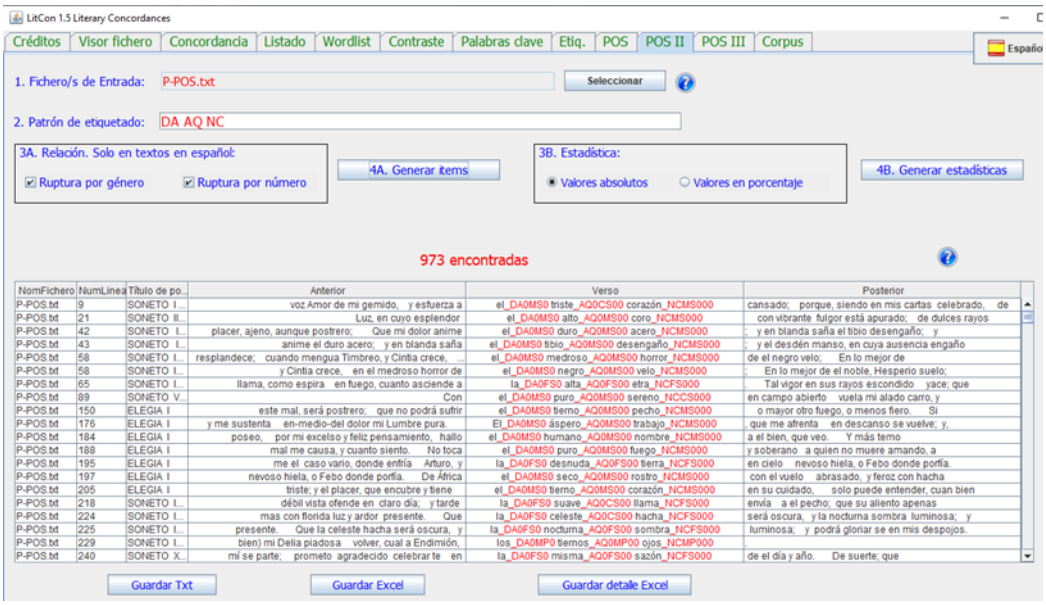

Figura 17. Resultados del patrón artículo + adjetivo calificativo + sustantivo en *Versos*, obtenidos a través de la herramienta POS II de Litcon. Se obtienen 973 resultados. Fuente: elaboración propia.

Además, la herramienta POS II de Litcon también aporta información más detallada sobre una categoría concreta. Para obtener estos datos, se ha introducido una única categoría o etiqueta en el cuadro de *Patrón de etiquetado*, en este caso, la de determinante artículo (DA). En el tercer paso, en lugar de atender al cuadro *3A. Relación…*, se han marcado los parámetros en el cuadro *3B. Estadística*, en el que se puede elegir entre valores absolutos y valores en porcentaje. En este caso, se han escogido los valores en porcentaje. Tras hacer clic en el botón *4B. Generar estadísticas*, han aparecido los resultados con los valores de ocurrencias (en porcentaje) de la categoría desglosados por poema y las veces que se repite cada valor (véase la Figura 18). Finalmente, en la parte inferior de la ventana, se encuentra el botón *Guardar detalle Excel*, que permite guardar los resultados de la columna de ocurrencias en un CSV que luego el usuario, por ejemplo, podrá llevarse a otros programas para aplicar test estadísticos.

