CORPUS DE NOVELAS DE LA EDAD DE PLATA, EN XML-TEI[1]

CORPUS OF NOVELS OF THE SPANISH SILVER AGE, IN XML-TEI

José CALVO TELLO
Universidad de Würzburg
jose.calvo@uni-wuerzburg.de

Resumen: En este artículo se presenta el *Corpus de novelas de la Edad de Plata*, una colección de 358 novelas publicadas por autores españoles entre 1880 y 1939. La selección de textos sigue criterios fijados por manuales de literatura. Los textos han sido codificados en XML-TEI, formato que también recoge los metadatos revisados de manera manual y las anotaciones lingüísticas realizadas por herramientas automáticas. El conjunto de datos permite realizar descripciones estadísticas, evaluar hipótesis propuestas por otros investigadores o explorar nuevas correlaciones. Finalmente, se darán descripciones semánticas de diferentes subgéneros de la novela.

Palabras clave: *Corpus de novelas de la Edad de Plata*. Géneros literarios. Corpus. Aprendizaje automático. Estadística.

Abstract: In this paper the *Corpus of Novels of the Spanish Silver Age* is presented, a collection of 358 novels published by Spanish authors between 1880 and 1939. The selection of the texts follows criteria from the studies of literature. The texts have been encoded in XML-TEI. In this format are also saved the manually curated metadata and the linguistic annotations by automatic tools. The data set allows calculating statistical descriptions of the novel, evaluate hypotheses by other researchers or explore new

[1] Esta publicación es parte de los resultados de investigación del Proyecto MNEMOSINE, "Hacia la Historia Digital de La Otra Edad de Plata: producción, almacenamiento, uso y difusión" (ref. RTI2018-095522-B-100), financiado por el Ministerio de Ciencia, Innovación y Universidades (MCIU), la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER).

correlations. Finally, it will be presented semantic descriptions about several subgenres of the novel.

Key Words: *Corpus of Novels of the Spanish Silver Age*. Literary genres. Corpus. Machine Learning. Statistics.

1. PRESENTACIÓN

El objetivo de este artículo es presentar el corpus de novelas de la Edad de Plata, titulado originalmente en inglés *Corpus of Novels of the Spanish Silver Age*, abreviado como CoNSSA. La creación de este corpus, así como su utilización para analizar principalmente los subgéneros de la novela (novela erótica, de aventuras...), se enmarca dentro del proyecto de investigación CLiGS[2], financiado por el Ministerio de Educación alemán. Este proyecto iniciado en 2015 por Christof Schöch (desde 2018 catedrático de Humanidades Digitales en la Universidad de Trier), está localizado en la cátedra de Fotis Jannidis sobre Filología Computacional y Humanidades Digitales de la Universidad de Würzburg. Los investigadores de este grupo analizamos mediante metodologías cuantitativas y computacionales, el género literario en diferentes lenguas (principalmente francés y español de España e Hispanoamérica), así como en diferentes períodos (Calvo Tello *et alii*, 2015, 2018).

El proyecto se enmarca dentro de los estudios literarios digitales, cuyo objetivo es aplicar a textos literarios metodologías computacionales originadas en la informática o métodos propios, para el análisis, evaluación, descripción, visualización o predicción de fenómenos literarios. El género literario es un objeto de estudio adecuado para metodologías cuyo objetivo es analizar cantidades de textos o cantidades de rasgos[3] demasiado

[2] El nombre completo es "Computergestützte literarische Gattungsstilistik", que se podría traducir como "Estilística computacional del género literario".

[3] Entiendo rasgo como cualquier propiedad que puede describir una instancia. Un rasgo del fonema /m/ es su nasalidad. Un rasgo de la forma *trayendo* es que es una forma verbal. Un rasgo de un texto puede ser que contiene una palabra tantas veces. Un rasgo de un autor es el hecho de ser hombre,

grandes como para basarse en una lectura filológica tradicional de todos los textos. De hecho, los trabajos de dos investigadores influyentes en esta rama de las Humanidades Digitales, Franco Moretti y su *distant reading* (2013), y Matthew Jockers y su *macro-analysis* (2013), dedican secciones al análisis de los subgéneros de la novela. Desde entonces, los estudios computacionales de género literario han aplicado metodologías provenientes del aprendizaje automático o *machine learning* (Müller y Guido, 2016). Aunque más adelante daré una explicación más completa, por ahora podemos entender que la aplicación de esta área de la informática a los géneros literarios busca combinaciones de rasgos que diferencian novelas de distintos géneros. Dentro de las Humanidades Digitales, el análisis de autoría mediante metodologías estilométricas, el aprendizaje automático y el análisis cuantitativo de géneros literarios han sido fruto de un intenso y fructífero intercambio de metodologías y conjuntos de datos (Underwood, 2014; Hettinger *et alii*, 2016).

Sin embargo, estas metodologías solo son capaces de ser evaluadas y testadas en aquellos conjuntos de datos formalizados que cada lengua tenga a su disposición. Por conjuntos de datos (*data sets*) me refiero a diferentes recursos entre los que se encuentran catálogos, corpus textuales, archivos visuales, diccionarios, etc. Además, los tipos de conclusiones y descripciones a los que lleguen dependen profundamente de los criterios con los que fueron creados. Por eso son importantes investigaciones como las que conforman este monográfico, en los que los investigadores que han trabajado en la creación de conjuntos de datos digitales puedan explicar sus principios y cómo estos afectan a su utilización para su consulta o análisis.

2. ESTADO DE LA CUESTIÓN

El aprendizaje automático ha generado un enorme interés tanto en círculos académicos, institucionales y empresariales. Lamentablemente, para el caso del español, este interés no se ha traducido en un impulso o apertura para que el investigador tenga acceso a grandes conjuntos de datos formalizados con los que trabajar. Instituciones financiadas con

o que haya nacido en 1880.

dinero público como la Real Academia, la Biblioteca Cervantes Virtual, o proyectos de financiación universitarios como TESO (Simón Palmer, 1997) o ArteLope (Oleza Simó, 2013) tienden a no dar acceso abierto a los datos. Cuando se comenzó a gestar CoNSSA, no había a disposición de los usuarios ningún corpus en español que diese acceso abierto a textos literarios de la época, ni que ofreciese información sistemática sobre el género literario al que pertenecen las obras.

A pesar de estas restricciones que las instituciones y la comunidad de investigación se han auto impuesto desde hace décadas, algunos investigadores han dado acceso en los últimos años a corpus textuales que pueden ser descargados en su conjunto[4]. Por ejemplo, los proyectos *Corpus of Spanish Golden-Age Sonnets*, liderado por Borja Navarro-Colorado (2015) y *Diachronic Spanish Sonnet Corpus*, liderado por Pablo Ruiz Fabo (2017), han publicado más de 4.000 sonetos en XML-TEI de manera abierta. También en este formato ha publicado la *Biblioteca Electrónica Textual del Teatro en Español 1868-1939* 25 obras de teatro de la Edad de Plata, proyecto dirigido por Teresa Santa María Fernández (2017). Otros investigadores han publicado conjuntos de textos en otros formatos, como se observa con los ejemplos del corpus de *Fábulas mitológicas del Siglo de Oro*, de Antonio Rojas Castro (2016), el *Corpus Lazarillo* (de Javier de la Rosa, 2016) o el *IMPACT-es Diachronic Corpus* de Felipe Sánchez-Martínez (2013).

