Revista de Humanidades Digitales

NAVARRO COLORADO, Borja, CHICO RICO, Francisco, TOMÁS DÍAZ, David, PRIETO DE PAULA, Ángel Luis, CARRASCO, Rafael C., SÁNCHEZ MARTÍNEZ, Felipe y RIBES LAFOZ, María. *Proyecto ADSO. Análisis distante del soneto castellano de los Siglos de Oro (XVI y XVII)*. Universidad de Alicante, 2016-2018.

http://adso.gplsi.es/index.php/es/proyecto-adso/

Reseña realizada por:
María Ángeles HERRERO
mangherr@gmail.com

1. INTRODUCCIÓN

El Proyecto ADSO tiene como principal objetivo analizar aquellos rasgos comunes, semánticos y métricos de los sonetos escritos en lengua castellana de los Siglos de Oro (XVI y XVII). Las composiciones que forman el corpus del proyecto van desde las primeras muestras renacentistas con Garcilaso de la Vega, hasta las últimas manifestaciones del Barroco que son encarnadas por la mexicana sor Juana Inés de la Cruz. Para ello, se aplican técnicas computacionales como el análisis distante[1] y el macro-análisis de Jockers[2] y métodos innovadores como el procesamiento del lenguaje natural y *Text Mining*.

El proyecto ha sido financiado por la Fundación BBVA, gracias a la concesión de una de las Ayudas a Equipos de Investigación Científica en Humanidades Digitales de la convocatoria del 2016, con duración de dos años (2016-2018).

[1] La referencia más esencial al respecto es la de Moretti, F. (2007). *La literatura vista desde lejos*. Barcelona: Marbot ediciones.
[2] Se desarrolla en Jockers, M. L. (2013). *Macroanalysis. Digital Methods and Literary History*. Illinois: University of Illinois Press.

ISSN 2531-1786

Revista Digital de Acceso Abierto http://revistas.uned.es/index.php/RHD/index
Recibido 10/07/2018 – Aceptado 10/10/2018

Actualmente el proyecto se encuentra en su fase de finalización. Ha sido dirigido por el profesor Dr. Borja Navarro Colorado (IP) y su realización se ha desarrollado dentro del Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI) del Departamento de Lenguajes y Sistemas Informáticos (DLSI) de la Universidad de Alicante.

Figura 1. Página web del GPLSI (DLSI, UA).

2. VISUALIZACIÓN DEL PROYECTO: WEB Y REDES SOCIALES

El proyecto cuenta con una web principal que es accesible en el dominio http://adso.gplsi.es/index.php/es/proyecto-adso/, y fue creada en WordPress. Se trata de una web minimalista, en la que se usa una misma tipografía, no presenta diseños complejos, imágenes, ni gráficos. Está disponible en tres idiomas: castellano, catalán e inglés. Sin embargo, no todas las descripciones están traducidas, por lo que se dejan entrever algunos pequeños errores de traducción en algunas opciones y pestañas. A pesar de su diseño sencillo, se trata de una web primordialmente funcional y muy accesible, con un menú básico que viene encabezado por las entidades que sostienen el proyecto: la Fundación BBVA y la Universidad de Alicante.

Fig. 2. Principal web del proyecto ADSO.

Quedan bien definidas en ella las secciones descriptivas del proyecto como son los objetivos y el desarrollo (Corpus de métrica > Información), las y los miembros del equipo de investigación, la bibliografía en la que se apoya y una sección dedicada a un blog con algunas publicaciones breves que se refieren a actividades desarrolladas dentro del mismo proyecto. Hay que destacar, sobre todo, los apartados que dirigen al corpus de descarga de sonetos en el repositorio GitHub y especialmente de interés es el de la demo del sistema de escansión, aunque éste sólo es efectivo con versos endecasílabos.

Figura 3. Apartado de demostración del sistema de escansión.

Destaca, por tanto, la practicidad de la web, pero sería más adecuado el acceso a descarga o consulta con el despliegue a una nueva pestaña (Corpus de métrica > Descarga; Corpus de métrica > Consulta), puesto que al hacerlo se sale de la web principal. Cabe señalar que el acceso a GitHub puede resultar poco adecuado para aquellos usuarios no familiarizados con su interfaz a la hora de realizar la descarga del corpus, por lo que su accesibilidad debería ser más clara en este sentido.