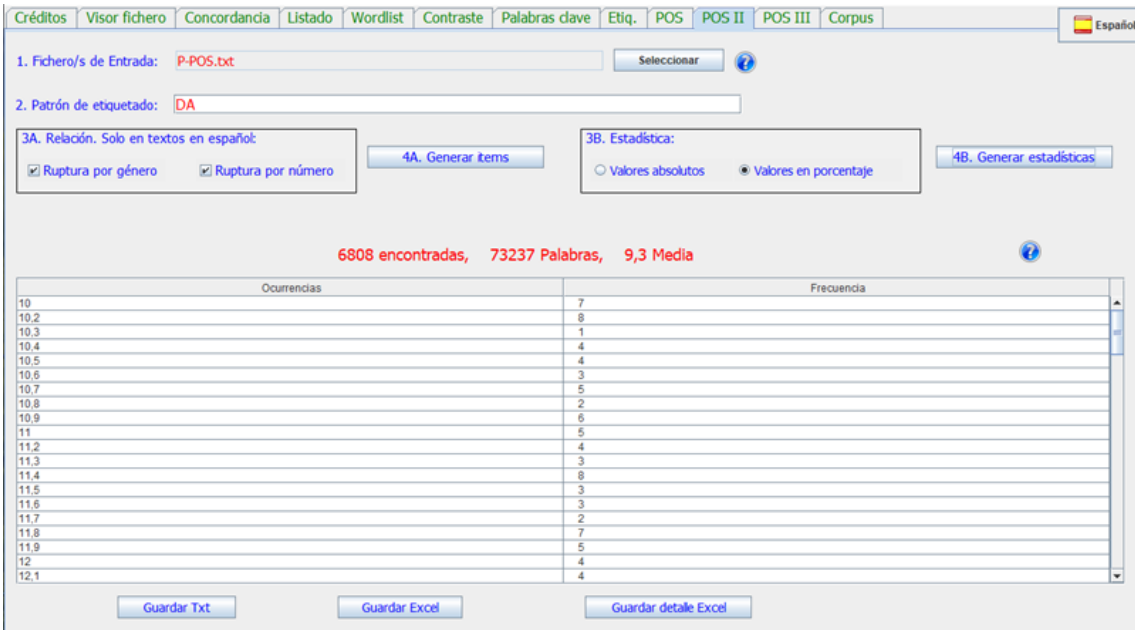

Figura 18. Resultados de cantidad de artículos por poema en *Versos*, obtenidos a través de la herramienta POS II de Litcon. Hay 6.808 ocurrencias de determinantes artículos en un total de 73.237 palabras y la media de porcentaje de ocurrencias de artículos por poema es 9,3. Fuente: elaboración propia.

Por último, se ha realizado un tercer análisis morfológico a través de la herramienta POS III de Litcon. Frente a las dos anteriores, POS y POS II, en las que el usuario debía introducir en el cuadro de búsqueda el patrón o patrones en los que estaba interesado, en esta última implementación, se selecciona un texto o varios (como se indica al clicar en el botón de ayuda), junto con el número de etiquetas en contacto en el que el usuario está interesado, y si se desea que se realice ruptura por género y número. Con estos datos, el programa genera automáticamente una lista de todas las relaciones existentes y sus frecuencias en tantos por mil que cumplen estos requisitos. En este caso, se han subido los tres ficheros etiquetados morfológicamente que contienen la totalidad de la poesía de Herrera, se ha seleccionado en el segundo paso que se desean generar relaciones de tres etiquetas, y en el tercer paso, se han marcado las casillas de ruptura por género y número. Tras hacer clic en *4. Generar relaciones*, se generan los resultados (véase la Figura 19). Estos pueden ordenarse pinchando en las cabeceras de las columnas: alfabéticamente (mediante *Relación*), por frecuencia total en todos los textos (mediante *Frecuencia 1/1000*), o por frecuencia en cada uno de los textos (mediante *F1*, *F2* o *F3*). En la parte inferior de la ventana, pueden guardarse los resultados obtenidos en formato TXT o CSV. Gracias a esta herramienta se ha podido comprobar, por ejemplo, que la relación de tres categorías morfológicas más frecuente en todo el corpus es la de preposición seguida de determinante artículo, seguido a su vez de nombre común (SP DA NC), con una frecuencia absoluta de 3.077 ocurrencias y una frecuencia por mil de 29,83. El mayor número de ocurrencias en valores absolutos se encuentran en el texto de P (1702), pero atendiendo a la frecuencia por mil, la frecuencia de aparición es mayor en el texto de B (32,05).