Aun así, en muchos casos los investigadores necesitan ir a fuentes más genéricas que publican en web o similares, como los libros electrónicos en ePub (Garrish, 2011), y a partir de ellos pueden convertirlos en otros formatos. Algunas de las fuentes más utilizadas son Cervantes Virtual o el proyecto Gutenberg. Menos frecuente es el acceso al portal *ePubLibre*, que contiene miles de libros electrónicos mediante *torrents*. En este mismo formato ha comenzado recientemente la Biblioteca Nacional de España a publicar parte de su catálogo. Si el investigador no encuentra versiones web del texto, el siguiente paso es acceder a versiones digitales, normalmente en PDF, como por ejemplo las que se encuentran en plataformas como Archive.org o Biblioteca Nacional de España.

En realidad, la investigación cuantitativa en géneros literarios

[4] En esta sección solo mencionaré los proyectos por nombre. Al final del artículo se encuentran referencias completas a cada recurso, incluidos enlaces.

está obligada a mirar al ámbito internacional para encontrar modelos, especialmente hacia las zonas de habla inglesa y alemana. Para el inglés, hay corpus como el *Shakespeare Folger Library* (Mowat y Werstine, 2010), el *CENLAB* aglutinado por Jonathan Reeve con más de 400 novelas (2016), o los corpus *LitBank* de David Bamman (2019) y el *Novels Project Corpus* realizado por Allen Riddell (2014), ambos con 100 novelas. Además de esto, nos encontramos que diferentes proyectos trabajan con conjuntos de datos aportados por empresas, como el que utiliza Ted Underwood, con decenas de miles de textos provenientes de Hathi Trust (Underwood, 2019).

El segundo gran foco internacional de publicación de textos literarios es el territorio de habla germana. El proyecto que da acceso abierto al mayor conjunto de datos es *TextGrid Repository* (2016), donde se encuentran decenas de miles de textos en alemán, codificados en formato XML-TEI. Otro proyecto de divulgación de grandes cantidades de texto, en este caso con un interés explícito en la historia de la lengua, es el *Deutsches Textarchiv* (Grötschel, 2007), que cuenta con casi 4.000 textos en diferentes formatos. En cuanto al francés, uno de los mayores portales sobre textos es *Théâtre Classique* (2007), que ofrece más 1.200 obras de teatro.

Hasta ahora he señalado ejemplos de recursos en un solo idioma, pero también se encuentran diferentes proyectos que dan acceso a obras en diferentes lenguas, como el *Novel 450* de Andrew Piper (2016), con subcorpus de 150 novelas en alemán, inglés y francés. Actualmente se está desarrollando el proyecto *DraCor* (Trilcke y Fischer, 2018), con más 700 obras de teatro, principalmente en alemán y ruso, aunque también en notable menor medida en griego, latín, inglés y español. En la actualidad, uno de los proyectos más interesantes está siendo llevado a cabo por el proyecto europeo *Distant Reading*, cuyo principal objetivo es desarrollar una *European Literary Text Collection* (ELTeC), que contendrá al menos diez subcorpus en lenguas europeas, cada uno con 100 novelas, seleccionadas con criterios relativamente estrictos.

3. OBJETIVO: ¿PARA QUÉ SE CREÓ EL CORPUS?

Antes de que este corpus se crease, el objetivo de estudio ya estaba definido: analizar los diferentes subgéneros de la novela, como de aventuras, histórica, de humor, etcétera. Sin embargo, la selección de las obras no se hizo basándose en su pertenencia o no a ciertos géneros literarios, sino a obras analizadas (no solamente mencionadas) en manuales de literatura. Este criterio no permite acceder a un conjunto de textos que represente correctamente la literatura que se publicó durante esa época, ya que para ello se necesitarían los catálogos de novelas publicadas en la época, con miles de registros, la mayoría de ellos sin posibilidad de digitalización. Sin embargo ese criterio sí permite representar aquellas obras que los estudios de literatura tienden a tener en cuenta, algunos cientos de obras, muchas de ellas ya digitalizadas. Este último objetivo no era solamente más factible de conseguir, sino también más cercano a los objetos más frecuentes de estudio de los estudios literarios.

Una de las dificultades para definir ese conjunto fue el hecho de que no demasiados manuales cubren de manera homogénea, generalista y en profundidad el período entre 1880 y 1939 en su conjunto. Por ejemplo, el manual de Mainer (2009) analiza el período entre 1902 y 1939, por lo que habrían faltado información anterior. La utilización de obras más específicas (como monografías sobre ciertas décadas, autores o géneros literarios) habría causado sesgos en los datos, dificultando la comparación de los resultados. Esto daría lugar a artefactos con diferentes de objetivos y metodologías. Por ejemplo, podría haber elegido un catálogo diferente para cada una de las seis décadas que analizo, tratando así de utilizar manuales que abarquen más obras. Las diferentes premisas en las que cada manual se basase se transformarían en sesgos para cada década. Si el manual sobre la década de 1930 tuviese un enfoque feminista y el manual de 1880 prefiriese tesis católicas, los análisis posteriores señalarían cómo los autores y textos pasan de posiciones más católicas a posiciones feministas. Sin embargo si el manual católico fuese el de 1930 y el feminista el de 1880, los resultados serían diametralmente opuestos. Estos resultados serían meramente producto de la selección de fuentes para las obras, es decir, un artefacto causado exclusivamente por la falta de homogeneidad metodológica. Este tipo de efectos se evitan seleccionando fuentes de datos que hayan analizado el período completo de manera homogénea.

El principal manual utilizado fue el *Manual de la literatura española* (MdLE), de Pedraza Jiménez y Rodríguez Cáceres, en concreto los volúmenes 7 al 13 (1980). Este manual efectivamente cumple los criterios anteriormente mencionados: trata de manera uniforme y completa la Edad de Plata, y continúa siendo uno de los manuales generalistas más extensos, con referencias a cientos de obras en cada década. Además, los diferentes volúmenes tienen una estructura uniforme y explícita, pudiendo definir con exactitud la sección de los manuales que trata cierto grupo, período, autor u obra. Para que la selección no dependiese de un único volumen y que arrastrase los sesgos que cualquier investigación tiene, tomo también como referencia la colección *Historia de la literatura española* (HdLE) en sus volúmenes 5 y 6, dirigida por Mainer (2010). En comparación con el anterior, el HdLE tiene una estructura mucho menos homogénea y explícita, así como una menor extensión. En total, hay 730 novelas de la época que son mencionadas en al menos uno de esos manuales, y 360 que son mencionadas en ambos.