También sería conveniente el acceso a una búsqueda avanzada, ya que sólo permite la búsqueda sencilla. Y quizá sería interesante permitir *feedback*, para poder compartir en redes sociales, aunque hay que mencionar que el proyecto cuenta con una cuenta en Twitter, https://twitter.com/proyectoadso.

Figura 4. Cuenta en Twitter del Proyecto ADSO

3. CORPUS

El corpus de textos con el que ha trabajado el proyecto proviene de la Biblioteca Virtual Miguel de Cervantes (BVMC), en concreto de la *Biblioteca del soneto*[3]. Se trata de una de las más grandes recopilaciones de textos digitalizados en HTML. Está disponible en un repositorio de GitHub: *Corpus of Spanish Golden-Age Sonnets (with metrical annotation) / Corpus de Sonetos del Siglo de Oro (con anotación métrica)*, https://github.com/bncolorado/CorpusSonetosSigloDeOro[4].

[3] Se puede consultar en la página principal de la *Biblioteca del Soneto*. Accesible desde: http://www.cervantesvirtual.com/bib/portal/bibliotecasoneto/index.html.

[4] El repositorio web fue reseñado anteriormente por Calvo Tello, J. (2017). Corpus of Spanish Golden-Age Sonnets. RIDE, 6: *Text Collections* IDE (September 2017). Accesible desde: https://ride.i-d-e.de/issues/issue-6/corpus-of-spanish-golden-age-sonnets/.

El corpus seleccionado pretende ser una compilación representativa de la sonetística en lengua castellana de los siglos áureos (XVI y XVII). Incluye un total de más de 5.000 sonetos pertenecientes a una cincuentena autores, entre los que figuran los canónicos de la poesía castellana del Renacimiento y del Barroco como Garcilaso de la Vega, Boscán, Cervantes, Quevedo, Lope de Vega, Herrera, Góngora, Calderón o sor Juana Inés de la Cruz, así como otros autores, algunos de ellos de menor renombre. Esto se debe a que, a la hora de conformar el corpus, no se ha tenido en cuenta la calidad literaria, sino que el criterio para su inclusión ha sido que al menos cada autor/a contase con un mínimo de 10 sonetos digitalizados.

Cada soneto ha sido anotado en XML siguiendo el estándar TEI (TEI Consortium, 2016)[5]. La información de cada poema incluye los metadatos, se marca la estructura del soneto por estrofas y versos y se usa una etiqueta para señalar la métrica de los versos (14, salvo los que llevan añadido un estrambote). Ésta es una de las principales novedades y aportaciones del desarrollo del proyecto, la de incorporar el patrón métrico de cada uno de los sonetos de los que consta el corpus.

La anotación métrica se ha llevado a cabo de manera semi-automática. Aparece representada a través de un patrón métrico en el que el símbolo "+" corresponde a las sílabas tónicas, y el "-", a las átonas.

```
<text>
    <body>
        <head>
            <title>- CVIII - </title>
        </head>
        <lg type="cuarteto">
            <l n="1" met="-+-+-+-+-+-">Amor por ese sol divino jura,</l>
            <l n="2" met="+-+---+---+-">siendo negro color vuestros despojos,</l>
            <l n="3" met="-+-+-+---+-">quizá por luto, más que por enojos,</l>
            <l n="4" met="-+---+---+-">de muchos que mató vuestra hermosura.</l>
        </lg>
```

Figura 5. Ejemplo: CorpusSonetosSigloDeOro/LopeDeVega_1/LopeDeVega_11.xml.

