Revista de Humanidades Digitales

Diseño de corpus literarios para análisis cuantitativos

Design of Literary Corpus for Quantitative Analysis

José CALVO TELLO
Universidad de Würzburg
jose.calvo@morethanbooks.eu
https://orcid.org/0000-0002-1129-5604

ABSTRACT

This article delves into literary corpus design, studying each stage of the process. Unlike most dissemination studies in the Digital Humanities which are usually focused on the process of data analysis or results, this article focuses on a previous stage, the data curation for corpus creation, task that defends as a central one. This work is a perfect methodological guide for the researcher interested in taking forward a project based on quantitative analysis, as it covers all stages of this process and guides the explanations towards practical cases, and has been written by a researcher trained and experienced in the field, author himself of Spanish textual corpus –Corpus of Spanish Short Stories (Calvo Tello, 2107) and Corpus of Novels of the Spanish Silver Age (CoNSSA) (Calvo Tello 2018)–.

RESUMEN

Este artículo profundiza en los diseños de corpus literarios, planteando y siguiendo cada parte del proceso completo a realizar. A diferencia de la mayoría de los estudios de divulgación de las Humanidades Digitales, que suelen tener por objeto el proceso de análisis de datos o los resultados del mismo, este artículo focaliza el estadio previo, el de preparación de datos para la confección del corpus, tarea que defiende como central. Este trabajo resulta una perfecta guía metodológica para el investigador interesado en sacar adelante un proyecto basado en análisis cuantitativos en tanto que cubre todos los estadios del proceso, orienta las explicaciones hacia casos prácticos, y ha sido confeccionado por un investigador formado y con experiencia en el campo, autor él mismo de corpus de textos en español –Corpus of Spanish Short Stories (Calvo Tello, 2107) y Corpus of Novels of the Spanish Silver Age (CoNSSA) (Calvo Tello, 2018)–.

RHD 4 (2019)

ISSN
2531-1786

KEYWORDS

Stylometry, Supervised and Unsupervised Automatic Learning, Data Analysis.

PALABRAS CLAVE

Estilometría, aprendizaje automático supervisado y no supervisado, análisis de datos.

Recibido 01/07/2019 – Aceptado 10/07/2019

1. INTRODUCCIÓN

Conceptos como lectura distante (Moretti, 2005) o macroanálisis (Jockers, 2013) han modificado algunas áreas de las Humanidades de los últimos años. Aunque las Humanidades (Digitales o no) difícilmente pueden trabajar con cantidades de datos que otras áreas considerarían big data (Schöch, 2013), la cantidad de datos disponibles o deseados se ha multiplicado. En ocasiones ha sido el traspaso de ciertas tecnologías, como *word embeddings* (Blei, Andrew y Jordan, 2003; Mikolov, Chen, Corrado y Dean, 2013) o aprendizaje automático (Müller y Guido, 2016), el que ha reclamado la búsqueda de mayores cantidades de datos. Esta manera cuantitativa de trabajar contrae nuevas rutinas, herramientas, filtros y visualizaciones a las que hasta ahora no estábamos acostumbrados y con las que debemos familiarizarnos antes de poder realizar tales análisis. El diseño de la recogida de datos define en buena medida los pasos o las conclusiones a las que podremos llegar. Un diseño deficiente de recogida de datos puede trastocar el análisis posterior en su conjunto.

Este trabajo pretende aportar luz en este camino desde la experiencia de haber trabajado en la composición de varios corpus que posteriormente han sido analizados mediante metodologías como estilometría, aprendizaje automático supervisado y no supervisado o grafos y redes sociales. Debido a mi experiencia y al marco de la publicación, me centraré en la recogida de textos literarios.

2. LOS DATOS: TAREA CLAVE EN LAS HUMANIDADES DIGITALES

Algunas especialidades profesionales tienen asociadas de manera clara algunos productos o tareas dentro de la sociedad: consideramos que los informáticos son principalmente quienes desarrollan programas informáticos, los bibliotecarios son los encargados de mantener y desarrollar las bibliotecas, los filólogos son los especialistas que preparan un texto histórico para su edición, los editores son los que seleccionan, invierten y publican un texto, etc. Sin embargo, la cuestión sobre quién debe preparar y publicar datos humanísticos para la investigación sigue siendo un misterio. Hay una laguna de responsabilidad sobre quién debe asumir la ingente tarea de seleccionar, filtrar y poner a disposición datos que el resto de la comunidad de investigadores puedan utilizar de manera completa.

Diferentes tipos de proyectos han asumido partes de esa tarea: la RAE (Real Academia Española) con corpus y diccionarios, Biblioteca Virtual Cervantes con la edición de textos, proyectos de investigación con corpus y colecciones de texto, bibliotecas públicas con la digitalización de sus fondos, empresas como ProRequest con TESO[1] o Google con Google Books, editoriales, comunidades de usuarios (Wikimedia, ePubLibre) u organizaciones (Project

[1] Accesible desde: http://teso.chadwyck.com/.

Gutenberg, Archive.org). Pero sigue sin haber nadie que lo haya asumido como su principal tarea, a quien –o a donde– los investigadores puedan ir en busca de datos. Esto se debe en parte a la influencia que la informática ejerce sobre nuestro campo: los informáticos trabajan principalmente con datos que usuarios o procesos automáticos generan, como la participación de usuarios en redes sociales, su comunicación, su navegación entre páginas, etcétera. La generación de estos datos en formato digital controlado es automática y su acceso es de relativa facilidad. Esto contrasta con la dificultad de los datos digitales en las Humanidades, que en algunos casos comienza por buscar el objeto en librerías de segunda mano o bibliotecas en algún lugar del mundo.

En mi opinión, los humanistas digitales debemos asumir la tarea de la selección, preparación y publicación de datos para la investigación. Debemos hacerlo colaborando estrechamente con el resto de los actores, especialmente con Bibliotecas, y trabajando en proyectos específicos en los que la formación de los datos sea el principal objetivo y no un mero paso preparativo de menor importancia.

3. CASO DE USO

Antes de comenzar a crear el corpus de textos literarios, debemos poder responder dos preguntas:

- ¿Qué tipo de análisis queremos realizar?
- ¿Qué textos queremos analizar?

Aunque son dos preguntas diferentes, las respuestas no son completamente independientes. El tipo de análisis determina en cierta medida tanto las épocas como los géneros que podemos investigar. Algunas de las nuevas metodologías desarrolladas en los últimos años, principalmente *deep learning* (Chollet, 2018), pero también *topic modeling* (Blei et al., 2003) o *word embeddings* (Mikolov et al., 2013), requieren de grandes cantidades de datos. Esto imposibilita el trabajar con ciertos géneros literarios o épocas de los que, en el mejor de los casos, solo podremos obtener algunos cientos de textos cortos. En estos casos, la opción más sencilla es trabajar en géneros literarios de los que ya dispongamos de cantidades aceptables de documentos digitalizados, los cuales posean, a su vez, suficiente texto (algunos cientos de novelas o piezas de teatro, algunos miles en el caso de poesía).