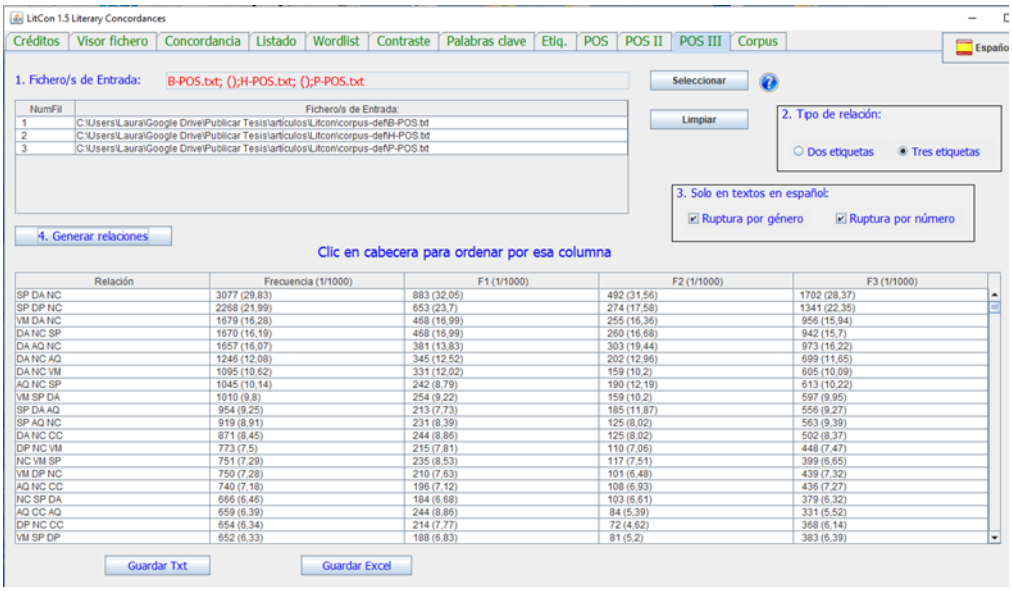

Figura 19. Resultados de relaciones de tres etiquetas morfológicas en *B*, *H* y *P*, obtenidos mediante la herramienta POS III de Litcon. Fuente: elaboración propia.

4.7. Muestreo aleatorio de poemas

Por último, aunque habitualmente en Estilística computacional y Estilometría se trabaja con los corpus completos, esto es, con la población completa de los textos que se quieren analizar, en ocasiones resulta conveniente hacer muestreo de los textos. Se ha producido una muestra aleatoria

de los poemas de *P* gracias a la herramienta Corpus de Litcon (véase la Figura 20). Esta última opción permite crear una muestra aleatoria de un fichero de texto con una extensión aproximada introducida por el usuario. Además, la muestra aleatoria se realiza tomando los textos completos de los poemas, de forma que se seleccionan composiciones aleatoriamente hasta que se completa una muestra de extensión aproximada al número de palabras marcado por el usuario. Y cada vez que se utiliza la opción, se genera una muestra diferente. En el caso de los poemas de *P*, tras subir el archivo con los poemas antes de ser etiquetado morfológicamente e indicar el nombre del fichero de salida con la muestra producida, se ha establecido el número de aproximado de palabras que debe tener la muestra (en este caso, en torno a 5.000). Al hacer clic en *4. Generar Corpus*, han aparecido como estadísticas de entrada el número total de palabras, líneas y poemas del texto completo de P, y como estadísticas de salida, los mismos datos referidos a la muestra aleatoria generada. Así, mientras que el texto completo de *P* consta de 365 poemas, que se traducen en 12.430 líneas y 71.834 palabras, la muestra aleatoria producida contiene 26 poemas, que constituyen un total de 826 líneas y 5.021 palabras.

Figura 20. Ventana para generar una muestra aleatoria de poemas con una cantidad aproximada de palabras a través de la herramienta Corpus de Litcon. *P* cuenta con 71.834 palabras, 12.430 líneas y 365 poemas, mientras que la muestra contiene 5.021 palabras en 826 líneas de 26 poemas. Fuente: elaboración propia.

5. CONCLUSIONES

En este artículo se ha realizado un breve repaso de los programas de análisis de corpus disponibles, procedentes de diferentes comunidades de investigación. Todos coinciden, sin embargo, en que se trata de herramientas genéricas que no tienen en consideración las particularidades formales de determinados tipos de textos, como es el caso de los textos poéticos en verso. Frente a los programas de análisis y explotación textual existentes, Litcon no solo presenta herramientas y opciones que no estaban disponibles en estos (como Contraste o Corpus), sino que en su diseño se ha prestado especial atención a las características de los textos literarios y, especialmente, poéticos. En este sentido, se ha atendido a detalles de gran importancia en estos textos, referentes a la disposición textual, como son las pausas de final de verso, que marcan la separación entre distintos patrones métricos, los números de verso y los títulos de los poemas.

Además, este artículo muestra cómo el uso de las diferentes herramientas que componen el programa Litcon pueden contribuir significativamente al estudio y análisis de textos literarios, y especialmente poéticos o textos en verso, como se ha visto en el estudio de caso de la poesía de Fernando de Herrera. Estas también pueden aplicarse y ser de utilidad para otras tipologías

textuales similares como las letras de canciones.

Por último, Litcon cuenta, como se ha visto, con una interfaz altamente intuitiva, que no requiere conocimientos previos de programación, por lo que resulta de fácil manejo tanto para humanistas digitales como para estudiosos con habilidades ofimáticas a nivel de usuario. La creación de Litcon contribuye, pues, al estado de la cuestión de los programas disponibles, ya que pone a disposición de los investigadores esta nueva herramienta y sus posibilidades.

REFERENCIAS BIBLIOGRÁFICAS

Alonso Ramos, M. (1994). Hacia una definición del concepto de colocación: De J. R. Firth a I. A. Mel'čuk. *Revista de Lexicografía, I*, 9-28. https://ruc.udc.es/dspace/bitstream/handle/2183/5383/RL_1-1.pdf?sequence=1

Anthony, L. (2013). A critical look at software tools in Corpus Linguistics. *Linguistic Research*, *30*(2), 141-161.