La selección de estos dos manuales de literatura como fuente es fácilmente criticable desde el punto de vista de estudios literarios. De igual manera, la informática puede criticar el tamaño final del corpus y señalar que un par de cientos de novelas no son conjuntos de datos realmente grandes para aplicar algoritmos de aprendizaje automático. Las dos áreas de investigación tienen objetivos diametralmente opuestos: las humanidades prefieren la calidad de la selección y la preparación, frente a la informática que favorece la cantidad. Esta tensión de criterios es característica de áreas interdisciplinares, como las Humanidades Digitales. Este dilema es especialmente notable en el caso de los trabajos de doctorado, si se compara con los trabajos de colaboración entre humanistas e informáticos. El doctorando, un único investigador, debe satisfacer las exigencias de varias comunidades científicas. CoNSSA no tienen en cuenta una gran cantidad de manuales de literatura, pero su diseño está definido siguiendo los estudios de literatura. De la misma manera, no es un corpus de miles de novelas, pero su tamaño es suficiente como para metodologías estadísticas o de aprendizaje automático como clasificación. Ni su tamaño ni su selección son óptimos para ninguna de las dos áreas, aunque no creo que hoy en día ningún corpus lo sea. Sin embargo sí permite llegar a conclusiones relevantes tanto sobre las obras literarias, así como sobre los algoritmos.

De esta manera, el MdLE y el HdLE fueron tomados como referencia para definir el conjunto de textos de la Edad de Plata que son típicamente tenidos en cuenta por los estudios de literatura. Esta puede ser entendida como una población completa en términos estadísticos (concepto que explicaré en el siguiente párrafo) de aquellas obras que los estudios de literatura tienden a analizar de esta época. La principal innovación de esta metodología es el hecho de llegar a una población estadística literaria concreta, lo que permite la correcta aplicación de metodologías estadísticas.

¿Qué es una población en términos estadísticos y por qué es relevante contar con una? En estadística, una población es un grupo de instancias que pueden definirse de alguna manera (Evans, 1996). La población de España es una población estadística. Los periódicos publicados en España en 2019 es otra. Las ovejas que viven actualmente en Reino Unido o las viñas de Francia son otros dos ejemplos. Con los pasos explicados en los anteriores, he llegado a definir de manera concreta una población estadística literaria: aquellas novelas publicadas entre 1880 y 1939 que son tratadas típicamente por los estudios de literatura (simplificado en los manuales HdLE y MdLE). Los estudios que aplican métodos cuantitativos a sus datos raramente disponen de datos de toda la población estadística: no existe ningún registro de todos los pacientes de asma, o ninguna base de datos contiene el peso de todos los españoles. Para analizar fenómenos como el asma o el sobrepreso, los investigadores crean muestras aleatorias con la máxima cantidad de encuestados o pacientes que se puedan obtener. De estas se pueden esperar que cubran la variación que se encuentra en la población completa. Así, se puede estimar valores sobre la población completa, aunque solo se tenga una muestra aleatoria.

Lo que es hábito en medicina o sociología, no debe serlo obligatoriamente para ciertas áreas de humanidades. Un estudioso puede decidir trabajar con la *Biblia* y trabajar con todos los versículos y capítulos que conforman el libro. Es decir, trabaja con la población estadística completa de la *Biblia*. De igual manera, un especialista en Unamuno puede trabajar con todas las novelas publicadas en vida del autor, que no excederá un puñado de obras. Muchas áreas de las humanidades sí están acostumbradas a trabajar con poblaciones estadísticas completas pero reducidas (entre un par y varios miles de instancias). Analizar de manera descriptiva todas las novelas de Unamuno lleva a conclusiones relevantes sobre ese conjunto de obras de ese autor. Sin embargo estas no pueden ser

extrapoladas a las novelas de otros autores, o a los textos de otros géneros de Unamuno. Es decir, aunque la relevancia de las conclusiones es alta, están restringidas a los criterios de la selección de datos.

Siguiendo la misma argumentación, el objetivo del corpus no solo es representar una colección de novelas, sino representar una población estadística concreta. Una manera de realizarlo es disponer de todas las novelas mencionadas en HdLE y MdLE, con un total de 360 textos. Sin embargo esto hubiese significado digitalizar varios cientos de novelas, objetivo no abarcable en este proyecto por falta de medios económicos. La estrategia que seguí fue crear dos corpus, uno anidado dentro del otro:

1. El CoNSSA, el mayor de los dos, contiene novelas que son mencionadas en el MdLE, conteniendo el número máximo de novelas posibles. En concreto conseguí 358 textos, que corresponden al 49% de todas las novelas mencionadas en MdLE.
2. El CoNSSA-canon, el menor, que contiene 138 novelas mencionadas tanto en el HdLE como en el MdLE, y son tratadas en cierta profundidad (dedicación de al menos una página para su descripción) en el segundo.

De esta manera, se forma un corpus que representa el 49% de la mayor población de novelas sobre la que tengo datos. Dentro de él, se encuentra un subconjunto de todas las novelas más canónicas, formalizando el canon de la manera arriba explicada.

La segmentación anidada del corpus permite realizar análisis concretos sobre ambos, y comparar los resultados. Por ejemplo, como señalaré más adelante, Ortega y Gasset realizó la hipótesis de que las buenas novelas son también las más largas. Puesto en términos estadísticos, la calidad y la extensión de las novelas tendrían una correlación. Esto puede ser formalizado y evaluado en los dos corpus, y observar si podemos observar la hipótesis en alguno de los dos corpus. De esta manera no solo podamos verificar si la hipótesis se da en la población de obras más importantes, sino también observar si el efecto es similar en un conjunto mayor de novelas. Si en ambos corpus observamos el mismo fenómeno, es factible que lo encontrásemos también en conjuntos de textos aun mayores, de miles de novelas que por ahora no han sido digitalizadas.

La hipótesis de Ortega es un buen ejemplo de como algunos datos (la extensión de una novela), son muy sencillos de contabilizar (las páginas de un libro, las palabras de un documento digital). Otros, sin embargo, son fenómenos abstractos y discutibles que normalmente no se encuentran accesibles de manera sencilla, como la calidad de una obra. Pero el principal objetivo del corpus es analizar géneros literarios, los cuales suelen estar descritos no en términos lingüísticos sino mediante conceptos literarios relacionados con el protagonista, el lugar y tiempo de la acción, el narrador, o relaciones con agentes externos al texto, tal como la calidad percibida por críticos. Por ejemplo, una novela histórica suele definirse como aquella que tiene lugar en una época pasada y cuyo objetivo es reflejarla de manera fehaciente. Nótese que esa definición no utiliza rasgos lingüísticos o textuales, tales como la extensión, o la proporción de verbos o vocabulario concreto que contiene el texto. Estos últimos tipos de datos lingüísticos pueden ser obtenidos por herramientas de procesamiento del lenguaje natural, pero ninguna herramienta automática señala si la representación del mundo es claramente fehaciente, hasta cierto punto realista o fantástica (al menos no por ahora). Por eso, el corpus debería contener también fenómenos literarios anotados de manera manual, codificados digitalmente como metadatos.