No solo la cantidad de datos tiene un papel fundamental en las primeras elecciones, sino también los rasgos o subunidades del texto que queramos analizar. Si nuestro objetivo es estudiar el desarrollo de ciertos rasgos a lo largo de las diferentes subunidades del texto –por ejemplo, el cambio de pronombres en las diferentes estrofas de un poema, o el cambio de personajes a lo largo de los diferentes capítulos de novelas–, deberemos marcar de manera unívoca esas subunidades. Si decidimos crear un corpus de novelas en texto plano donde los

encabezamientos de capítulos no se distingan de manera unívoca del resto de párrafos, y una vez terminado queremos utilizar la información de capítulos, deberemos reestructurar cada texto del corpus, probablemente asumiendo otro formato más descriptivo.

El hecho de que queramos utilizar un dato específico en nuestro análisis no significa que tengamos que marcarlo nosotros a mano. Por supuesto podremos utilizar herramientas automáticas que anoten y extraigan buena parte de los datos que queremos, como por ejemplo información morfológica, sintáctica, léxica, entidades personales, etc. Aun así, hay que tener en cuenta que este tipo de herramientas suelen estar diseñadas y testadas con textos periodísticos contemporáneos. En muchos casos su aplicación a textos literarios o históricos puede contraer importantes deficiencias. En este punto del desarrollo del trabajo sería necesario informarse sobre qué herramientas anotan la información deseada, cuáles son sus resultados, si las herramientas son abiertas, si los desarrolladores las han documentado u ofrecen talleres sobre ellas y hasta qué punto son aplicables para los textos literarios que estamos analizando. Si alguno de los puntos no está claro, suele ser de ayuda ponerse en contacto con los desarrolladores, creándose quizás una posibilidad de colaboración en la que la herramienta gane en cobertura y nuestros datos sean anotados como queríamos.

Con la información que no esperamos anotar automáticamente existen dos tendencias igualmente peligrosas:

1. Querer anotar de manera cuidadosa todos los rasgos e informaciones posibles del texto, al no estar seguros de qué se querrá analizar posteriormente.
2. Mantener poca información marcada de manera explícita para terminar el corpus lo más rápido posible.

Cada proyecto debe encontrar un equilibrio entre ambos extremos que permita avanzar al proyecto. Es habitual que, una vez terminada la primera versión de un corpus, se revise cada uno de los textos añadiendo algún metadato o haciendo explícita o unívoca cierta información. Una posible estrategia realista y eficiente es borrar el mínimo posible de información que nos aporta el formato de entrada. Por ejemplo, si queremos crear un corpus de novelas que en su mayoría proviene de una comunidad que las publica en formato de libro electrónico ePub, en los que están marcados tanto cursivas como encabezamientos, es sensato mantener ambas informaciones en nuestro corpus: no tenemos que marcarlo nosotros y nuestro corpus tendrá más datos.

4. TIPO DE CORPUS

Otra cuestión importante para cada proyecto es ¿hasta qué punto podré generalizar mis análisis? El hecho de que trabajemos con un corpus de obras de teatro del Siglo de Oro no quiere decir que trabajemos con todas las obras de teatro de esa época. El objeto que

analizamos (el corpus) no tiene que ser el objeto que nos interesa (el teatro del Siglo de Oro). Otras ramas científicas abordan este problema acudiendo a la estadística y sus conceptos de población y muestra (Evans, 1996, pp. 14-16; Haslwanter, 2016, pp. 75-76). La población sería el conjunto de entidades por el que nos interesamos (por ejemplo: la población de Italia, los enfermos de cáncer entre 2005 y 2007 en Madrid, las obras de teatro del Siglo de Oro), mientras que una muestra es un conjunto concreto de casos de los que tenemos los datos (cierta cantidad de italianos a los que hemos encuestado, los datos médicos de los pacientes de un hospital, un corpus electrónico de obras de teatro del Siglo de Oro). Dependiendo de cómo diseñemos nuestra muestra, el corpus, tendremos mayor o menor capacidad de generalizar nuestros resultados sobre el resto de la población. Definiendo la población y la muestra estaremos definiendo el tipo de corpus con el que vamos a trabajar. Recientemente una valiosa publicación de Schöch (2017) ha tematizado esta cuestión en relación con los corpus literarios. Existen cuatro posibilidades principales en que un corpus representa o no una población completa.

En primer lugar, y a diferencia de otras muchas áreas científicas, hay que tener en cuenta que las Humanidades (especialmente las Digitales) sí suelen trabajar con poblaciones completas. Un ejemplo sencillo es la Biblia: si queremos analizar la aparición de la palabra *Dios* en la Biblia, podremos acceder a una versión digitalizada y realizar esa búsqueda. No debemos trabajar con una selección de libros o versículos bíblicos, podemos analizar la población completa y aplicar estadística descriptiva que llegue a conclusiones sobre ese aspecto. De una manera similar, hoy en día podemos realizar estudios léxicos sobre todas las novelas publicadas de Pardo Bazán: estaremos trabajando con la población completa y no con una muestra. Una población concreta puede contener a su vez poblaciones más reducidas pero también completas: por ejemplo, podemos decidir trabajar con el Nuevo Testamento, o incluso con una población completa aún más pequeña: los cuatro Evangelios. De esta manera tenemos tres poblaciones completas de diferentes tamaños en las que las menores son contenidas por las mayores, elegidas según nuestros intereses de análisis.

En algunas ocasiones trabajaremos con un conjunto de datos sobre los que no estaremos seguros de que realmente representen la población completa o sean los mejores acercamientos actualmente posibles a la población completa. Esto ocurre principalmente con textos no impresos. Es difícil definir la población de obras de teatro escritas por Cervantes ya que hay textos que se creen perdidos o cuya autoría es discutida (Calvo Tello y Cerezo Soler, 2017). El proyecto ArteLope[2] (ha digitalizado decenas de obras de teatro adscritas con mayor o menor seguridad a Lope; sin embargo, se sabe que escribió numerosas obras más, de las cuales a menudo solo conocemos su título. La estadística no da respuestas sencillas a este tipo de

[2] Accesible desde: https://artelope.uv.es/biblioteca/.

problemas típicos de las Humanidades. Posiblemente estadistas y humanistas debamos trabajar de manera conjunta intentando solucionar estos problemas. En muchos casos es aún mucho más complejo: ¿Cómo se puede definir la población de obras de teatro escritas en el Siglo de Oro? ¿De entradas de blogs literarios? ¿De cartas del siglo XV? ¿De poesía oral? Cuanto más efímero sea el medio, y cuánto más remota sea su composición, mayor será la dificultad de trabajar con una población definida de manera estricta.

La segunda manera de relacionar un corpus con una población completa es la manera estándar de la estadística de crear una muestra aleatoria y utilizar estadística inferencial. Para ello deberemos tener una lista de las unidades que pertenecen a nuestra población, definir una cantidad de datos que queremos recoger, y elegir de manera aleatoria de qué casos recogeremos la información. La variación de los datos de la muestra estará normalmente distribuida, por lo que los test estadísticos podrán ayudarnos a observar si nuestras conclusiones son esperables no solo en la muestra sino en la población completa a la que no tenemos acceso.

