Procesamiento del lenguaje natural y fijación del texto. Experiencias en torno a la constitución de un corpus diacrónico de sonetos

Helena Bermúdez Sabel, Clara I Martínez Cantón, Pablo Ruiz Fabo

Procesamiento del lenguaje natural y fijación del texto. Experiencias en torno a la constitución de un corpus diacrónico de sonetos

Helena Bermúdez Sabel
JinnTec
helena.bermudez@jinntec.de

Clara Isabel Martínez Cantón
UNED
cimartinez@flog.uned.es

Pablo Ruiz Fabo
Université de Strasbourg
ruizfabo@unistra.fr

Resumen
Esta contribución surge en el contexto de desarrollo del corpus de sonetos DISCO (*Diachronic Spanish Sonnet Corpus*), un corpus de 4530 sonetos en español compuestos entre el siglo XVI y el XX por autores de diversas procedencias (Europa, Latinoamérica y Filipinas). Este recurso contiene las anotaciones de diferentes fenómenos de versificación que han sido obtenidas a partir de técnicas del procesamiento del lenguaje natural (PLN). En este artículo presentamos cómo los resultados de la anotación automática pueden ser utilizados para detectar problemas de transmisión textual. Uno de los objetivos de esta contribución es el de proporcionar claves sobre posibles flujos de trabajo que, ayudándose de herramientas de PLN, permitan detectar posibles errores textuales, centrando así los esfuerzos de revisión manual en pasajes concretos.

Palabras clave
corpus; soneto; procesamiento del lenguaje natural; anotación; edición digital; versificación.

Abstract
Natural Language Processing and Text Curation. Experiences within the Development of a Diachronic Sonnet Corpus.

We present work carried out within the development of DISCO, the *Diachronic Spanish Sonnet Corpus project*, which consists of 4,530 sonnets in Spanish from Europe, Latin America and the Philippines, including texts from the 15th to the 20th centuries. The resource offers versification annotations obtained automatically through tools based on Natural Language Processing (NLP). In this article, we present how automatic annotation results can be exploited to detect textual transmission errors. Drawing on our experience with DISCO, we present observations towards the creation of workflows assisted by NLP-based tools, which can help detect possible textual errors, thus allowing us to focus on specific passages for our manual correction effort.

Keywords
Corpus; sonnet; natural language processing; annotation, scholarly editing, versification.

Introducción

El propósito de trabajar con corpus que reúnan una colección numerosa de textos literarios, que permitan la comparación de sus características genéricas, métricas, estilísticas, etc., se había visto condicionada, hasta la era digital, por la capacidad limitada del ser humano para recopilar, clasificar y analizar datos. En otras disciplinas, como la lingüística de corpus, este enfoque metodológico tuvo más presencia. Sin embargo, también en los estudios literarios encontramos previamente a la utilización de bases de datos o internet, trabajos de investigación que llevan a cabo cuantificaciones sobre determinadas características de un corpus. En el área de la métrica poética cabe nombrar, por ejemplo, repertorios que recogían los esquemas métrico-rimáticos de un corpus, y que fueron creados mediante un análisis manual, como *Die nicht lyrischen Strophenformen des Altfranzösischen* de *Gotthold Naetebus* (1891), repertorio métrico que reunía con un método riguroso la poesía narrativa francesa altomedieval, conservado en un comienzo únicamente en papel.

La edición de textos es un campo que se ha visto enriquecido por la utilización de métodos digitales de muy diversas maneras. En este artículo queremos exponer cómo la recopilación y anotación con herramientas digitales del

corpus DISCO (*Diachronic Spanish Sonnet Corpus*)[1] ha servido para mejorar la edición y curación de este corpus, pues ha permitido encontrar y corregir errores en la transmisión textual. Presentaremos, así, cómo ciertos indicios que percibimos gracias a la anotación automática pueden ser indicativos para la detección de pasajes problemáticos que pueden (o deben), después, ser cotejados manualmente.

DISCO es un corpus en continuo crecimiento,[2] que recoge en el momento presente 4530 sonetos en español de distintas épocas y lugares. Recoge poemas de 1215 autores y autoras procedentes de 22 países distintos de Europa, Latinoamérica y Filipinas. Cronológicamente se recopilan sonetos desde los inicios de esta forma poética (los primeros ensayos del Marqués de Santillana) hasta sonetos modernistas con versos alejandrinos o incluso hexadecasílabos.