4. METODOLOGÍA DE RECOPILACIÓN DE TEXTOS

Para componer el CoNSSA, en primer lugar se descargaron los archivos en diferentes formatos, como web (HTML), libro electrónico (ePub) o PDF. En el caso de los textos que escaneé, los archivos fueron guardados en formato de imagen (JPG). Tanto estas como la mayoría de PDFs descargados no disponen del texto accesible constitucionalmente; en ambos casos, se utilizaron programas de reconocimiento óptico de caracteres (OCR), en concreto FineReader de Abbyy, que es capaz de producir versiones en HTML. De esta manera cualquier otro formato original (ePub, PDF o papel) pasaba en primer lugar por el muy extendido formato de la web: HTML. Este lenguaje de marcado permite diferenciar de manera unívoca aspectos como la cursiva, los párrafos o las imágenes. Sin embargo tiene claras limitaciones para la investigación en humanidades, como la falta de señalar unívocamente versos, acotaciones y parlamentos

(estos dos últimos frecuentes en novelas dialogadas), por mencionar solo un par de fenómenos textuales, además de su limitada capacidad de integrar metadatos.

Aunque similar y emparentado con el HTML, el formato XML-TEI ofrece numerosas ventajas (Agenjo, 2015; Allés Torrent, 2015; Mueller *et alii*, 2015). En primer lugar, permite marcar de manera explícita, unívoca y compartida por la comunidad la mayoría de fenómenos textuales. Facilita la personalización (*customization*), ya sea mediante la selección de elementos y atributos existentes, o mediante la integración de nuevos aspectos que el proyecto requiera. El uso de XML-TEI, además, propicia que terceros proyectos puedan reutilizar los textos de manera más sencilla a si se utilizan otros formatos menos conocidos o con menor documentación. Por otra parte, esta tecnología prevé la aplicación de maneras de validación de los archivos, con lo que se puede confirmar de manera automática que todos los archivos siguen una serie de criterios (como si sigue las reglas de sintaxis básica de XML, si todos elementos y atributos siguen las normas de TEI, si muchos de los valores están en listas predefinidas). En concreto, se aplicaron tantos esquemas que controlan elementos y atributos (esquema *RelaxNG*), así como esquemas que controlan mediante taxonomías de metadatos qué valores pueden contener ciertos elementos (esquema *schematron*). Es decir, se utilizaron herramientas computacionales que realizan un control de calidad automático sobre el corpus.

Todos los metadatos de cada novela fueron integrados en este archivo XML-TEI. Entre ellos se encontraba información sobre la obra (título, identificadores de BNE y VIAF), el autor, las ediciones sobre las que se basa (primera edición, edición digital, edición en la que basa la digitalización), así como diferentes metadatos administrativos (fecha de creación y cambios, situación legal del texto, etcétera). Posteriormente, se añadió un resumen lo más extenso posible, ya sea de propia elaboración (tras leer la novela) o de fuentes externas, en cuyo caso se integraba una referencia. Disponer de resúmenes de todos los textos permite trabajar con corpus que no se pueden leer en su integridad, como un corpus de 358 novelas, pero pudiendo abrir una ventana al contenido básico.

El siguiente paso era la anotación manual de diferentes metadatos. Se anotaron también ciertos metadatos del autor, como su sexo o las fechas y lugar de nacimiento y muerte. En cuanto al texto, se anotaron decenas de campos diferentes referentes a:

- El género literario, proveniente de diferentes fuentes, como manuales de literatura, plataformas en Internet, anotación propia, etcétera.
- El protagonista: su sexo, su edad aproximada al comenzar la novela, su nivel socio-cultural, su profesión, etcétera.
- El lugar de la acción: continente, país, región, lugar, tamaño aproximado de la población y si el lugar existe en la realidad.
- El tiempo de la acción: época y duración de la acción.
- Otros: narrador, tipo de final, si la obra es autobiográfica o no, etcétera.

Además de estos, se añadió la cantidad de páginas que el MdLE utiliza para analizar la obra. El objeto de este dato cuantitativo sencillo es una aproximación a lo que los estudios de literatura consideran importante o de buena calidad. Aquellas obras cuyo análisis ocupa más espacio en MdLE (*La Regenta, Niebla, Fortunata y Jacinta, Tirano Banderas*) ocupan posiciones más altas del canon que los que menos párrafos ocupan (*La vida rota* de Barga*, Pachín González* de Pereda*, Los gusanos* de Lanza).

Además de los textos y los metadatos revisados manualmente, el corpus fue anotado con herramientas lingüísticas de manera automática. En primer lugar se utilizó Freeling (Padró y Stanislovsky, 2012), una herramienta de procesamiento de lenguaje automático que, entre otras funciones, anota información sintáctica y morfológica, y clasifica las entidades nominales en personas, lugares, instituciones u otros. Esta herramienta da acceso también a los identificadores de WordNet (Fellbaum, 1998), un proyecto lexicológico desarrollado desde hace décadas. Estos identificadores fueron utilizados para conseguir las categorías semánticas básicas (lex_names) de cada palabra. Además de esto, el corpus fue anotado automáticamente con los catálogos léxicos que se encuentran en el *Diccionario de uso del español* (DUE) de María Moliner (1966), del que se utilizó su versión electrónica. Toda la anotación lingüística fue cargada también en otro documento XML-TEI, con lo que el investigador dispone de todos los datos en un único documento: texto, metadatos y anotación lingüística.

El CoNSSA, como el resto de corpus desarrollados en el proyecto CLiGS, está organizados mediante la aplicación GitHub. Este servicio está

principalmente pensado para desarrollar y publicar código y programas informáticos. Sin embargo este y otros servicios similares como GitLab están siendo cada vez más utilizados por diferentes proyectos de Humanidades Digitales para controlar y publicar conjuntos de datos, entre ellos corpus textuales. Estos sistemas ofrecen numerosos servicios y características a sus usuarios, como son la posibilidad de controlar exactamente la divergencia de ediciones, la reconstrucción de todos los puntos de desarrollo de los archivos, la descarga completa, la actualización automática para todos los usuarios, etcétera. GitHub además está integrado con otros sistemas, por lo que tener datos allí facilita utilizar terceras herramientas. Una de las más interesantes es Zenodo, un proyecto de archivo de datos de investigación a nivel europeo. De esta manera, cualquier proyecto puede archivar ciertas versiones de sus datos, con lo que se recibe un Identificador de Objetos Digitales (DOI) que lleva a un enlace desde el que cualquier persona puede y podrá en el futuro continuar descargándose los datos.

El proyecto publicará en su fase final todos los datos de la manera descrita, con la excepción de aquellos textos cuyos autores murieron hace menos de 80 años. Sin embargo no queríamos esperar hasta el final del proyecto para publicar parte de los datos, con el fin de favorecer su utilización, la colaboración con otros proyectos y la visibilidad del proyecto. Por eso publicamos nueve corpus de textos literarios en francés, italiano y español. Las características descritas hasta aquí sobre el CoNSSA pueden observarse en un conjunto de 39 novelas dentro del subcorpus específico de *Textbox* (Schöch *et al*., 2015).