Si aplicamos esta manera de trabajar a las Humanidades, seguiríamos los siguientes pasos: en primer lugar, conseguimos una lista de todas las novelas publicadas en el siglo XIX (filtrando aquellas que sean nuevas –El Lazarillo o El Quijote se volvieron a publicar en el siglo XIX, pero probablemente no queramos analizarlas como obras del siglo XIX–, y filtrando también textos traducidos). Elegimos una cantidad o porcentaje de los textos que queremos tener en nuestra muestra, por ejemplo el 10%. Seleccionamos de manera aleatoria qué textos concretos serán parte de nuestra muestra: es decir, aplicamos alguna función que genere números aleatorios sin tener en cuenta si esos textos están digitalizados o no, o la importancia del autor o del texto. Una vez tengamos un índice de qué libros deberán estar en nuestro corpus, conseguimos digitalizaciones de cada uno de ellos (digitalizando nosotros mismos los que no lo estén). Al terminar, conseguiríamos un corpus al que puede aplicarse no solo estadística descriptiva, sino la estadística inferencial con la que podremos llegar a conclusiones no sobre el 10% de textos que se encuentran en nuestro corpus, sino sobre el total de la población. No conozco ningún proyecto de humanidades que haya creado una muestra aleatoria real de una población definida. Este hecho arroja algunas preguntas: ¿Por qué? ¿Es acaso incompatible esta metodología con los análisis literarios? ¿O sencillamente los humanistas no se han concienciado de trabajar de esta manera?

En tercer lugar, nos encontramos con muestras balanceadas según ciertos parámetros, muy frecuentes en algunas áreas de investigación más cercanas a la Lingüística. La lingüística de corpus suele elegir una dimensión, ya sea del texto o del productor del texto, y se diseña el corpus para que cierto valor represente cierto porcentaje del total del corpus (ya sea en cantidad de textos o en cantidad de tokens). Por ejemplo, los corpus de la RAE están balanceados geográficamente: la mitad de las formas (unidad similar a token) del corpus CREA (Corpus de Referencia del Español Actual) provienen de España, la otra mitad de los países

hispanoamericanos. Este diseño permite analizar posteriormente de manera sencilla si una palabra está sobrerrepresentada o subrepresentada en parte del corpus. El diseño de este tipo de corpus permite llegar a conclusiones sobre las dimensiones con las que ha sido diseñado, pero no sobre otras dimensiones que no han sido influyentes en ese proceso. Por ejemplo, el CORDE (Corpus Diacrónico del Español) no está balanceado en cuanto al género del texto de manera diacrónica. En parte esto es imposible: un corpus diacrónico no puede tener la misma cantidad de novelas a lo largo de todos los siglos ya que durante muchos siglos ese género no se practica.

En cuarto y último lugar nos encontramos el caso más habitual en las Humanidades Digitales, los corpus oportunísticos: proyectos que coleccionan tantos textos como ha sido posible. Estos provienen o bien de colecciones preexistentes como Project Gutenberg[3], TextGrid Repository[4], Théâtre Classique (Fièvre, 2007), Cervantes Virtual[5], o de una digitalización propia. A estos corpus no subyace ninguna población completa. Aunque estas colecciones sean bienintencionadas y en muchos casos razonables, las conclusiones a las que se llega mediante su análisis no deben generalizarse a conjuntos mayores de textos.

5. DEFINIR LA POBLACIÓN

Como hemos observado, la pregunta sobre la población con la que trabajamos es central. Si decidimos trabajar con una población concreta, tendremos que realizar pasos consecutivos: definir la población, su fuente de datos y listar los textos que la componen.

En primer lugar, debemos formalizar qué población queremos analizar, por ejemplo: las novelas nuevas publicadas en el siglo XIX por autores españoles. Como se aprecia, la población no solo utiliza rasgos de los textos (novelas, nuevas, publicadas, siglo XIX) sino también de quien produce el texto (autores españoles). En segundo lugar, deberemos decidir de qué tipos de fuentes de información extraeremos el listado. Para ello podemos ir a fuentes primarias y recolectar de ellas el listado, por ejemplo, accediendo a los catálogos de las editoriales. Esto contrae el paso de tener que listar las editoriales de la época y posteriormente acceder a sus catálogos. En contraposición, podemos acceder a fuentes secundarias como catálogos o manuales (bibliotecarios, literarios) lo más extensos posibles.

Una vez hemos elegido la fuente de datos para nuestra población, deberemos enlistar el conjunto de obras. Para ello sugiero realizarlo en un formato tabular (similares a las de Microsoft Excel, como Calc en LibreOffice) y recoger de cada obra ciertos metadatos básicos como son el título, el autor, el año de publicación y otros metadatados básicos (género del

[3] Accesible desde: https://www.gutenberg.org/.
[4] Accesible desde: https://textgridrep.org/.
[5] Accesible desde: http://www.cervantesvirtual.com/.

autor, género literario) si la fuente para nuestra población los ofrece de manera sencilla. También es útil añadir una columna que sirva de identificador numérico.

Una vez tenemos el listado de nuestra población y su extensión, deberemos decidir qué tipo de corpus queremos construir: si queremos aspirar a la población completa, si queremos intentar una muestra aleatoria, si queremos construir un corpus balanceado o si nos es suficiente con un corpus oportunístico. Un dato fundamental para saber la cantidad de trabajo que representará cada uno de estos pasos es saber el estado de la digitalización de los textos. Por eso puede ser interesante buscar en Internet si los textos están digitalizados, en qué fuentes y en qué formatos. Una posible manera de hacerlo sería elegir una cantidad de textos (por ejemplo, 30), seleccionarlos a partir de nuestra población de manera totalmente aleatoria, y buscar digitalizaciones tanto en buscadores genéricos así como en los principales proyectos de digitalización que esperamos podrían haberlo digitalizado. Con estos datos básicos estaremos en la posición de decidir hasta qué punto nuestro proyecto puede coleccionar o digitalizar textos y podremos definir qué tipo de corpus queremos realizar.

6. OBTENCIÓN DE DIGITALIZACIONES

El siguiente paso es conseguir las digitalizaciones para construir nuestro corpus. Por razones prácticas preferiremos reutilizar textos digitalizados por otros proyectos. Por eso es inteligente, en primer lugar, no solo buscar cada uno de los textos sino también si ha habido otros proyectos que hayan digitalizado parte de nuestros textos. En la actualidad el sistema de control de versiones GitHub se está convirtiendo en una manera muy habitual de compartir código y datos entre diferentes proyectos, como lo demuestran los corpus BETTE (Santa María Fernández, Jiménez Fernández y Calvo Tello, 2017), DISCO (Ruiz Fabo, Bermúdez Sabel, Martínez Cantón y Calvo Tello, 2017), Corpus of Spanish Golden-Age Sonnets (Navarro-Colorado, 2015), *Dracor* (Fischer, Trilcke y Orekhov, 2018), o Théâtre classique (Fièvre, 2007). Esta plataforma además está perfectamente interconectada con el sistema de repositorios europeos a largo plazo apoyado por el CERN (European Organization for Nuclear Research), Zenodo, del que hablaré más adelante. Por ejemplo, en esta plataforma se encuentran archivados los corpus de Textbox (Schöch, Henny, Calvo Tello y Popp, 2015) mediante identificadores permanentes. Otros corpus se encuentran disponibles para descarga completa en servidores privados o universitarios, como es el caso de IMPACT-es (Sánchez-Martínez, Martínez-Sempere, Ivars-Ribes, y Carrasco, 2013). Otros proyectos no publican sus textos de manera abierta, pero sí pueden aceptar darlos para colaboraciones específicas. Conseguir el contacto correcto en el momento correcto puede ahorrar a nuestro proyecto cientos de horas de trabajo.