Los sonetos recogidos en DISCO se extrajeron en la mayoría de los casos de la Biblioteca Virtual Miguel de Cervantes, de distintas colecciones de sonetos (García González 2006a; 2006b), u otros recursos, como el Portal de Literatura Hispanofilipina. Algunos textos del siglo xviii proceden de Wikisource u otras fuentes, y para la elaboración del subcorpus modernista el equipo de investigación ha realizado la digitalización de diferentes ediciones impresas. Estas diversas fuentes han sido homogeneizadas y estandarizadas gracias a su transformación en XML, siguiendo las directrices de la Text Encoding Initiative (TEI).[3] Otro valor añadido de DISCO, sin embargo, consiste en enriquecer estos textos, añadiéndoles, de manera sistemática, distinta información. Por una parte, se han añadido metadatos para los autores, como las fechas de nacimiento, muerte, procedencia y el identificador de VIAF (Virtual International Authority File). El Identificador VIAF[4] (Fichero de Autoridades Virtual Internacional) agrupa en un solo registro los diferentes registros de autoridad provenientes de diversas instituciones internacionales, en su mayoría bibliotecas nacionales. El identificador VIAF permite la identificación y referenciación consistente de los y las poetas del corpus, por lo que es relevante incluirlo.

Además, y lo que, creemos, hace más interesante al corpus, DISCO incluye anotaciones literarias, creadas con distintas herramientas digitales (ver, para más detalle Ruiz Fabo *et al.* 2021). El corpus está analizado métricamente, recogien-

1. https://github.com/pruizf/disco (último acceso el 26 de septiembre de 2023).
2. Ver https://github.com/pruizf/disco/#versions (último acceso el 26 de septiembre de 2023).
3. La Iniciativa de Codificación de Texto (TEI, por sus siglas en inglés) es un consorcio que desarrolla y mantiene de manera colectiva unas directrices para la representación de textos en formato digital. Este sistema de representación es el lenguaje de marcado, actualmente serializado como un vocabulario XML comúnmente denominado XML-TEI, que cuenta con una guía de uso mantenida por el consorcio. Para obtener más información sobre la Text Encoding Initiative (TEI), puede visitar su sitio web oficial: https://tei-c.org/.
4. https://viaf.org/

do el número de sílabas de cada verso, su acentuación, su rima e incluso la existencia o no de encabalgamiento.[5]

Este corpus, además, puede consultarse a partir de una interfaz que permite la búsqueda y exploración sencilla de los sonetos y la visualización con gráficos dinámicos de sus características métrico-literarias, DISCOver[6] (Bermúdez Sabel *et al.*, 2022).

Buscando la reutilización y la posible investigación con el corpus, DISCO se ofrece tanto en texto simple como en XML-TEI, este último enriquecido con RDFa, un formato de datos enlazados,[7] utilizando por tanto estándares con gran proyección en las humanidades.

Así, este corpus contribuye al desarrollo de recursos digitales para el estudio de la poesía en español, especialmente, desde una perspectiva comparativa, diacrónica y polisistémica.

La anotación automática va seguida de una labor de revisión y edición manual para la que ciertos aspectos de la anotación son clave, como indicamos en un comienzo. Es interesante recalcar que estos aspectos, además, no son los mismos para el Siglo de Oro que para el Modernismo, porque las tendencias literarias y la métrica también van evolucionando.

Presentaremos aquí el proceso y algunas muestras del trabajo de edición realizado con el subcorpus del Siglo de Oro, que recoge un total de 477 poetas que son autores de 1088 sonetos. Es interesante señalar la existencia de otro corpus distinto de sonetos de Siglo de Oro, el *Corpus de Sonetos del Siglo de Oro* (Navarro-Colorado *et al.*, 2016), que se solapa solo en una pequeña parte con DISCO, pues 17 de los autores se hallan representados en ambos corpus y encontramos 333 sonetos en común.

Retos

Los desafíos encontrados para la edición de este corpus de sonetos han sido de diversa naturaleza.