5. ANÁLISIS Y RESULTADOS

En su conjunto, el CoNSSA contiene 358 obras, que suman más de 22 millones de *tokens* (es decir, palabras entendidas como cadenas textuales, numéricas o tipográficas, separadas por espacio o signos tipográficos). El CoNSSA-canon contiene 138 novelas, con casi 10 millones de *tokens*. ¿Qué representan estos datos en comparación con otros corpus similares o las secciones literarias de otros corpus más genéricos? Ambas versiones del corpus son notablemente mayores a colecciones anteriormente mencionados, como DISCO, ADSO, Corpus Lazarillo o Impact. El corpus TESO tendría alrededor de 10 millones de palabras, con lo que quedaría

entre la versión estándar del CoNSSA y su sección canónica. Los corpus de referencia de la Real Academia, CORDE y CREA, contienen secciones de obras literarias, cada uno con más de 27 millones de palabras. De esta manera, la versión completa de CoNSSA quedaría como uno de mayores corpus literarios de literatura española. Además, sería el mayor marcado en XML-TEI.

Las dos versiones del corpus pueden ser utilizadas para hacer una radiografía de las características principales de las obras de este período. Para ello, se puede utilizar estadística descriptiva, cuyo objetivo es resumir información de cantidades grandes de datos. El más conocido de ellos es la media, que se puede utilizar por ejemplo para describir la típica extensión de las novelas de esta época: las novelas del corpus tienden a tener 63.096 palabras. Uno de los aspectos negativos de la media es que es demasiado sensible a casos excepcionales (ya sea porque sean excepcionalmente altos o bajos). De esta manera si un par de novelas son excepcionalmente largas o breves, podrían estar distorsionando la representación de la media. Para una descripción estadística más robusta, puede utilizarse la mediana, que se obtiene al ordenar todas las instancias por su valor, y coger el caso medio. En el corpus, la extensión mediana del corpus es de 59.629 (similar a obras como *La ciudad de la niebla*, de Baroja, o *La voluntad*, de Azorín). El hecho de que la mediana sea menor que la media confirma que algunas obras son excepcionalmente extensas (como *La araña negra* o *Los Argonautas* de Blasco Ibáñez, *La Regenta* de Clarín o *Ángel Guerra* de Galdós).

La media o la mediana dan un dato único sobre la extensión típica de las novelas, pero esconde el hecho de que el corpus es variable: algunas novelas son mucho más extensas o mucho más breves que esas 60.000 palabras. La estadística pone a disposición una serie de herramientas como la desviación estándar para expresar esa variabilidad. En concreto, la desviación estándar de la extensión de las novelas es de 37.579. Sumando la desviación estándar a la media obtendríamos el límite típico que las novelas suelen tener, en este caso 100.615 palabras, novelas como *La bodega* de Blasco Ibáñez, o *Madrid de corte a checa*, de Foxá. Por su lado, restando la desviación estándar a la media, obtendremos el límite típico de novelas breves de esta época, que estaría en 25.457 palabras, con ejemplos como *El resplandor de la hoguera* de Valle-Inclán o *Niño y grande* de Miró con extensiones similares. De esta manera se ha obtenido

la radiografía buscada, una descripción estadística de la extensión típica de las novelas de esta época: suelen tener más de veinticinco mil palabras, y no más de cien mil. Estos datos concretos ahora pueden ser comparados con otros géneros literarios, otras lenguas u otras épocas. Sin embargo, ¿qué podemos aprender exactamente de la extensión de las novelas? ¿Es esto realmente interesante?

Antes de que responder a esa pregunta, hay que observar si se obtienen resultados similares mirando la población estadística completa que representa CoNSSA-canon. Si se observa la extensión de las novelas más canónicas, ¿se obtienen datos similares o la radiografía cambia notablemente? La media de las 138 novelas de CoNSSA-canon tienen una extensión media superior, de 70.458 *tokens*, con una desviación estándar de 36.036. Esto señala que las obras canónicas son más extensas que las que ocupan escaños más bajos en el canon. La versión completa del corpus está sesgada a favor de las novelas más canónicas, ya que estas son las que se encuentran digitalizadas. Esto quiere decir que de tener a disposición la mayor población completa de novelas de esta época, es previsible que su extensión sea más breve de lo descrito hasta aquí. De esta manera se han obtenido tres datos: 1) la extensión de las novelas canónicas; 2) la extensión de las novelas del corpus; 3) que una población mayor tendería hacia una extensión menor, aunque no se pueda deducir con exactitud cuán menor sería.

Regresemos a la pregunta sobre si se puede trabajar con variables y preguntas más pertinentes que la extensión de las novelas. Una de las maneras de obtener resultados estadísticos más interesantes es trabajar con variables más informativas. Por ejemplo, se puede utilizar la información sobre metadatos literarios anotada de manera manual. El grado de felicidad del final de la novela está codificado para cada texto en términos ordinales, con cinco posibles valores: final claramente triste, parcialmente triste, neutro[5], parcialmente positivo, claramente positivo. Esos valores pueden ser recategorizados numéricamente, por ejemplo señalando que finales claramente tristes representan el valor 0, mientras que las novelas con finales claramente positivos tendrían el valor 5. Una vez los valores están expresados de manera numérica, podemos calcular el final feliz típico de la novela de la Edad de Plata utilizando para ello la mediana: 0, es decir,

[5] Es decir, no se puede decir claramente si el final representa o no los objetivos del protagonista.

las novelas de esta época tienden a tener un final claramente triste. De la misma manera que con la extensión, se puede observar qué variabilidad tienen la felicidad de los finales: su rango intercuartílico[6] es de 2, es decir, los finales son proclives a ser entre claramente tristes, parcialmente tristes y neutros. Los mismos resultados se obtienen tanto de ambas versiones del corpus, lo que reafirma el análisis. El saber que la novela de la Edad de Plata tiende a terminar de manera triste es una radiografía obtenida de la misma manera que la anterior, pero al ser una variable más interesante, sus implicaciones también lo son. De nuevo, este dato puede ser ahora comparado con otras lenguas, otros géneros u otras épocas, pudiendo constatar tendencias en el desarrollo pesimista u optimista de la literatura, o en las diferentes tendencias en diferentes espacios geográficos o géneros.

Sin embargo, otra manera de obtener resultados más relevantes es poner en relación diferentes variables. El objetivo en este tipo de preguntas es observar si dos variables diferentes tienen alguna relación, o más propiamente dicho, si tienen correlación. Se ha señalado más arriba que las novelas de CoNSSA-canon tienden a tener mayor extensión que las de CoNSSA. Esto parece señalar que las obras canónicas son más extensas que las no canónicas. Ortega y Gasset menciona en *La deshumanización del arte* que "todas las grandes novelas que hoy preferimos son, desde otro punto de vista, libros un poco pesados" (2009: 177). Es decir, lanza la hipótesis de que hay una correlación entre el grado de canonización y la extensión del texto, sin especificar exactamente el grado de la relación ni determinar el período en el que esto ocurre. Para poder evaluar esta hipótesis estadísticamente, ambas variables deben ser formalizadas y expresadas de manera numérica. Es obvio que frente a la simplicidad de formalizar la extensión de un libro, la formalización de la canonización o calidad de una obra es profundamente problemática. Como he señalado más arriba, creo que la cantidad de páginas que manuales de literatura utilizan para describir una obra puede ser una posible manera de cuantificar este grado de canonización, aunque sea de manera demasiado sencilla o discutible. Si se acepta que puede ser una posible manera, se puede tratar de observar si hay una correlación entre ambas variables.