Si comprobamos que no hay ningún proyecto académico específico que haya trabajado la época o el género en los que estamos interesados, tendremos que recolectar las

digitalizaciones de diferentes fuentes. Algunos proyectos académicos de mediano tamaño han conseguido digitalizar decenas de textos en formatos digitales nativos con un mayor o menor control filológico. Entre estos proyectos se cuentan algunos como Clásicos Hispánicos (Jauralde Pou, 2013) o Canon 60[6]. Ambos proyectos cuentan con digitalizaciones en XML-TEI de los que exportan versiones en XHTML o eBook.

De menor calidad tanto en lo digital como en lo filológico, pero de una cobertura mucho mayor, son proyectos generalistas como Project Gutenberg o Cervantes Virtual. Lamentablemente este último proyecto, pionero en España digitalizando miles de textos en XML-TEI, nunca ha compartido de ninguna manera sus fondos, ni siquiera para investigación. Otro de los proyectos con mayor cobertura –aunque ya fuera del mundo académico, e incluso al margen de cualquier tipo de institucionalización– es el portal ePubLibre[7]. Este interesante proyecto está formado por una comunidad de usuarios que formatea y pone a disposición miles de textos en formato eBook.

Un proyecto que publica sus textos desde 2017 en formato libro electrónico es la Biblioteca Digital Hispánica de la BNE. Aun así, la enorme mayor parte de las digitalizaciones de este fondo se encuentra en PDF. También en este formato se encuentran los fondos de Archive.org[8], un vasto proyecto de archivo digital que colabora, entre otros agentes, con bibliotecas institucionales alrededor del mundo que ponen a disposición sus fondos. A su vez podemos encontrarnos otras bibliotecas similares, más o menos regionales o específicas.

El formato PDF ofrece algunas ventajas frente a otros formatos, como es el hecho de que muchos aspectos visuales del documento digitalizado se vean recogidos sin intermediación. Aun así, cuando analizamos el texto de una obra, el formato PDF no es el punto de origen deseado: en muchos casos tendremos que utilizar software de OCR para poder tener una versión digital del texto, aunque también es posible que el PDF contenga tanto la imagen original digitalizada como una versión del texto, reconocido automáticamente, y sobrepuesto en la imagen. En ese caso es necesario evaluar la calidad del texto reconocido.

Por último, para los casos en los que no hayamos encontrado digitalizaciones previas o sean de baja calidad, deberemos digitalizar nosotros mismos los libros. Para ello necesitaremos en primer lugar conseguir imágenes de buena calidad (el estándar es 300 dpi). Estas pueden conseguirse utilizando escáneres específicos de digitalización de libros, en muchas ocasiones accesibles en bibliotecas universitarias. Este tipo de escáneres tienen la ventaja de ser más rápidos y menos agresivos para los libros que los escáneres tradicionales; además, el libro queda intacto (aunque los lomos pegados pueden verse afectados por la apertura al escanear). Las desventajas de estos escáneres recaen tanto en la velocidad como en la calidad: un libro de

[6] Accesible desde: https://tc12.uv.es/?page_id=3626.
[7] Accesible desde: https://epublibre.org/.
[8] Accesible desde: https://archive.org/.

unas 200 páginas suele tardar unos 40 minutos en ser escaneado y el resultado no suele alcanzar los 300 dpi, pudiendo estar las páginas deformadas por la apertura. Frente a estos escáneres se encuentran las máquinas que digitalizan por las dos caras, una por una cada página del libro que previamente ha sido guillotinado. Las ventajas de estas máquinas son por supuesto una mayor velocidad y una mayor calidad. Como desventaja está el hecho de que el proyecto debe adquirir ejemplares que puedan guillotinarse sin representar una pérdida material notable. De una categoría aún mayor son los escáneres-robot que pasan las páginas al mismo tiempo que las digitalizan. Este tipo de máquinas se encuentran por ahora fuera de la mayoría de los circuitos universitarios debido a su coste.

Una vez tengamos las imágenes digitalizadas, deberemos procesarlas mediante un programa de reconocimiento de OCR. El capítulo de Percillier (2017) sobre creación de corpus literarios ofrece comparaciones sobre diferentes opciones. En mi experiencia, los productos de la empresa Abbyy, como FineReader, funcionan de manera muy positiva cuando se parte de digitalizaciones ediciones y de buena calidad. Este tipo de productos tienen licencias asumibles para proyectos de investigación y exportan sus datos a numerosos formatos, entre ellos varias versiones de XML o XHTML, desde los cuales podremos continuar convirtiendo el archivo hasta el formato maestro de nuestro proyecto.

7. TEXTO

Antes de comenzar a convertir textos, debemos tomar una serie de decisiones relativas al texto y los documentos que formarán nuestro corpus. Probablemente la más importante es la pregunta relacionada con el formato maestro (master format). Con este término me refiero al formato principal del proyecto que es controlado manualmente y que, en caso de encontrarse errores o de quererse ampliar la información, puede ser modificado. Los diferentes tipos de formatos pueden organizarse en los siguientes grupos:

- Formatos Word: aunque denostados por muchos, este tipo de formato (DOC, DOCX o sus versiones de LibreOffice) siguen siendo la base para ediciones de muchos grupos de investigación y el formato elegido por casi cualquier humanista para escribir prosa científica. Ofrecen una interfaz visual que todos conocemos, pero resultan poco útiles para analizar o buscar, y su capacidad para ser convertidos a otros formatos es limitada. Algunas de las versiones de este formato están basadas a su vez en formatos XML.
- PDF: es el formato más extendido para digitalización de documentos. A pesar de su pobre estructura textual, este formato ofrece la ventaja de recoger en un solo documento de manera sencilla información visual que en otros formatos se tiende a perder o cuya codificación explícita resulta costosa.