Hay algunos retos que pueden parecer comunes a la recopilación de un corpus, sea este digital o no, como todo el flujo de trabajo relacionado con la

5. Para estas anotaciones se han utilizado herramientas como Jumper (Marco Remón y Gonzalo 2021), el sistema de ADSO (Navarro-Colorado 2017) y Rantanplan (de la Rosa *et al.* 2020) para la escansión y la acentuación métricas, Rhyme Tagger (Plecháč 2018) para la rima y ANJA (Martínez Cantón *et al.* 2021; Ruiz *et al.* 2017) para el encabalgamiento. Las anotaciones métricas han sido corregidas manualmente en los sonetos modernistas, mientras que las de rima han corrido la misma suerte en el subcorpus del Siglo de Oro.

6. https://prf1.org/disco/ (último acceso el 26 de septiembre de 2023).

7. El corpus ha sido utilizado para diversos estudios de humanidades digitales, como se recoge en https://github.com/pruizf/disco/#resources-reusing-disco (último acceso el 26 de septiembre de 2023).

selección de los elementos, en este caso los sonetos, del corpus, y la información que sobre ellos se proporciona. Sin embargo, como es bien sabido, el medio condiciona el contenido, y ofrece posibilidades y limitaciones. En nuestro caso, condicionó la selección de un solo tipo de composición poética, el soneto. Esta elección ha sido en parte debida a la incontestable relevancia y tradición de esta forma composicional en la literatura española. No es únicamente una forma poética, sino que puede ser considerada un subgénero lírico (Romero Luque 2017; Spang 1993; Zavala 1995). No obstante, las razones que nos guiaron a recoger un corpus solo de sonetos tienen que ver también con su tratamiento digital. Se trata de una forma manejable desde el punto de vista computacional, ya que obedece a restricciones claras. La variabilidad se mantiene dentro de unos límites, lo que facilita la comparación de rasgos métricos entre poemas, o entre distintos subcorpus de poemas.

Además, cuando este corpus comienza a gestarse, las herramientas de escansión y asignación del acento métrico estaban más avanzadas para el endecasílabo, mientras que para otras formas métricas la precisión era menor o los resultados eran poco significativos. No en vano, usando métodos computacionales, el soneto había sido estudiado ya por distintos investigadores (Navarro-Colorado 2015; Navarro-Colorado, Lafoz, y Sánchez 2016; Navarro-Colorado 2017; Agirrezabal, Alegria, y Hulden 2017). Por estas razones, un nuevo corpus de sonetos enriquecía el panorama ya existente y abría un diálogo con trabajos anteriores en los estudios literarios tradicionales, en el desarrollo de corpus digitales y en el análisis computacional de la poesía.

La elección de las fuentes de las que se tomarán los sonetos supuso también un reto, ya que, por una parte, debíamos aprovechar el material ya digitalizado disponible en bibliotecas digitales, pero éramos conscientes de que debía construirse un corpus que siguiera unos determinados criterios de selección con el fin de ofrecer a la comunidad científica un recurso de gran utilidad. Así, y dado que existía ya el corpus del Siglo de Oro (siglos xv-xvii) de Navarro-Colorado (2015) centrado en autores canónicos, DISCO se ocupa de recopilar sonetos de todo tipo de autores, incluyendo autores menores, que completen (y complementen) así otro tipo de acercamientos.

Con la idea de recoger una muestra de cómo se ha cultivado el soneto en español desde el siglo xv hasta el Modernismo en el siglo xx, no solo desde los autores canonizados, sino desde una perspectiva amplia, se recogieron sonetos de todas las épocas, potenciando así la dimensión diacrónica. Además, las dos últimas versiones de DISCO,[8] de 2021 y 2023 se recogen poemas de autores filipinos y se le da más cabida a escritores latinoamericanos, tratando de ampliar también la procedencia de los poetas del corpus.

8. Consultar aquí: https://github.com/pruizf/disco/#versions (último acceso el 26 de septiembre de 2023).

Otro reto que se nos planteó fue el de incrementar la usabilidad del corpus y, de manera particular, hacer visible la existencia de poetas mujeres, y la procedencia de los autores. Para ello, a través de la interfaz de exploración del corpus DISCOver se facilitó la consulta y navegación de corpus a través de dos funciones de búsqueda principales: búsqueda por autor y búsqueda facetada basada en metadatos de autor. El buscador dispone de filtros para periodos, género (femenino o masculino) y origen del autor o autora (Latinoamérica, Asia o Europa).