Anthony, L. (2022a). What can corpus software do? En A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (2nd ed.) (pp. 103-125). Routledge.

Anthony, L. (2022b). *Antconc* (4.1.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Baker, P., McEnery, A., & Hardie, A. (2006). *A Glossary of Corpus Linguistics*. Edinburgh University Press.

Corpas Pastor, G. (2001). En torno al concepto de colocación. *Euskera*, *XLVI*(1), 89-108.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, *8*(1), 107-121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Gutiérrez, S. (2019). Análisis de corpus con Voyant Tools. *The Programming Historian en español*, 3. https://programminghistorian.org/es/lecciones/analisis-voyant-tools

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. En *24th Pacific Asia Conference on Language, Information and Computation* (pp. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764

Heiden, S. (2018). *TXM* (0.7.9) [Computer software]. http://textometrie.ens-lyon.fr/?lang=fr

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. En S. Bolasco (Ed.), *Proc. Of 10th International Conference on the Statistical Analysis of Textual Data-JADT 2010* (Vol. 2, pp. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto. https://halshs.archives-ouvertes.fr/halshs-00549779

Hernández-Lorenzo, L. (2020). *Los textos poéticos de Fernando de Herrera: Aproximaciones desde la Estilística de corpus y la Estilometría*. https://idus.us.es/handle/11441/93465

Hernández-Lorenzo, L. (2021). Nueva Luz para la problemática de Versos: Una aproximación a su

léxico desde las Humanidades Digitales y los estudios de corpus. En J. Montero & P. Ruiz Pérez (Coords.), *De Herrera. Estudios reunidos con motivo del IV Centenario de Versos (1619)* (pp. 151-206). Universidad de Sevilla.

Herrera, F. de. (1975). *Obra poética* (J. M. Blecua, Ed.). Boletín de la Real Academia Española.

Hockey, S. (2004). The History of Humanities Computing. En S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 3-19). Blackwell.

Jannidis, F. (2014). *PyDelta* [Python]. Universität Julius-Maximilians Würzburg. https://github.com/cophi-wue/pydelta

Juola, P. (2005). A Prototype for Authorship Attribution Software. En P. Liddell, R. Siemens, A. Bia, M. Holmes, P. Baer, G. Newton, & S. Arneil (Eds.), *The International Conference on Humanities Computing and Digital Scholarship. The 17th Joint International Conference* (pp. 97-99). University of Victoria.

Jurafsky, D., & Martin, J. H. (2021). Sequence Labeling for Parts of Speech and Named Entities. En *Speech and Language Processing*. Prentice Hall.

Macrí, O. (1972). *Fernando de Herrera*. Gredos.

McEnery, A., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McEnery, A., & Wilson, A. (2001). *Corpus Linguistics. An Introduction*. Edinburgh University Press.

Montero, J. (2021). La transmisión de los textos poéticos de Fernando de Herrera: Estado de la cuestión y nuevas perspectivas. En *De Herrera. Estudios reunidos en el centenario de versos (1619)* (pp. 107-149). Editorial Universidad de Sevilla.

Padró, L. (2011). Analizadores multilingües en freeling. *Linguamática*, *3*, 13-20. http://upcommons.upc.edu/handle/2117/14772

Pincemin, B., & Heiden, S. (s. f.). *Qu'est-ce que la textométrie? Présentation. Site du projet Textométrie*. https://pages.textometrie.org/textometrie/Introduction

Rayson, P., & Garside, R. (2000). Comparing Corpora using Frequency Profiling. *Proceedings of the Workshop Comparing Corpora, held in conjuction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, *9*, 1-6.

Rockwell, G., & Sinclair, S. (2016). *Hermeneutica. Computer-Assisted Interpretation in the Humanities*. MIT Press.

Rojas Castro, A. (2013). Las Humanidades Digitales: Principios, valores y prácticas. *Janus: estudios sobre el Siglo de Oro*, *2*, 74-99. http://www.janusdigital.es/articulo.htm?id=24

Rüdiger, J. O. (2018a). *Corpus Explorer* (2.0) [Computer software]. https://notes.jan-oliver-ruediger.de/software/corpusexplorer-overview/

Rüdiger, J. O. (2018b). CorpusExplorer v2.0-Seminartauglich in einem halben Tag. *DHd2018. Kritik der digitalen Vernunft*, 28-30. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf

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

Scott, M. (2010). What can corpus software do? En A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (1st ed.) (pp. 136-151). Routledge.

Scott, M. (2016). *WordSmith Tools* (7.0) [Computer software]. Lexical Analysis Software. https://www.lexically.net/wordsmith/

Sinclair, S., & Rockwell, G. (2016). *Voyant Tools* [Web application]. https://voyant-tools.org/