- Bases de datos: aunque en claro retroceso en círculos académicos de humanidades, muchos proyectos siguen conteniendo sus textos en bases de datos relaciones (típicamente SQL). Las ventajas de este formato son su sencilla integración con lenguas de programación para web como PHP, y que casi cualquier servidor, tanto privado como institucional, los soporte –en comparación con bases de datos NoSQL–. Alguna de sus desventajas: dificultad de ingresar datos de manera manual, dispersión de los datos en diferentes tablas y dificultad de representar al mismo nivel texto y nodos (por ejemplo, cursiva en un texto).
- Formatos web: en sus diferentes versiones (HTML, XHTML) sirve de puente entre el resto de formatos: todos los demás o bien pueden convertirse, o bien pueden ser convertidos en un HTML aceptable de manera sencilla. Ofrece las ventajas de ser un formato extendidísimo, sencillo de aprender, de sintaxis flexible, integración con otras tecnologías (CSS, JavaScript, PHP, etc.) y que contiene elementos semánticos. Como desventaja principal está el hecho de que sus elementos no están diseñados para textos históricos o literarios: aunque hay un elemento específico para vídeo, no hay manera estándar de codificar versos.
- eBook: hijo directo de los formatos web –con sus fortalezas y debilidades–, sus principales ventajas son la de encapsular en un solo archivo todas las partes de un texto, así como metadatos –codificados en XML–, además de su elegante visualización en libros electrónicos. Tiene, sin embargo, dos desventajas. La principal es la división de formatos: el estándar (ePub), y el más extendido (mobi), propiedad de Amazon. La otra es su desarrollo tortuoso durante los últimos años: la institución que dirige su creación, llamada IDPF, lanzó el formato ePub3 en 2007 (Garrish, 2011), pero por diferentes razones no fue asumido mayoritariamente por editoriales ni instituciones públicas. En 2017 el IDPF fue subsumida por el W3C (Reid, 2017). Desde entonces se ha anunciado el desarrollo de un nuevo formato, Packaged Web Publication (Gylling, 2017) que en principio vendrá a ocupar el lugar del ePub.
- Texto plano: es el formato más extendido entre informáticos y lingüistas computacionales, principalmente cuando el texto ha de ser entregado a algún programa que debe analizarlo o anotarlo. Su principal ventaja es su sencillez: no pueden ocurrir errores de formato ya que no hay ningún formato. Su principal desventaja es también su sencillez: debemos eliminar la información de tipos de textos (como encabezamientos, párrafos normales, notas, versos, etc.).
- XML: es en realidad solamente una sintaxis de cómo debe funcionar la estructura del archivo: los elementos podremos definirlos nosotros mismos. En realidad, algunos de los formatos arriba mencionados están basados –de

manera más o menos estricta– en la sintaxis XML. Uno debe tener una razón poderosísima para comenzar de nuevo el trabajo de definir elementos y atributos XML en lugar de utilizar un conjunto ya definido, y compartido por una comunidad. Aunque hay diferentes conjuntos definidos para texto (como DocBook) (Harold y Means, 2004, pp. 95-98), el más extendido, principalmente en las Humanidades Digitales, es el de la Text Encoding Initiative (TEI).

- TEI: siglas de Text Encoding Initiative (Burnard, 2014; Agenjo, 2015), es un lenguaje de marcado que permite recoger de manera semántica tanto los metadatos como el texto en un mismo archivo. Nos permite tanto hacer complejas ediciones (Rojas Castro, 2017; Fradejas Rueda, 2018), como realizar versiones sencillas del texto con una fuerte similitud a HTML, pero pudiendo codificar de manera explícita elementos básicos de teatro o poesía que no tienen correspondiente en web. Como desventajas están el hecho de que esté menos extendido que HTML, sea más complicado de aprender, falte soporte en muchas herramientas –aunque algunas como Voyant Tools o Stylo lo acepten–, y sea necesario que la sintaxis XML sea respetada escrupulosamente (en comparación con web o eBooks). Su conversión a otros formatos se realiza mediante tecnologías como XSLT o xQuery (Allés Torrent, 2015), o expresiones regulares, aunque el TEI Consortium ofrece numerosas posibilidades online. En los últimos años diferentes proyectos (como eXist-db) han lanzado maneras de construir interfaces que hagan de nuestros archivos XML una base de datos sobre la que crear páginas webs.

La complejidad intertextual de los géneros analizados repercute en la elección del formato. Cuantos menos tipos textuales diferentes contenga cada obra, menos problemática será la elección de un formato semánticamente sencillo como HTML o texto plano. Principalmente si elegimos este último formato hay que cuidar que nuestros intereses inmediatos puedan llevarnos a eliminar cualquier elemento textual que no queramos en la actualidad, pero que podamos echar de menos en poco tiempo. Por ejemplo, si elegimos hacer un corpus de teatro en texto plano, quizás decidamos eliminar todas las acotaciones y nombres de personajes en obras de teatro porque en un primer momento nos interese estudiar el vocabulario de los parlamentos. Sin embargo, en un par de años podremos estar interesados en analizar otros aspectos de las obras o el vocabulario en relación con escenas o personajes, información que habíamos eliminado al crear el corpus y que debemos reestablecer. De utilizar un formato más complejo tendríamos los datos accesibles.

Además de sobre el formato, deberemos tomar una serie de decisiones sobre los archivos, los metadatos, la unidad central de texto, el tipo de edición, etc. La revista RIDE ha

publicado en 2017 y 2018 varios números (6 y 8) de reseñas de colecciones de textos y, junto a ellas, publicaron los criterios para revisarlas (Henny-Krahmer y Neubar, 2017). Estos resultan útiles también como *checklist* para corpus en desarrollo. Algunos de los puntos más importantes son los que siguen:

- ¿Qué datos exactamente se encuentran en el archivo maestro: metadatos, anotación lingüística, texto, versiones del texto, imágenes?
- ¿Cuál es la unidad principal de texto de la que queremos recoger metadatos? Por ejemplo, si queremos realizar un corpus de cuentos: ¿utilizamos un archivo por cada cuento, por cada colección de cuentos, por cada autor? ¿Queremos anotar metadatos solo a nivel de la colección, de cada cuento, de ambos?
- En el caso de que, por ejemplo, trabajemos con textos anteriores al siglo XVIII: ¿qué modernización deberán tener los textos? Si queremos anotar lingüísticamente los textos será mejor modernizarlos, mientras que si estamos interesados en aspectos históricos concretos de la forma del texto, los necesitaremos sin modernizar. De cualquier manera, cualquier corpus debería tener un tipo de modernización homogéneo, ya que una parte considerable de los tipos totales pueden verse afectados y esa información puede causar ruido en nuestros resultados.
- ¿Qué tipo de edición necesita mi proyecto: filológica crítica, filológica, profesional, social, oportunística? Dependiendo de la época con la que trabajemos, las dificultades propias del texto (pensemos en casos como La Celestina o El Buscón) y el tipo de análisis que queramos analizar, necesitaremos elevar las exigencias sobre la edición.

8. FLUJO Y CONTROL DE VERSIONES

El formato maestro no tiene que ser el único formato con el que trabajemos: podemos derivar a partir de él otros formatos para diferentes usos específicos. Dependiendo de lo rico que sea semánticamente el formato maestro, podremos exportarlo a otros formatos correctamente; el paso inverso no es posible. De esta manera, si hemos codificado el corpus en XML-TEI podremos convertirlo en cualquier otro formato –con mayor o menor dificultad: la conversión de XML-TEI a PDF, aunque teóricamente sencilla, en la práctica resulta compleja–, mientras que si hemos elegido texto plano la conversión es imposible.