Además, con la intención de facilitar la selección y reutilización de subcorpus derivados de DISCO, DISCOver permite seleccionar los sonetos o autores deseados y etiquetar el subcorpus creado. Este nombre se utilizará para identificar los resultados del subcorpus en los gráficos que la herramienta genera. Para los poemas seleccionados por el usuario, se crean gráficos dinámicos que muestran diferentes características métrico-rimáticas de la selección. Pueden combinarse en un mismo gráfico los resultados de diferentes selecciones con el fin de facilitar la comparación cuantitativa. Es posible además acceder además al texto de los propios sonetos en los que se ofrece la anotación de diferentes elementos de versificación (patrón acentual de cada verso, esquema de rima y encabalgamiento). Por otra parte, la creación de nuevas herramientas de anotación automática ha dado nuevas posibilidades a este corpus. Así, si antes nos habíamos ceñido al soneto hasta el siglo XIX, pues son casi sin excepción endecasilábicos, las versiones nuevas de DISCO incluyen sonetos modernistas, con sus distintas medidas, versos compuestos, etc. Su etiquetado automático ha sido posible gracias a la existencia de nuevas herramientas de escansión, como Jumper (Marco Remón y Gonzalo 2021) o Rantanplan (de la Rosa *et al.* 2020). Un tratamiento eficaz del verso compuesto ha sido posible gracias a Jumper, cuyos resultados con este tipo de versos (como evaluado en Marco Remón y Gonzalo 2021) superan los de otras herramientas.

La detección de la rima se añadió, asimismo, únicamente a partir de la versión 3 del corpus, publicada en octubre de 2019, con la adaptación de la herramienta Rhyme Tagger (Plecháč 2018) para el español.

La edición digital de textos cuenta ya con directrices específicas, por lo que DISCO, desde un primer momento ha seguido de cerca los criterios de RIDE para Colecciones de Textos Digitales (Henny-Krahmer y Neuber 2017).

La creación de un flujo de trabajo y unos criterios homogéneos para la edición y limpieza de los textos será, precisamente, el reto del que hablaremos en este artículo.

Metodología

Las anotaciones prosódico-rimáticas del corpus DISCO han sido realizadas a partir de diferentes herramientas de procesamiento del lenguaje natural. En el

caso particular de la rima, se ha utilizado el sistema multilingüe *Rhyme Tagger* (Plecháč 2018).

Uno de los mecanismos implementados en la elaboración del corpus DISCO para mejorar la calidad de este recurso ha sido el de revisar manualmente los versos no rimados. La rima se presenta como un fenómeno especialmente apropiado para focalizar los esfuerzos de revisión manual por diferentes motivos relacionados tanto con la historia de la versificación como con el procesamiento automático. Por una parte, dependiendo del período, la presencia de versos no rimados en el soneto en español puede considerarse una rareza (Pardo 2014). Un estudio sobre cómo se componía un soneto en el Siglo de Oro español señala precisamente la rima como piedra de toque sobre la que se construía el edificio poemático: "cuando uno se ponía a hacer un soneto, lo primero que hacía era apuntar en el margen las palabras rimantes. Esto no debe sorprender: es obvio que lo que rige todo soneto es la rima" (Dadson 1998: 512). Resulta llamativo, de hecho, que Kurt Spang (1993: 99-103), en el espacio que dedica al soneto como subgénero lírico, haga alusión a las posibles muchas variedades de la rima y que no se considere la posibilidad de los sonetos sin rima.

Por otra parte, el rendimiento de la anotación automática de la rima es muy elevado, con una precisión de 0.98 y una exhaustividad de 1. Esto quiere decir que la herramienta, en principio, detecta todos los casos de rima y que solo en 2 de cada 100 casos anota como rima algo que en realidad no lo es.

Teniendo en cuenta lo anterior, la presencia de un verso no rimado en el corpus de sonetos del Siglo de Oro puede relacionarse con tres posibles escenarios: 1) innovación formal, 2) problema de transmisión textual de la obra o 3) error generado durante el proceso de conformación del corpus. En otros períodos del corpus DISCO, en los que se ha generado una versión electrónica del texto a partir de fuentes impresas, tendríamos un caso específico del escenario 3 que es muy habitual: errores generados durante el proceso de digitalización (OCR). Sin embargo, la fuente del subcorpus del Siglo de Oro es nativamente digital (García González 2006). Es posible, no obstante, que nos encontremos con errores textuales en esta fuente que en su origen provengan de errores de OCR y que han pasado desapercibidos por el editor: en todo caso, en nuestro contexto específico tienen que ser categorizados como problemas de transmisión del texto, pues estamos utilizando una fuente digital curada.