Cabe señalar que la disponibilidad del corpus anotado en GitHub facilita la copia de datos, así como su modificación en el caso de que fuese necesario. El repositorio se estructura en carpetas y archivos clasificados y distribuidos según el nombre del autor/a de las composiciones. Cada soneto es codificado como un solo archivo y son agrupados los del mismo poeta en una carpeta única, aunque en algún caso lo hace en dos como en el caso de Lope de Vega por la gran cantidad de sonetos de su autoría (más de 1300). Esto, sin duda, ayuda enormemente al estudio de los textos desde diferentes perspectivas: *topic modeling*,[6] análisis de los patrones comunes tanto en los sonetos de un mismo autor como de modo comparativo con el

[5] TEI CONSORTIUM (Ed.). (2016). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Version 3.1.0., TEI Consortium. http://www.tei-c.org/Guidelines/P5/.
[6] Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM, 55*(4), 77-84.

resto. En este sentido, precisamente se han llevado algunos estudios como los publicados por el investigador principal del proyecto, Navarro Colorado (2015 y 2016)[7].

En la BVMC se codifican los sonetos como un libro digitalizado en HTML, pero en el repositorio del proyecto cada poema se presenta por separado, esto es, de forma individual en un fichero TEI. Sin embargo, los metadatos son comunes para todos los sonetos. Puesto que los archivos presentan la misma estructura, esto permite un análisis comparativo de manera efectiva. Tratar los sonetos de forma individual refuerza el peso de cada texto y su singularidad dentro del corpus. Sin embargo, queda en segundo plano la importancia de la colección o la antología poética en sí misma.

4. SISTEMA Y PROCEDIMIENTO DE ESCANSIÓN

El sistema de anotación métrica se encuentra bajo licencia de Creative Commons Reconocimiento-NoComercial 4.0 Internacional. Para proceder a la anotación métrica de los sonetos, en primer lugar, éstos se procesan mediante un sistema automático de escansión[8]. Si bien, por un lado, esto permite unos resultados a gran escala de forma rápida, por otro, el procesamiento puede generar errores sistemáticos en una buena parte de los sonetos escandidos. Para minimizarlos, en segundo lugar, se pasa a una fase de validación manual de los resultados automáticos del patrón métrico. Lógicamente, esta fase es más lenta, y no se concibe ni es factible hacerlo de todo su conjunto, sino de un porcentaje mínimo del 10%. En la web principal del proyecto se indica que es una tarea que se lleva a cabo por diferentes anotadores. Una anotación manual que se basa en una *Guía de anotación* propuesta por los miembros del equipo (Navarro, Ribes y Sánchez), cuya última versión publicada en la web es del 2/10/2016. No obstante, esta guía no coincide con la que está disponible en el repositorio de GitHub, siendo esta última anterior a la de la web ADSO. La *Guía de anotación* pretende resolver dudas en el proceso de anotación métrica y se fundamenta, principalmente, en la aportación clásica de referencia como es *Métrica española* (1984) de Antonio Quilis[9].

El mismo equipo del proyecto fue consciente de los errores que se encontraban en la anotación manual de un porcentaje significativo del corpus, tal como ejemplificaron a través de la presentación de la comunicación "Cuestiones y problemas en la anotación métrica de un corpus de poesía castellana" (Herrero et al., 2017) en el *III Congreso Internacional de la*

[7] Navarro Colorado, B. (2015). A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. En *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver (EEUU), y Navarro Colorado, B. (2016). Hacia un análisis distante del endecasílabo áureo: patrones métricos, frecuencias y evolución histórica. *Rhythmica. Revista española de métrica comparada*, 14, 89-118.

[8] Navarro Colorado, B. (2018). A Metrical Scansion System for Fixed-Metre Spanish Poetry. *Digital Scholarship in the Humanities*, 33(1), 112-127. doi.org/10.1093/llc/fqx009.

[9] Quilis, A. (1984). *Métrica española*, Barcelona: Ariel.

Sociedad Humanidades Digitales Hispánicas[10]. Los errores se encuentran, primordialmente, en tres niveles: 1. Los que proceden directamente del texto, es decir, errores textuales como erratas de la propia edición digital; 2. Errores derivados del propio sistema de escansión (aparición de algunos versos no endecasílabos como en el estrambote; errónea clasificación categorial de palabras; errónea resolución de diéresis y sinalefas); 3. Versos con ambigüedad métrica cuyo criterio depende del anotador/a. Quizá este último nivel de errores encontrados resulte el más débil para el sistema de escansión, puesto que depende de la apreciación de un anotador/a, pero se intentan minimizar los casos teniendo como referencia y apoyo la *Guía de anotación métrica* propuesta.