Sea cual sea nuestro proceso de conversión, es una estrategia sabia archivar también los documentos intermedios que manejamos –principalmente los documentos originales, sean de la fuente o formato que sean–. La historia de la cultura está llena de ejemplos de cómo se han desechado documentos intermedios creyendo que no tenían ninguna información o valor. Años

después los mismos agentes (por ejemplo, las editoriales de finales del siglo XX guardando solo PDF de sus impresos) o investigadores descubrían que todos esos documentos sí que tenían valor histórico o funcional. No podemos estar seguros de qué nos interesará investigar en algunos años, ni de qué procesos de conversión serán más sencillos dentro de algunas décadas. De esta manera cada proyecto no debe tener un solo archivo en el que realice tanto la anotación como los análisis, sino que puede tener un conjunto de formatos que se vayan convirtiendo de unos a otros. Esta conversión debería ser automática a partir del formato maestro, que debe ser el que se corrija o mejore temporalmente. Aunque estos flujos de trabajo no suelen ser tematizados en conferencias o artículos, su correcto diseño e implementación puede ser de enorme valor para el proyecto. Veamos un par de ejemplos de proyectos diferentes que conozco de primera mano.

El primer ejemplo proviene de la colección Clásicos Hispánicos (Jauralde Pou, 2013). El objetivo de esta colección es publicar ediciones filológicas cuidadas por un especialista, pensadas para la lectura en libro electrónico. El punto de partida debía ser un formato en el que filólogos expertos pudiesen trabajar cómodamente, por lo que decidimos partir de formatos Word. Estos son transformados en XML-TEI de manera semi-automática, confirmando que cierta información filológica básica es codificada de manera correcta. Este es el formato maestro de la colección, sobre el que corregimos posibles errores generados en cualquier punto del proceso. A partir de él lo convertiremos de manera automática al formato estándar de libro electrónico ePUB, desde el que a su vez generamos el mobi. En caso de que el texto vaya a ser utilizado para investigación y se necesiten fragmentos específicos en texto plano, estos son extraídos directamente desde el formato maestro.

El segundo ejemplo proviene de la experiencia de la creación de los dos corpus de textos en español Corpus of Spanish Short Stories (Calvo Tello, 2017) y Corpus of Novels of the Spanish Silver Age (CoNSSA, Calvo Tello, 2018), el primero publicado de manera íntegra y el segundo de manera parcial como parte de Textbox (Schöch et al., 2015), realizados en la Universidad de Würzburg. Para estas colecciones descargamos los textos desde fuentes con diferentes formatos (HTML, ePub, PDF), y otros fueron digitalizados por nosotros mismos. A partir de ellos los convertimos a XML-TEI, nuestro formato maestro, revisando manualmente que el texto estuviese completo. Este formato es transformado en el resto de los formatos utilizados por el proyecto: XML-TEI anotado, PDF o texto plano. Todos los formatos de cada texto se encuentran en nuestra cuenta de GitHub.

De nuevo mencionamos GitHub, en este caso como manera de evitar los conflictos de versiones. La cantidad de coautores en proyectos está incrementándose; mover los archivos de un colaborador a otro mediante correos o dispositivos externos (discos duros, USB) es un sistema precario que rápidamente forma conflictos de versiones: cada investigador puede realizar modificaciones en paralelo de manera independiente. Por eso cada proyecto debe utilizar herramientas que no solo posibiliten el acceso a distancia y su edición –como sería el popular

Dropbox o servicios similares–, sino que también permitan que los diferentes estadios del proyecto queden correctamente guardados, pudiendo regresar a cada punto histórico concreto. Esto es lo que las herramientas basadas en Git permiten. Esta tecnología, originalmente creada para organizar el diseño de software –un ejemplo dentro de las DH es stylo–, (Eder, Kestemont, y Rybicki, 2013), está siendo utilizada por proyectos académicos que también trabajan en la edición de datos: Dracor (Fischer et al., 2018), Théâtre classique (Fièvre, 2007), Las Soledades (Rojas Castro, 2017) o Las siete partidas digitales (Fradejas, 2018).

La implementación más utilizada en la actualidad es la ya mencionada GitHub, una plataforma que permite crear perfiles personales y repertorios abiertos de manera gratuita. Disponen también de la posibilidad de solicitar cada año una cuenta como investigador, con la que se disponen de repositorios privados –tanto personales como de organización–. Los repositorios permiten que los datos se encuentren en el ordenador de cada investigador a la vez que en los servidores de GitHub. De esta manera, cada investigador puede realizar cambios de manera controlada (*commits*) y que el resto pueda descargarse (*pull*) las actualizaciones. Los diferentes cambios, versiones (*branches*) y publicaciones (*releases*) van creando una estructura arbórea histórica que puede ser rescatada en cualquiera de sus puntos por cualquier investigador. Si el proyecto no está cómodo con que nuestros datos se encuentren en los servidores de una empresa, podemos optar por otras implementaciones de Git que se encuentran en nuestros propios servidores, como GitLab. Estas herramientas ofrecen más posibilidades en cuanto a la publicación e integración con herramientas de archivo que veremos más adelante.

9. METADATOS Y ANOTACIÓN

Además de los textos, necesitaremos datos sobre los mismos que posibiliten su análisis (Burnard, 2004; NISO, 2004). Dependiendo del tipo de datos que sean o del nivel al que se refieran los llamaremos metadatos –datos sobre los datos– o anotación –lingüística o de otro tipo–.

Los metadatos especifican información normalmente de la unidad principal de texto. Por ejemplo, si estamos creando un corpus de obras de teatro, querremos anotar de cada obra de teatro quién es su autor, el título o su año de publicación. Esta información puede encontrarse en el texto (cubierta, paratextos legales) o puede ser que no se encuentre en él (porque le falten páginas, porque esa información no se encuentre); pero de manera independiente querremos tenerla a nuestra disposición. Otros datos que no se encuentra en el texto original pero que podremos querer en nuestras ediciones digitales son ciertas informaciones sobre unidades textuales menores ya sean lingüísticas (como oraciones, sintagmas o palabras) o de otro tipo (como versos). Por ejemplo, es posible que queramos anotar lingüísticamente la categoría gramatical de las palabras y utilizar la información en nuestros análisis. A este tipo de

información suele llamarse anotación, aunque en ciertos casos las fronteras entre anotación y metadatos es poco clara, como sucede en la especificación geográfica de cada escena o capítulo.

Los metadatos básicos que cada texto debería contener son título, autor y fecha de composición o de publicación. Aunque estos datos puedan parecer obvios, pensemos en casos como Santa Teresa o Vallé-Inclán, en los nombres más habituales poco tienen que ver con su nombre de nacimiento. Por eso es necesario anotar la información de título y autor mediante identificadores unívocos que instituciones terceras otorguen, como lo son VIAF (Virtual International Authority File) a nivel internacional, o la BNE para autores españoles. De hecho, VIAF aporta 167 formas que las diferentes bibliotecas alrededor del mundo han otorgado para Valle. Estos identificadores permiten comparar de manera más sencilla corpus y catálogos, además de conceder un primer punto de referencia para Linked Open Data y poder extraer de otras fuentes más información sobre textos y autores.