Esta contribución explora cómo podemos utilizar los resultados de la anotación automática relativos a la rima para detectar posibles problemas en la transmisión textual de una obra. La búsqueda de estos potenciales problemas se centra en la presencia de versos no rimados que serán categorizados atendiendo a los tres posibles escenarios previamente descritos.

Algunos de los casos encontrados presentan una lectura que de manera inmediata puede ser corregida. Véase como ejemplo los siguientes versos de Gonzálvez de Andrade (en negrita, palabra final del verso no rimado):

Des que en lechos de zafir **reposan**
y que por sendas de cristal caminas,
derramando tus urnas cristalinas
en favor de las playas arenosas;
(Paulo Gonzálvez de Andrade, *Soneto*, vv. 1-4)

Desde un punto de vista semántico y formal, la última palabra del primer verso tiene que corregirse en "reposas". La presencia de "reposan" es muy probablemente resultado del proceso de conformación de la edición digital de la Biblioteca Virtual Miguel de Cervantes, ya que esta es la única edición que hemos encontrado con la lectura "reposan" (y en los recursos de ella derivados).

Como decíamos, la búsqueda automática de versos no rimados nos ha proporcionado ejemplos que responden a un problema de fijación de texto y no a simples errores de digitalización. Este es el caso de uno de los versos no rimados encontrados en la obra de Esteban Manuel de Villegas en la que uno de sus sonetos presenta la siguiente lectura:

Quién me dijera, Clori, que algún día
te pudiera olvidar tan fácilmente,
mientras soltero crin hizo en tu frente
con hilos de oro lazos de **rabia**.
(Esteban Manuel de Villegas, *Eróticas*, Soneto II, vv. 1-4)

Ediciones recientes de esta antología de Villegas (p. ej. Villegas 2010) reproducen también la lectura "rabia". Sin embargo, en las primeras ediciones de la obra de Villegas podemos encontrar una lectura mucho más adecuada, tanto desde el punto de vista semántico como formal: "taujía" como palabra rima (véase Villegas 1618; 1774; 1797). Taujía aparece hoy en día en el DRAE como palabra en desuso para "ataujía", que significa, a su vez, damasquinado, es decir, una técnica artesanal de ornamentación que consiste en incrustar hilos de oro o plata en superficies metálicas, como acero o hierro, creando intrincados diseños decorativos. Como vemos, este término, además de encajar métricamente mucho mejor, tiene sentido en su contexto y resulta mucho más lírico.

Se reproduce a continuación una transcripción semi-diplomática de estos versos en la edición príncipe (Villegas 1618, 80):

Quien me dixera, Clori, que algun dia
te pudiera olvidar tan facilmente,
mientras ſoltero crin hiço en tu frente
con hilos de oro laços de tauxia,
(Esteban Manuel de Villegas, *Eróticas*, Soneto II, vv. 1-4)

La transmisión de la obra de Villegas parece mostrar una cierta complejidad, pues encontramos en los tercetos de uno de sus sonetos dos versos no rimados que no podemos corregir de manera inmediata:

Cuando mostrabas de azucena, **fruta**,
la tez bruñida, o como sin cuidado
de mí solicitabas tus placeres,
déjame pues, que si te quise **hermosa**,
ya no es posible, puesto que has llegado
a tiempo, que a ti misma no te quieres.
(Esteban Manuel de Villegas, *Eróticas*, Soneto X, vv. 9-14)

Una vez más, en una de las ediciones contemporáneas a Villegas (Villegas 1618, 82) encontramos una lectura con rima:

Cuando moſtrabas de açuçena, i roſa
la tez bruñida; ó como ſin cuidado
de mi ſolicitabas tus placeres,
Dejame pues, que ſi te quiſe hermoſa,
ya no es poſible, pueſto que has llegado
a tiempo, que a ti miſma no te quieres.
(Esteban Manuel de Villegas, *Eróticas*, Soneto X, vv. 9-14)

En este caso, la hipótesis que manejamos para explicar cómo la lectura "azucena, y rosa" ha pasado a "azucena, fruta," pasa por considerar que se trata de un error generado durante la creación de un recurso electrónico usado como fuente, causado por un error de OCR, y que los editores posteriores no han detectado. Es posible que el caso anteriormente comentado, la sustitución de *taujía* por *rabia,* responda a la misma casuística.