5. AMPLIACIÓN DEL PROYECTO

A pesar de que la web no visualiza la evolución posterior del proyecto, éste se ha visto motivado por una serie de novedades en su desarrollo durante los dos años de realización. Tal y como se especifica en la comunicación de Herrero et al. (2017), otros de los objetivos perseguidos era su ampliación respecto a dos aspectos: el del corpus de poemas y el del sistema de escansión.

Por una parte, en cuanto a la ampliación del corpus, se pretendía la incorporación de nuevos metros, nuevas estructuras estróficas, es decir, dar la bienvenida a otros tipos de poemas como églogas, silvas, octavas reales, romances, liras, madrigales, redondillas, coplas, etc. Esto ha supuesto para el proyecto trabajar con un corpus mucho más extenso, con la incorporación de alrededor de 700 poemas a los sonetos con los que ya contaba. La mayoría de los poemas digitalizados, también en estándar TEI-XML, han sido recogidos de la BVMC y pertenecen a los mismos autores de los siglos XVI y XVII de los sonetos del corpus base.

Por otra, el sistema de escansión muestra una nueva visualización y la información que incluye de cada soneto escandido es más compleja. De esta manera, se incluye la *url* del texto (en el <TEI Header>), el tipo de poema, la separación silábica de cada uno de los versos, así como el lema y la categoría gramatical de las palabras de las que se compone (categorizada a través de las etiquetas EAGLES).

La implementación del nuevo sistema de escansión supone un avance cualitativo en cuanto al análisis de los patrones semánticos y métricos de la poesía áurea castellana, pero se vislumbran ciertos errores que pueden surgir en dos aspectos: a. los derivados del patrón métrico; b. los provenientes de la categorización morfológica. A su vez, estos dos tipos de errores se generan atendiendo a tres categorías: 1. El corpus digitalizado; 2. El sistema de

[10] Herroro, M. Á, Navarro Colorado, B., Ribes Lafoz, M., Prieto De Paula, Á L. y Chico Rico, F. (2017). Cuestiones y problemas en la anotación métrica de un corpus de poesía castellana. En *III Congreso Internacional de la Sociedad Humanidades Digitales Hispánicas*. Málaga, octubre 2017.

escansión; 3. Según el criterio del anotador/a. Desafortunadamente, el repositorio de GitHub del proyecto no dispone actualmente de ejemplos de este nuevo sistema de escansión.

6. CONCLUSIONES

No cabe duda de que el proyecto, uno de los pioneros en Humanidades Digitales en España, es un recurso muy válido para el estudio de la poesía áurea castellana. El proyecto se ha desarrollado formalmente durante dos años tras los cuales se ha demostrado que la aplicación de nuevos métodos de análisis distante basados en la lingüística computacional y *Text mining* son de enorme utilidad para extraer rasgos y patrones comunes en una época y en un género literario determinado. El uso de un gran corpus digitalizado de textos en TEI-XML y estar disponible en GitHub, así como su facilidad de descarga abierta y de citación, ayudan a su accesibilidad. Es incuestionable la gran aportación que supone la anotación métrica de cada verso, pero la información sobre el corpus en su totalidad y las novedades de los últimos meses de trabajo se echan en falta.

El desarrollo de un proyecto como ADSO en dos años resulta viable para la implementación de los objetivos marcados en su inicio. Sin embargo, evidencia que éste tiene un recorrido mucho más amplio y las investigaciones que se puedan derivar del mismo sobrepasarán este período de evolución.

Aunque muchos aspectos del desarrollo de estas técnicas se encuentran todavía en un estado embrionario y se necesitan mejorar algunos detalles en concreto del sistema de escansión, para los que el criterio del anotador/a no sea determinante en ciertos casos, queda demostrado que su aplicación en el campo de la Filología supone un salto cuantitativo y cualitativo en el análisis de los textos poéticos. Pues esto invita a dejar de lado el tradicional *close reading* en el análisis textual, y la aplicación del *Distant Reading* facilitará que éste sea mucho más preciso en estudios de temas, tópicos y patrones métricos.