He mencionado el año de publicación o de redacción, aunque hay otras fechas de relevancia: fecha de inicio o fin de redacción, fecha de estreno, fecha de revisión por parte del autor, etc. Las secciones de metadatos del *teiHeader* estructuran la complejidad de facetas del texto: podemos trabajar con un texto escrito durante el siglo XVII, publicado por primera vez en XVIII (edición perdida), publicado de nuevo en el siglo XIX, haber sido digitalizado como PDF en el siglo XX y haber sido convertido a XML-TEI en el siglo XXI. Todas las fechas son relevantes en el camino genético del texto, y por lo tanto todas deberían codificarse correcta y específicamente. De no hacerse estaremos perdiendo información, como ocurre al ir a la ficha de la edición de El Quijote en el Cervantes Virtual[9]. Esta señala que la publicación utilizada para la digitalización fue publicada en 1911-1913, y que su digitalización tuvo lugar en 1999, pero no aparece que ese texto fue publicado por primera vez en 1605. De esta manera, si nos descargamos los textos y metadatos del Cervantes Virtual, *El Quijote* podría acabar siendo catalogada y analizada como una obra del siglo XX. En los corpus de Textbox, cada texto contiene cuatro fechas diferentes referenciadas de manera específica: la fecha de la primera publicación, la fecha de la publicación del ejemplar que fue digitalizado, la fecha de la digitalización y la fecha en la que el texto pasó a formar parte al corpus. El caso de BETTE (Santa María Fernández et al., 2017) también contiene la fecha de estreno.

En cuanto a la anotación, la más frecuente es la anotación gramatical (morfológica o sintáctica). Normalmente las herramientas de anotación lingüística requieren como formato de entrada texto plano, pero suelen producir diferentes formatos de exportación, como texto plano, tablas CSV o XML (sea TEI o no). La herramienta FreeLing[10] produce desde 2016 XML; se

[9] Accesible desde: https://bit.ly/2pgkTfZ.
[10] Accesible desde: https://bit.ly/2yKTNQ8.

puede acceder a la anterior versión por navegador mediante las APIS amigables del laboratorio IULA de la Univerisdad Pompeu Fabra. Freeling también ofrece otros tipos de anotaciones lingüísticas: anotación fonética o anotación semántica (mediante Wordnet). Finalmente, es posible anotar otros tipos de información: los corpus DISCO (Ruiz Fabo et al. 2017) y ADSO (Navarro-Colorado, 2015) han anotado información métrica sobre sus corpus poéticos; y el proyecto de los Observadores[11] (Ertler, 2013; Semlak, 2014) ha anotado información narratológica sobre las secciones de los textos.

10. LICENCIA, PUBLICACIÓN Y ARCHIVO

Finalmente debemos enfrentarnos a una serie de cuestiones en cuanto a la publicación y archivo de los datos. En primer lugar, deberemos aclarar hasta qué punto deben respetarse derechos de explotación del autor. En España los herederos retienen los derechos de explotación durante 80 años tras la muerte del autor, tras lo cual los textos pasan a estar en dominio público (Bercovitz Rodríguez-Cano, 2012, pp. 92-95): en 2017 los autores muertos en 1936 quedaron libres de derechos de explotación, pudiendo publicarse de manera libre.

La preparación y selección de los datos es una labor intelectual sobre la que podemos reclamar cierta propiedad intelectual. En las últimas décadas se han extendido las licencias Creative Commons, que de una manera sencilla permiten al creador de datos definir exactamente qué derechos entrega a terceros y cuáles restringe. El proyecto europeo de infraestructura para herramientas lingüísticas CLARIN, del que España todavía no es miembro, ha puesto a disposición una plataforma de información legal (Kamocki y Ketzal, 2017) con artículos introductorios y posts sobre aspectos legales dirigidos específicamente a humanistas.

La publicación de nuestros datos es importante por diferentes razones: revierte en la sociedad la inversión pública en investigación, posibilita la ciencia replicable, refuerza las Humanidades Digitales en español, pero también mejora las posibilidades de que nuestros datos se mantengan en el tiempo. La Universidad Stanford tiene un programa dirigido a las bibliotecas cuyo título tiene un mensaje claro y sencillo: LOCKSS, Lots of Copies Keep Stuff Safe o muchas copias mantienen las cosas seguras. Si decenas de usuarios han clonado o descargado un corpus, las posibilidades de que en unas décadas ese corpus siga de alguna manera accesible son mucho mayores que si el corpus se encontraba únicamente en el ordenador de un investigador. Hasta hace pocos años los proyectos disponían en la práctica de dos posibilidades de colocar sus datos online: un servidor universitario o un servidor privado contratado de manera personal. Ambas opciones tienen importantes carencias: cualquier modificación en el organigrama de informáticos de la universidad puede borrar nuestra carpeta y que nuestros datos desaparezcan. Por otro lado, un servidor contratado de manera personal requiere que

[11] Moralischen Wochenschriften en su título original en alemán.

cada año se pague los costes (usualmente entre 50 y 100 euros). El final del proyecto o de su financiación, o incluso el descuido de no pagar a tiempo, pueden hacer que los datos desaparezcan.

Ya hemos visto en secciones anteriores nuevas posibilidades como GitHub o GitLab, aunque estas herramientas no solucionan el problema de archivo a largo plazo. Por eso resultan necesarias iniciativas como Zenodo[12], un proyecto europeo apoyado por el CERN que se encarga de que los datos de investigación estén disponibles de manera abierta y archivados en el largo plazo. Su identificación se realiza mediante DOI y los datos pueden ser descargados de manera sencilla. A finales de 2017, la versión alemana del proyecto de infraestructura europeo para las Humanidades Digitales (DARIAH)[13], del que España tampoco es parte, ha lanzado un repositorio cuyo objetivo es archivar los datos de investigación de manera abierta y mantenida en el tiempo.

11. CONCLUSIONES

La cantidad de profesionales formados en Humanidades Digitales desde el nivel de grado sigue siendo reducida: la enorme mayoría de investigadores en HD comenzamos a interesarnos en estas metodologías al nivel de máster. Los artículos y conferencias tienden a mostrar exclusivamente resultados de análisis, ocultando (por falta de tiempo, espacio o interés) cientos de detalles sobre la preparación de los datos. En este artículo defiendo que la preparación y publicación de datos humanísticos en formato digital para su análisis debería ser asumida como tarea central por las Humanidades Digitales. He intentado señalar algunas de las preguntas claves que debemos hacernos al comenzar este camino: elección del análisis, decisión sobre el tipo de corpus, definición de población, obtención de digitalizaciones, modelización del texto, elección de flujo de trabajo, metadatos y anotación, y finalmente publicación. Este camino no es rápido ni fácil, pero ni debe ser un secreto ni es más complicado de lo que aquí se detalla. El recorrerlo nos permite ampliar nuestro horizonte de análisis: no debemos quedarnos en lo que otros ya hayan digitalizado, tenemos la libertad de analizar lo que nos interese.