Discusión

El trabajo de revisión de los versos no rimados, detectados automáticamente, ha sido una estrategia eficaz para mejorar la calidad de los textos contenidos en el corpus DISCO. Cabe destacar, no obstante, que la mayor parte de los versos no rimados detectados en el corpus respondía a errores de la herramienta de anotación, generalmente relacionados con rimas imperfectas (p. ej. *serenas* / *cantinelas*) o con grafías homófonas (p. ej. *donna* / *razona, divino* / *digno*). Por lo tanto, este proceso de revisión también nos ha llevado a implementar correcciones en la anotación rimática.

Si bien es cierto que los flujos de trabajo automatizados requieren de la intervención humana para mejorar los resultados de la automatización, en esta contribución argumentamos que las herramientas de procesamiento del lenguaje natural también pueden ser utilizadas para detectar algunos de los errores humanos. Esa misma sugerencia es hecha por los autores de dos herramientas de escansión automática para español (Rantanplan y Jumper) cuando señalan:

> Remarkably, using both systems [Rantanplan and Jumper] on the same corpora has proved to be particularly useful in our experimentation: when both systems make the same mistake, it is a reliable signal that the manual annotation might be erroneous. Therefore, used in conjunction they can be a useful tool for corpus clean-up (Marco *et al.* 2021)

El corpus DISCO nace ante la necesidad de crear una gran colección de poemas en español (concretamente de sonetos) para poder responder a nuevas preguntas de investigación utilizando las metodologías habilitadas por la computación. Nos atrevemos a afirmar que entre las tareas necesarias para poder investigar en disciplinas de Humanidades Digitales más ingratas destaca la creación de los corpus: es un trabajo arduo y cronofágico, y que muchos organismos de evaluación académica en Humanidades ni siquiera consideran como producción científica.

El desarrollo de grandes colecciones de recursos electrónicos está generalmente vinculado a procesos de automatización que hagan esta tarea abarcable en el menor tiempo posible. La presencia de errores en el texto o en la anotación es prácticamente inevitable, de ahí que sea necesario establecer procesos de control de calidad en cada una de las etapas del flujo de trabajo (especialmente en las automatizadas).

Si bien es evidente que la calidad de los resultados de investigación adquiridos dependerá de la calidad científica del recurso utilizado, también es cierto que las metodologías computacionales nos permiten establecer márgenes de error aceptables: la presencia de errores no es incompatible con la calidad científica ni con la consecución de resultados pertinentes (véase, p. ej., Franzini *et al.* 2018). A la hora de desarrollar un recurso electrónico, resulta fundamental que el equipo de investigación de Humanidades defina cuáles serán los tests que permitan evaluar la calidad de los resultados de cada uno de los procesos de automatización y establecer un margen de error aceptable. A la hora de explotar dicho recurso, se necesitará hacer una reevaluación del margen de error atendiendo a los objetivos de la investigación específica que se va a desarrollar.

La Biblioteca Virtual Miguel de Cervantes es un recurso de gran valor para los estudios literarios en español. Aunque en esta contribución hayamos reflejado algunos de los problemas textuales que presenta una de las ediciones disponibles en este portal, consideramos que esta edición (dada su extensión) se encuentra dentro de los parámetros del margen de error aceptable.

En la elaboración del corpus DISCO hemos establecido un sistema de versiones documentado que permite una mejora continua (pero organizada) de este recurso. Esta estrategia ha permitido que una versión inicial del corpus estuviese disponible para poder ser utilizada por la comunidad científica ya en 2017. Seis años más tarde, hemos publicado la versión 5.0 del corpus (Ruiz Fabo, Bermúdez Sabel, y Martínez Cantón 2023). Este sistema de versiones permite que las y los usuarios de este recurso puedan citar la versión concreta sobre la que se basa

su trabajo y que el estado del corpus en ese momento siempre esté disponible. Al mismo tiempo, también pueden comprobar qué cambios ha sufrido el corpus en versiones posteriores y evaluar las nuevas posibilidades de investigación que les ofrece cada una de las actualizaciones.