[6] El rango intercuartílico expresa la variabilidad, como la desviación estándar. La diferencia entre ambas es que la desviación estándar utiliza el valor de las instancias (como la media) y el rango intercuartílico utiliza la posición de las instancias (la mediana). Estos últimos son las maneras correctas de describir variables ordinales como la felicidad de las novelas.

Todos estamos acostumbrados a manejar correlaciones en el día a día: cuanto más calórica sea nuestra dieta, más engordaremos, lo que señala una correlación positiva entre calorías y peso. Cuanto más ejercicio aeróbico hagamos, más adelgazaremos, lo que presenta una correlación negativa entre deporte y peso. Sin embargo hasta ahora no estamos acostumbrados a ver fenómenos literarios como variables en correlación. Para evaluar si hay una correlación, qué tipo y cómo de fuerte es, podemos utilizar medidas estadísticas como el coeficiente de correlación de Pearson, normalmente referido mediante la letra *r* (Evans, 1996: 130-131), que expresa el grado de la correlación con valores entre -1 y 1, representando 0 que ambas variables no están asociadas.

El resultado al aplicarlo a la cantidad de *tokens* por novela, y la cantidad de páginas dedicadas en MdLE es de un valor *r* de 0.21 (siendo esta estadísticamente significativa, valor $p < 0.001$). Esto quiere decir que, como decía Ortega, hay una correlación positiva entre ambas variables. El test permite cuantificar este efecto, y señala que no es demasiado fuerte. Para la interpretación de este valor de 0.21 podemos recurrir a Evans (1996: 146), que señala que este valor representa una correlación débil. En otras palabras, se observa cierta correlación, pero lógicamente hay otros factores entre ambas variables: algunas novelas altamente canonizadas son breves, algunas novelas poco canonizadas son muy extensas. De esta manera he evaluado la hipótesis de Ortega sobre que la extensión de las novelas tendría una correlación con la calidad. Las formalizaciones de los datos aquí presentadas señalan en esa misma dirección, aunque la matizan: la correlación es débil.

La confirmación de lo que Ortega dijo hace muchas décadas, sin tener datos cuantitativos ni medios informáticos, puede ser de nuevo uno de los argumentos más frecuentes contra las Humanidades Digitales: ¿estamos aprendiendo algo nuevo? ¿O solo usamos complejos medios informáticos para confirmar las intuiciones que ya teníamos? Veamos otras dos hipótesis, en concreto sobre el estilo de la época. Estas fueron lanzada por Germán Gullón en su ensayo "Limites de la novela moderna" (1994), en los que analiza el desarrollo de la novela en general del paso del siglo XIX al XX. En concreto me quiero fijar en el siguiente párrafo:

> *La expresión del escritor moderno se caracteriza precisamente por la presencia de puntos suspensivos, interrogaciones, diálogos internos, la*

> *paradoja, la ironía, la frase pasional. Cuando la sinceridad sustituye a la verdad, el discurso adquiere un carácter sincopado. El hilván que atravesaba la frase, el párrafo, se frunce, las palabras se amontonan, se acumulan configurando unos párrafos irregulares, muy cortos o excesivamente largos* (Gullón, 1994: 200).

De esta cita pueden formalizarse dos hipótesis sobre la evolución del estilo del paso del siglo XIX al XX: 1) La cantidad de caracteres tipográficos en las novelas aumenta con el tiempo. 2) La variación de extensión en los párrafos y frases de las novelas (frases muy largas y frases muy cortas) aumenta con el tiempo.

Para evaluar la primera hipótesis utilizo el año de publicación para formalizar el paso del tiempo. La otra variable se obtuvo sumando la frecuencia relativa de todos los *tokens* que representan caracteres tipográficos. El test de ambas variables muestra efectivamente una correlación, aunque no la esperada: el coeficiente *r* es de -0.26 (valor $p < 0.001$). El signo negativo señala que según pasa el tiempo, la frecuencia relativa de la tipografía no aumenta, sino que en realidad disminuye.

La segunda hipótesis señalaba que la extensión de las frases o párrafos tienden a aumentar con el tiempo: frente a los decimonónicos "párrafos amplios, construidos con solidez y equilibrios" (Gullón, 1994: 200), con la modernidad se encontrarían tanto frases muy breves como muy largas. Esta variación es precisamente lo que mide la desviación estándar. Así, se calcula la extensión de todas las oraciones de cada novela, calculando posteriormente su desviación estándar. Esta variable es puesta en correlación de nuevo con el año de publicación. El resultado es una correlación estadística ($r = -0.36$, valor $p < 0.001$), pero, de nuevo, no en la dirección esperada: la desviación estándar de las oraciones no aumenta con el paso del tiempo, sino que los datos señalan que disminuye.

Estas dos hipótesis son buenos ejemplos de que las Humanidades Digitales no están limitadas a confirmar lo que ya sabíamos, sino que pueden señalar, como en el caso de Gullón, que lo que pensábamos no era correcto, e incluso que los datos señalan justo lo contrario. El hecho de que hayamos visto algunas hipótesis refutadas, subraya la importancia de la confirmación de la hipótesis de Ortega. Evaluar una hipótesis positivamente no es "volver a decir lo que ya sabíamos", sino que niega que los datos digan algo diferente.

Además, las herramientas computacionales permiten buscar correlaciones que hasta ahora no habíamos sospechado. Uno de los aspectos más claros es el hecho de que las novelas se hacen estadísticamente más breves con el tiempo: las novelas tienden a encoger con el tiempo, en concreto tienden a hacerse 562 *tokens* más breves por cada año que pasa (valor $p < 0.001$). Es decir, entre 1880 y 1939 la extensión de la novela mengua, tendencia observable tanto en la versión CoNSSA como CoNSSA-canon. Este hecho era desconocido hasta ahora.

Uno de los objetivos actuales de mi trabajo no solo es conseguir que algoritmos clasifiquen correctamente los diferentes géneros literarios, sino también obtener una descripción empírica sobre estas categorías. Es decir, no solo aplicar algoritmos, sino obtener las características del objeto analizado. Para ello utilizo la anotación semántica mencionada más arriba de las herramientas WordNet y una versión digital del diccionario de María Moliner. Estas dos herramientas agrupan diferentes palabras en grupos que tienen ciertas características comunes, como que están relacionadas con partes del cuerpo, comida, etcétera. ¿Cómo se pueden utilizar estos rasgos para describir los géneros literarios? En primer lugar, estos rasgos deben contabilizarse en cada texto: por ejemplo, el texto *Mayorazgo* de Baroja contiene 1.174 unidades léxicas relacionadas con la categoría de verbos de comunicación de WordNet. Esta frecuencia absoluta es dividida por la extensión total de la novela, obteniendo la frecuencia relativa de cada rasgo. El siguiente paso es seleccionar los textos asociados a cada género. De esas novelas, se calcula la frecuencia media de cada rasgo semántico. Por ejemplo las novelas históricas tienen una frecuencia relativa media de 0.0032 en cuanto a verbos de comunicación, frente a las novelas poéticas, cuya frecuencia es mucho menor: 0.0004. Sin embargo esta media sigue afectada por la frecuencia general de cada rasgo: verbos de comunicación son muy frecuentes en todas las obras, por lo que su media tenderá a ser más alta. El vocabulario sobre narcóticos puede ser mucho menos frecuente en términos generales: un par de referencias puede ser suficiente para ayudar a caracterizar un género específico como la novela de aventuras. Para integrar la frecuencia de cada rasgo, calculo la unidad tipificada, también conocida como *z-score*, en cuyo cálculo se utiliza el valor en el texto analizado, la media en el conjunto de la muestra y la desviación estándar. Esta unidad tipificada expresa lo típica que es la frecuencia de cada rasgo teniendo en cuenta el corpus completo. Esto permite que el vocabulario de