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.

[12] Accesible desde: https://zenodo.org/.
[13] Digital Research Infrastructure for the Arts and Humanities. Accesible desde: https://www.dariah.eu/.

Bercovitz Rodríguez-Cano, A. (2012). *Manual de Propiedad Intelectual*. Manuales. Valencia: Tirant lo Blanch.

Blei, D. M., Andrew Y. N. y Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Burnard, L. (2004). Metadata for Corpus Work. En M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: AHDS Literature, Languages and Linguistics. Recuperado de https://bit.ly/2KhSUFk el 06/08/2019.

_____ (2014). *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. Marseille: OpenEdition Press. doi:10.4000/books.oep.

Calvo Tello, J. (2017). *Corpus of Spanish Short Stories from 1880-1940*. doi: 10.5281/zenodo.597430.

_____ (2018). *Corpus of Novels of the Spanish Silver Age*. Würzburg: Universidad de Würzburg. Recuperado de: https://github.com/cligs/textbox el 06/08/2019.

Calvo Tello, J. y Cerezo Soler, J. (2017). La Conquista de Jerusalén ¿de Cervantes? Análisis estilométrico sobre autoría en el teatro del siglo de oro español. *Digital Humanities Quarterly*, *10*. Recuperado de http://www.digitalhumanities.org/dhq/ el 06/08/2019.

Calvo Tello, J. y Santa María, T. (2017, October 13). GHEDI/BETTE: Versión 1.0 de BETTE. doi:10.5281/zenodo.1010140.

Chollet, F. (2018). *Deep Learning with Python*. New York: Manning.

Eder, M., Kestemont, M. y Rybicki, J. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, *16*(1), 1-15. Recuperado de https://bit.ly/2YtMc83 el 06/08/2019.

Ertler, K. D. (2013). *Moralischen Wochenschriften*. Graz: Universidad de Graz. Recuperado de https://bit.ly/31kDciA el 06/08/2019.

Evans, J. D. (1996). *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove: Brooks/Cole Pub. Co.

Fischer, F., Trilcke, P. y Orekhov, B. (2018). *Dracor*. Recuperado de https://dracor.org/ el 06/08/2019.

Fièvre, P. (2007). *ThéâTre Classique*. Recuperado de http://www.theatre-classique.fr el 06/08/2019.

Fradejas Rueda, J. M. (2018). *7PartidasDigital/XML-TEI*. doi: 10.5281/zenodo.1195641.

Garrish, M. (2011). *What Is EPUB 3?* Sebastopol: O'Reilly Media. Recuperado de https://oreil.ly/1W6OBf6 el 06/08/2019.

Gylling, M., Meester, B., Herman, I., Siegman, T., Cramer, D. y Rosenthol, L. (2017). Web Publications for the Open Web Platform: Vision And Technical Challenges. Recuperado de https://www.w3.org/TR/pwp/ el 06/08/2019.

Harold, E. R. y Scott Means, W. (2004). *XML in a Nutshell*. O'Reilly Media: O'Reilly. Recuperado de https://oreil.ly/2M4ngxo el 06/08/2019.

Haslwanter, T. (2016). *An Introduction to Statistics with Python. With Applications in the Life Sciences*. Cham: Springer International Publishing.

Henny-Krahmer, U. y Neuber, F. (2017). Criteria for Reviewing Digital Text Collections, Version 1.0. *A Review Journal for Digital Editions and Resources*, *6*. Recuperado de https://bit.ly/2KknOgA el 06/08/2019.

Jauralde Pou, P. (Dir.) (2013). *Clásicos Hispánicos. Madrid-Würzburg: More Than Books*. Recuperado de http://www.clasicoshispanicos.com/ el 06/08/2019.

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

Kamocki, P. y Ketzan, E. (2017). Legal Information Platform. *CLARIN ERIC*. Recuperado de https://www.clarin.eu/content/legal-information-platform el 06/08/2019.

Mikolov, T., Chen, K., Corrado, G. y Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [cs.CL]*. Recuperado de https://bit.ly/2T7Px6l el 06/08/2019.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Müller, A. C. y Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Beijing: O'Reilly.

National Information Standards Organization (NISO) (2004). *Understanding Metadata*. Bethesda: NISO. Recuperado de https://bit.ly/1yMOPJx el 06/08/2019.

Navarro-Colorado, B. (2015). A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects. En A. Feldman, A. Kazantseva, S. Szpakowicz y C. Koolen (Eds.), *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver. Recuperado de https://bit.ly/2OJ3d9N el 06/08/2019.

Percillier, M. (2017). Creating and Analyzing Literary Corpora. En S. Hai-Jew (Ed.), *Data Analytics in Digital Humanities. Multimedia Systems and Applications* (p. 91-118). Cham: Springer International Publishing. Recuperado de https://bit.ly/2M30iXq el 06/08/2019.

Reid, C. (2017, febrero 01). *W3C and IDPF Officially Combine Organizations*. [Entrada de blog]. Recuperado de https://bit.ly/2M3OMei el 06/08/2019.

Ruiz Fabo, P., Bermúdez Sabel, H., Martínez Cantón, C. y Calvo Tello, J. (2017). *DISCO. V2.0* [Data set]. doi:10.5281/zenodo.1069844.

Rojas Castro, A. (2017). La edición crítica digital y la codificación TEI. Preliminares para una nueva edición de las Soledades de Luis de Góngora. *Revista de Humanidades Digitales*, *1*, 4-19. doi:10.5944/rhd.vol.1.2017.16379.

Sánchez-Martínez, F., Martínez-Sempere, I., Ivars-Ribes, X. y Carrasco, R. C. (2013). An Open Diachronic Corpus of Historical Spanish. *Language Resources and Evaluation, 47*(4), 1327-42. doi:10.1007/s10579-013-9239-y.

Santa María Fernández, M. T., Jiménez Fernández, C. M., y Calvo Tello, J. (2017). *Biblioteca Electrónica Textual del Teatro Español, 1868-1936*. Recuperado de: https://zenodo.org/record/1010140#.XYuXJkZKjIU el 07/10/2019.

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities*, *2*(3), 2-13. Recuperado de https://bit.ly/191nMDf el 06/08/2019.

_____ (2017). Aufbau von Datensammlungen. En F. Jannidis, H. Kohle y M. Rehbein (Eds.), *Digital Humanities: eine Einführung* (pp. 223-33). Stuttgart: J.B. Metzler Verlag.

Schöch, C., Henny, U., Calvo Tello, J. y Popp, S. (2015). *The CLiGS Textbox*. Recuperado de https://github.com/cligs/textbox el 06/08/2019.

Semlak, M. (2014). Digitale Edition als Instrument für literaturwissenschaftliche Forschung. En C. Schöch y L. Schneider (Eds.) *Literaturwissenschaft im digitalen Medienwandel* (pp. 36-48). Recuperado de www.fu-berlin.de/phin/beiheft7/b7t02.pdf el 06/08/2019.