Conclusiones

La presente contribución es un ejemplo de cómo la evaluación de los resultados de anotaciones automáticas (usando una o varias herramientas para la comparación) se puede utilizar como estrategia para mejorar el contenido textual de grandes corpus. En este caso, la anotación rimática automática nos ha guiado en tareas de fijación de texto limitadas a la palabra rima. No obstante, las correcciones manuales de anotaciones automáticas pueden explotarse de distintas maneras, pues cabe la posibilidad de que sirvan para mejorar herramientas de anotación existentes, para mejorar datos de entrenamiento y también para la creación de modelos de corrección automática.

Esta contribución y el trabajo de revisión textual llevado a cabo dentro del corpus DISCO es igualmente útil para otros recursos derivados de la Biblioteca Virtual Miguel de Cervantes como ADSO (Navarro-Colorado, Ribes-Lafoz y Sánchez 2016).

Las ediciones automatizadas son imperfectas, pero valiosas. En el análisis donde se busca evaluar estadísticamente una tendencia (lectura distante, lingüística de corpus), la existencia de cierta proporción de errores no es un obstáculo insalvable, ya que las tendencias se pueden detectar dentro de ciertos márgenes de error.

En este sentido, una edición automatizada de un corpus amplio puede ayudar a detectar tanto las tendencias generales como los valores atípicos Así, la detección, gracias a la anotación automática, de pasajes en los que encontramos características no esperadas (número de versos, de sílabas, rimas inesperadas) puede ayudar también a focalizar nuestros esfuerzos críticos e interpretativos en las partes más innovadoras, y llegar a una mayor comprensión de la experimentación poética, con el objetivo de acometer estudios y ediciones críticas.

Como ampliación y generalización de este trabajo inicial, planeamos utilizar la anotación métrica para buscar irregularidades (o "rarezas") en la escansión que nos permitan detectar errores de transmisión textual, así como errores generados durante la creación del corpus (mucho más frecuentes). El objetivo es precisamente el de eliminar todas las irregularidades causadas por errores, para poder centrarse en el estudio de las innovaciones o particularidades métrico-rimáticas.

Bibliografía

Agirrezabal, Manex, Iñaki Alegria, y Mans Hulden, "A Comparison of Feature-Based and Neural Scansion of Poetry", *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, noviembre, 2017, pp. 18-23. <https://doi.org/10.26615/978-954-452-049-6_003>.

Bermúdez Sabel, Helena, Pablo Ruiz Fabo, et Clara Martínez Cantón, "DISCOvering Spanish Sonnets: A circular reading experience", en *Computational Stylistics in Poetry, Prose and Drama*, ed. Anne-Sophie Bories, Petr Plecháč, y Pablo Ruiz. De Gruyter. <https://doi.org/10.1515/9783110781502-004>

Dadson, Trevor J. 1998. "Cómo se hacía un soneto en el siglo de oro: el caso de 'amor, la red de amor digo que es hecha' ", en *Siglo de Oro. Actas del IV Congreso Internacional de la AISO*, Alcalá de Henares, 1996, pp. 2: 509-524.

Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk, y Jan Rybicki, "Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm", *Frontiers in Digital Humanities*, 5 (2018). <https://www.frontiersin.org/articles/10.3389/fdigh.2018.00004>.

García González, Ramón, ed., *Sonetos del siglo XIX*, Alicante, Biblioteca Virtual Miguel de Cervantes, 2006a, <https://www.cervantesvirtual.com/nd/ark:/59851/bmc4q861>.

García González, Ramón, ed., *Sonetos del siglo XV al XVII*, Alicante, Biblioteca Virtual Miguel de Cervantes, 2006b, <https://www.cervantesvirtual.com/nd/ark:/59851/bmc2r439>.

Henny-Krahmer, Ulrike, y Frederike Neuber, "Criteria for Reviewing Digital Text Collections, version 1.0", *IDE. A Review Journal for Digital Editions and Resources*, 6 (2017), <https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/>.