enfermedades pueda obtener unidades tipificadas notablemente altas en las novelas naturalistas, incluso cuando en términos totales su frecuencia no es notable: lo importante es que en el resto de novelas este tipo de vocabulario es aún menos frecuente.

De esta manera, cada subgénero puede ser descrito en rasgos semánticos cuya unidad tipificada es alta. En otras palabras: ¿qué rasgos semánticos son específicos para cada género? Por ejemplo, la novela de guerra tiene como vocabulario específico vocabulario militar (*tropa, artillería, guerra, soldado, milicia, grado*), personas (*gente, enemigo*), vocabulario en relación a armas (*disparar, luchar, arma*) o medios de transporte (*ferrocarril, automóvil*). Para el género de la novela de aventuras los rasgos semánticos más específicos están relacionados con el mar (*costa, marina, barco*), viajes (*geografía, viajar, raza, geología*), sustancias y objetos (*narcótico, agua, palo, armadura*) o medios de transporte (*barco, automóvil*). Estos rasgos semánticos son fácilmente interpretables en comparación con palabras funcionales como conjunciones, preposiciones o artículos. Este tipo de palabras pueden resultar notablemente útiles para los algoritmos, pero resulta opaca a los humanos: ¿qué significa que en un género literario la frecuencia de la palabra *con* sea muy alta? Sin embargo podemos entender la relación entre los barcos y las novelas de aventuras. De esta manera la anotación lingüística está facilitando la clasificación automática, pero también permitiendo que los investigadores tengamos una descripción semántica de lo que particulariza los diferentes géneros literarios.

6. PERSPECTIVAS

El proyecto CLiGS está comprometido a hacer accesible para el final del proyecto todo lo que se pueda publicar, respetando el estado de los derechos de autor de cada texto. Se estudiará qué hacer con aquellas obras cuyos derechos de explotación no hayan expirado para entonces. Las posibilidades van desde la publicación de ciertos datos lingüísticos para investigación que no afecten su estatus legal, hasta la publicación paulatina en el futuro, según los textos vayan quedando libres de derechos.

Está previsto que el canal de publicación siga siendo GitHub y Zenodo de manera similar hasta ahora. Un aspecto positivo de estos

canales es que permiten la flexibilidad de modificar o corregir los datos, manteniendo un control exacto de qué se modificó en qué momento. GitHub incluso permite que otros usuarios señalen errores o mejoras, por lo que la corrección motivada por la comunidad es posible.

Más importante que las posibles correcciones que yo pueda asumir, es que otros investigadores tengan acceso completo a los datos. Esto facilita que colegas puedan utilizar hoy en día y en el futuro una sección del corpus, de sus metadatos o de la anotación, y puedan modificarlo para sus propios objetivos. Por ejemplo, la colección podría ampliarse hacia décadas anteriores, décadas posteriores, podría ampliarse en extensión o en calidad filológica (realizando cotejos de cada novela con una edición de referencia). De la misma manera los metadatos pueden ser expandidos o estandarizados, o aplicar nuevas herramientas lingüísticas que por ahora no existen. Está en nuestras manos que la comunidad de Humanidades Digitales trabajando en textos en español avance más rápido y de manera más sólida. Si cada vez que un proyecto digitalizase, revisase o corrigiese datos, estos se pusiesen a disposición de la comunidad, nuestros esfuerzos serían notablemente más eficaces. Por eso entiendo que el CoNSSA es un paso en el análisis de las novelas de esta época, pero la comunidad de investigadores es responsable de continuar avanzando en esa senda.

7. CONCLUSIONES

En este trabajo he presentado un conjunto de novelas españolas publicadas entre 1880 y 1939 por autores españoles. La recolección de los datos ha sido diseñada para reproducir por un lado, un conjunto completo y representativo de las novelas más canónicas de la época, dentro de un corpus mucho mayor pero sin que la presencia de cada texto esté motivada. Los textos fueron convertidos de diferentes formatos digitales (principalmente formatos web o PDF) a XML-TEI, cuyo formato recoge la totalidad de los datos. El principal objetivo del corpus es analizar los géneros literarios de la novela de esta época. Para esto he utilizado diferentes herramientas computacionales para el análisis lingüístico, así como la anotación manual de numerosos metadatos sobre el protagonista, el lugar y momento de la acción o el narrador, entre otros.

Disponer de datos cuantitativos sobre las novelas permite resumir

grandes cantidades de datos mediante estadística descriptiva. En concreto, he analizado la extensión típica de las novelas de la época, o el grado de felicidad de sus finales: tienden a ser entre claramente tristes y neutros. Los datos también permiten evaluar hipótesis, por ejemplo mediante correlaciones. En este trabajo he observado si los datos no falsifican la hipótesis de Ortega y Gasset sobre que la calidad literaria está en correlación con la extensión del texto. Efectivamente hay una correlación estadística, aunque débil. Con la misma metodología, he analizado otras dos hipótesis: la primera que con el tiempo el porcentaje de puntuación va aumentando; la segunda, que la variación de la extensión de las oraciones también aumenta. Los datos no solo falsifican estas hipótesis, sino que además apuntan en la dirección contraria. Finalmente he explorado y evaluado positivamente una hipótesis que hasta ahora nadie había propuesto: que las novelas se van haciendo más cortas en este período.

¿Cuándo comienza este proceso de acortamientos de las novelas? ¿Se siguió extendiendo durante el siglo XX? ¿Está todavía hoy en activo y cada año las novelas tienden a ser de media más cortas que el año anterior? Los datos aquí presentados no permiten responder a esta pregunta, pero gracias a que este corpus existe, podemos estar más cerca de observar ciertas tendencias que se expanden por varios siglos. Además, los ejemplos muestran el interés que puede tener analizar estadísticamente hipótesis. Esta manera de trabajar no es específica de las Humanidades Digitales, ya que en décadas anteriores había perspectivas sobre humanidades fuertemente formalizadas. La digitalización y los ordenadores sencillamente ponen a nuestra disposición herramientas cada vez más robustas, más rápidas y más sencillas de utilizar. El grado de interés de los análisis computacionales que hagamos dependerá del rigor de nuestra metodología, de la calidad de nuestros datos y de la creatividad de nuestras hipótesis.