Marco Remón, Guillermo, y Julio Gonzalo, "Escansión automática de poesía española sin silabación", *Procesamiento del Lenguaje Natural* 66 (2021), pp. 77-87, <https://doi.org/10.26342/2021-66-6>.

Martínez Cantón, Clara Isabel, Pablo Ruiz Fabo, Elena Gonzalez-Blanco, Elena, y Thierry Poibeau. 2021. "Automatic Enjambment Detection as a New Source of Evidence in Spanish Versification", en *Plotting Poetry: On Mechanically Enhanced Reading*, ed. Anne-Sophie Bories, Gerald Purnelle, y Hugues Marchal, Liège, Presses Universitaires de Liège, pp. 93-112.

Naetebus, Gotthold, *Die nicht-lyrischen strophenformen des altfranzösischen*, Leipzig, Druck von J.B. Hirschfeld, 1891.

Navarro-Colorado, Borja, "A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects", en *Proceedings of the Computational Linguistics for Literature Workshop at NAACL 2015*, Denver, U.S., Association for Computational Linguistics, 2015.

Navarro-Colorado, Borja, "A metrical scansion system for fixed-metre Spa-

nish poetry", *Digital Scholarship in the Humanities* 33.1 (2017), pp. 112-27. <https://doi.org/10.1093/llc/fqx009>.

Navarro-Colorado, Borja, María Ribes Lafoz, y Noelia Sánchez, "Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation", en *Proceedings of the Tenth International Conference on Language Resources and Evaluation, Portoroz, Slovenia*, Portorož, Slovenia, 2016, pp. 4630-34. <http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf>.

Pardo, Arcadio, "De la diversidad del soneto". *Rhythmica. Revista Española de Métrica Comparada*, 12 (2014), pp. 127-172, <https://doi.org/10.5944/rhythmica.14247>.

Plecháč, Petr, "A Collocation-Driven Method of Discovering Rhymes (in Czech, English, and French Poetry)", en *Taming the Corpus: From Inflection and Lexis to Interpretation*, ed. Masako Fidler y Václav Cvrček, Cham, Springer International Publishing, 2018, pp. 79-95. <https://doi.org/10.1007/978-3-319-98017-1_5>.

Romero Luque, Manuel, "El soneto modernista (Manuel Machado como paradigma)», Rhythmica, 15 (2017), pp. 113-45. <https://doi.org/10.5944/rhythmica.21190>.

Rosa, Javier de la, Álvaro Pérez, Laura Hernández, Salvador Ros, y Elena González-Blanco, "Rantanplan, fast and accurate syllabification and scansion of spanish poetry", *Procesamiento del Lenguaje Natural*, 65 (2020), pp. 83-90.

Ruiz Fabo, Pablo, Helena Bermúdez Sabel, Clara Martínez Cantón, y Elena González-Blanco, "The Diachronic Spanish Sonnet Corpus: TEI and Linked Open Data Encoding, Data Distribution, and Metrical Findings", *Digital Scholarship in the Humanities*, 36 (Supplement_1) (2021), pp. i68-80, <https://doi.org/10.1093/llc/fqaa035>.

Ruiz Fabo, Pablo, Clara Martínez Cantón, Thierry Poibeau, y Elena González-Blanco, "Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets", en *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2017, pp. 27-32, <http://www.aclweb.org/anthology/W17-2204>.

Spang, Kurt, *Géneros literarios*, Madrid, Síntesis, 1993.

Villegas, Esteban Manuel de, *Las eroticas o amatorias*, Nájera, por Ivan de Mongaston, 1618.

Villegas, Esteban Manuel de, *Las Eroticas; y Traduccion de Boecio*, Madrid, por Don Antonio de Sancha, 1774.

Villegas, Esteban Manuel de, *Las Eróticas y traduccion de Boecio*, Segunda edición, Madrid, por Don Antonio Sancha, 1797.

Villegas, Esteban Manuel de, *Las eróticas o Amatorias*. ed. María Ángeles Díez Coronado y Emilio Magaña Orúe, Logroño, Instituto de Estudios Riojanos, 2010.

Zavala, Iris M, "El amor es una aventura en el mal: los sonetos de Sor Juana", *Tropelías: Revista de Teoría de la Literatura y Literatura Comparada*, 5-6 (diciembre) (1995), pp. 443-52, <https://doi.org/10.26754/ojs_tropelias/tropelias.19955-65581>