REFERENCIAS BIBLIOGRÁFICAS

AGENJO, X. (2015). "Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos". *Ínsula: Revista de Letras y Ciencias Humanas* 822, 12-15.

ALLÉS TORRENT, S. (2015). "Edición digital y algunas tecnologías aliadas". *Ínsula: Revista de Letras y Ciencias Humanas* 822, 18-21.

BAMMAN, D. (2019). *LitBank*. Berkeley: University of California, *https://github.com/dbamman/litbank* [30/06/2019].

CALVO TELLO, J.; HENNY-KRAHMER, U. y SCHÖCH, C. (2018). "Textbox: análisis del léxico mediante corpus literarios". En *Historia del léxico español y humanidades digitales,* D. Corbella Diaz, A. Fajardo Aguirre y J. Langenbacher-Liebgott (eds.), 223-251. Berlín: Peter Lang.

CALVO TELLO, J.; SCHÖCH, C.; RIßLER-PIPKA, N. y KRAFT, T. (2015). "Humanidades Digitales y estudios hispánicos en Alemania". *Voy y Letra 26, 45-61.*

DE LA ROSA PÉREZ, J. (2016). *Corpus Lazarillo*. Londres / Ontario: The University of Western Ontario, *http://ir.lib.uwo.ca/etd/3486/* [30/06/2019].

EVANS, J. D. (1996). *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove: Brooks.

FELLBAUM, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

GARRISH, M. (2011). *What is EPUB 3?* Sebastopol: O'Reilly Media, *http://shop.oreilly.com/product/0636920022442.do [30/06/2019]*.

GRÖTSCHEL, M. (2007). *Deutsches Textarchiv*. Berlín: Berlin-Brandenburgische Akademie der Wissenschaften, *http://www.deutschestextarchiv.de/* [30/06/2019].

GULLÓN, G. *et alii* (1994). "Límites de la novela moderna". En *Historia y crítica de la literatura española, 199-207. Barcelona: Crítica.*

HETTINGER, L.; REGER, I.; JANNIDIS, F. & HOTHO, A. (2016). "Classification of Literary Subgenres". *Digital Humanities im deutschsprachigen Raum Konferenz,* Leipzig (Universität Leipzig) 7-12, Marz, 154-158, *http://dhd2016.de/boa.pdf* [30/06/2019].

JOCKERS, M. L. (2013). *Macroanalysis - Digital Methods and Literary History*. Champaign: University of Illinois Press.

MAINER, J.-C. (2009). *La Edad de Plata (1902-1939). Ensayo de interpretación de un proceso cultural.* Madrid: Cátedra.

MAINER, J.-C. (ed.) (2010). *Historia de la literatura española.* Madrid: Crítica.

MOLINER, M. (1966). *Diccionario de uso del español.* Madrid: Gredos.
MORETTI, F. (2013). *Distant Reading*. Londres: Verso.
MOWAT, B. & WERSTINE, P. (2010). *Shakespeare Folger Library*. Washington: Folger, *https://www.folgerdigitaltexts.org* [30/06/2019].
MUELLER, M.; RAHTZ, S.; PYTLIK ZELLIG, B.; CUMMINGS, J. & TURSKA, M. (2015). "TEI Simple: An Introduction", *http://htmlpreview.github.io/?https://github.com/TEIC/TEI-Simple/blob/master/teisimple.html* [30/06/2019].
MÜLLER, A. C. & GUIDO, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientist*. Beijing: O'Reilly.
NAVARRO-COLORADO, B.; RIBES LAFOZ, M. & SÁNCHEZ, N. (2015). *Corpus of Spanish Golden-Age Sonnets*. Alicante: Universidad de Alicante, *https://github.com/bncolorado/CorpusSonetosSigloDeOro [30/06/2019]*.
OLEZA SIMÓ, J. (2013). *Biblioteca Digital Arte Lope*. Valencia: Universitat de València, *artelope.uv.es/biblioteca* [30/06/2019].
ORTEGA Y GASSET, J. (2009). *La deshumanización del arte, ideas sobre la novela.* Madrid: Castalia.
PADRÓ, L. & STANISLOVSKY, E. (2012). "FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA* (Istanbul) 21-27, May, 2473-2479.
PEDRAZA JIMÉNEZ, F. B. y RODRÍGUEZ CÁCERES, M. (1980). *Manual de literatura española. Pamplona: Cénlit Ediciones.*
PIPER, A. (2016). *Novel450*. Montreal: McGill University, *https://txtlab.org/data-sets/* [30/06/2019].
REEVE, J. (2016). *Corpus of English-language novels (CENLAB)*. Columbia: Columbia University, *https://github.com/JonathanReeve/cenlab [30/06/2019]*.
RIDDELL, A. (2014). *Novels Project Corpus*. Bloomington: Indiana University, *https://github.com/novels-project/novels-corpus* [30/06/2019].
ROJAS CASTRO, A. (2016). *Fábulas mitológicas del Siglo de Oro.* Barcelona: Universidad Pompeu Fabra*, https://github.com/arojascastro/fabulasmitologicas* [30/06/2019].
RUIZ, P.; MARTÍNEZ CANTÓN, C. & CALVO TELLO, J. (2017).

DISCO: Diachronic Spanish Sonnet Corpus. Madrid: UNED, *https://github.com/pruizf/disco [30/06/2019]*.

SÁNCHEZ-MARTÍNEZ, F. (2013). *IMPACT-es diachronic corpus.* Alicante: Universidad de Alicante, *https://www.digitisation.eu/tools-resources/language-resources/impact-es/* [30/06/2019].

SANTA MARÍA FERNÁNDEZ, M. T.; JIMÉNEZ FERNÁNDEZ, C. M. y CALVO TELLO, J. (2017). *Biblioteca Electrónica Textual del Teatro Español, 1868-1936,* Madrid: Universidad Internacional de La Rioja*, https://github.com/GHEDI/BETTE* [30/06/2019].

SCHÖCH, C.; HENNY, U.; CALVO TELLO, J. & POPP, S. (2015). *The CLiGS Textbox*. Würzburg: University of Würzburg, *https://github.com/cligs/textbox* [30/06/2019].

SIMÓN PALMER, M.ª del C. (1997). *Teatro Español del Siglo de Oro.* Ann Arbor: ProQuest, *teso.chadwyck.com* [30/06/2019].

TEXTGRID CONSORTIUM (2016). "TextGrid: A Virtual Research Environment for the Humanities", *https://textgridrep.org/en/* [30/06/2019].

THÉÂTRE CLASSIQUE (2007). París: Université Paris-IV Sorbonne, *http://www.theatre-classique.fr* [30/06/2019].

TRILCKE, P. & FISCHER, F. (2018). *Dracor*. Potsdam-Moscú, *https://dracor.org/* [30/06/2019].

UNDERWOOD, T. (2014). "Understanding Genre in a Collection of a Million Volumes, Interim Report", *https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251* [30/06/2019].

____ (2019). *Distant horizons: digital evidence and literary change*. Chicago: The University of Chicago Press.

Recibido el 13 de marzo de 2020.

Aceptado el 29 de abril de 2020.