Pedro Mármol Ávila
Editor

Literatura, didáctica y humanidades digitales: aportaciones para la docencia y la investigación

Nicolás Asensio Jiménez
José Calvo Tello
Isabel María Gómez-Trigueros
Laura Hernández-Lorenzo
Rebeca Lázaro Niso
Noelia López Souto
Pedro Mármol Ávila
Mónica Martín Molares
Guadalupe Nieto Caballero
Gimena del Rio Riande
María Sánchez Cabrera
Ángela Torralba Ruberte

Dykinson, S.L.

Este libro ha sido sometido a evaluación por parte de nuestro Consejo Editorial
Para mayor información, véase www.dykinson.com/quienes_somos

Cada uno de los capítulos de este volumen
ha superado un proceso de revisión por pares

Editorial DYKINSON, S.L. Meléndez Valdés, 61 - 28015 Madrid
Teléfono (+ 34) 91 544 28 46 - (+ 34) 91 544 28 69
e-mail: info@dykinson.com
http://www.dykinson.es
http://www.dykinson.com

ISBN: 978-84-1170-479-3
Depósito legal: M-31165-2023
ISBN electrónico: 978-84-1170-635-3
DOI: 10.14679/2121

Preimpresión:
Besing Servicios Gráficos S.L.
besingsg@gmail.com

TEXT ENCODING INITIATIVE (TEI) COMO FORMATO PARA DATOS CUALITATIVOS A ESCALA CUANTITATIVA: EL CASO DE XML-TEI BIBLE

José Calvo Tello
Biblioteca Estatal y Universitaria de Göttingen

DOI: 10.14679/2124

1. TEI: un éxito de las Humanidades Digitales

La editorial de tecnología O'Reilly es una de las más influyentes en su área. Su nombre proviene de su fundador, Tim O'Reilly, uno de los impulsores de, por ejemplo, el concepto de *web 2.0*. La editorial O'Reilly publicó una introducción a XML (Harold y Means, 2004), uno de los formatos estándares más extendidos para el intercambio de información. En esta publicación, se encuentra un apartado titulado «Text Encoding Initiative», junto a otros estándares como DocBook, o los formatos de la familia OpenOffice.

No hay muchas iniciativas nacidas en las Humanidades Digitales (a partir de ahora HD) que hayan conseguido establecerse hasta el punto de que editoriales informáticas tengan que mencionarlos como relevantes. Por otro lado, no es extraño que la informática se beneficie de la experiencia y punto de vista de las Humanidades en ámbitos como la descripción de textos. Por supuesto, tampoco hay muchos proyectos de HD que hayan continuado desarrollándose durante varias décadas.

Text Encoding Initiative nació en los años 80 como una iniciativa motivada por investigadores en Humanidades y bibliotecarios (Burnard, 2014).

Desde ese momento se ha creado una asociación y una comunidad activa (con conferencia anual) que continúan desarrollando TEI, así como otras tecnologías y herramientas. Además de esto, la comunidad mantiene una revista de investigación, *Journal of TEI* (*jTEI*), en la que caben artículos de reflexión, presentación de nuevas propuestas o aplicaciones de TEI a conjuntos de datos concretos.

Pero ¿qué es exactamente TEI y para qué se usa? Cuando se habla de TEI como herramienta, se está haciendo referencia a una serie de directrices (o guía de referencia, en inglés *guidelines*) sobre cómo se pueden describir textos y sus metadatos. Tomemos un pasaje concreto para explicar esto, proveniente de la edición de *La Celestina* que llevó a cabo José Luis Canet (2017) para la colección Clásicos Hispánicos, que explicaré más adelante:

> SEMPRONIO.–
>
> (*Ap.*) No me engaño yo, que loco está este mi amo.

Cualquier edición de una obra de teatro distingue qué elementos textuales representan el nombre del personaje que habla, qué fragmentos son parte de una acotación y qué es lo que cada personaje dice. Normalmente esas distinciones están marcadas en la página impresa mediante diferencias tipográficas, como el uso de versalitas para el nombre del protagonista, la raya para el comienzo de lo que se dice y la cursiva en paréntesis para marcar acotaciones.

Esta marcación tipográfica (Hockey, 2000: 4) permite al lector diferenciar estos tipos textuales, pero entraña diferentes problemas. El primero es su limitación: si el editor quisiese además añadir información lingüística, aspectos de la génesis del texto o marcar información sobre las entidades (como personas o lugares mencionados), el abanico tipográfico se quedaría corto o resultaría tan complejo que la mayoría de los lectores no podrían decodificarlo. Además, estas normas son en muchos casos implícitas, por lo que requiere que un investigador confirme que en cada edición efectivamente la cursiva solo se ha utilizado para acotaciones y no para otros usos. Finalmente, aun teniendo el texto en formato digital, resulta dificultoso extraer fragmentos; por ejemplo, todos los nombres de los personajes y nada más.

Las guías de referencia de TEI explican cómo los investigadores pueden aplicar ciertos elementos para describir textos y sus metadatos. Estas guías se implementan en concreto en el formato XML. Este formato contiene únicamente una sintaxis sencilla en la que se prevén diferentes tipos, de los que los más importantes son texto, elementos, atributos, valores y comentarios. En teoría, cada proyecto debería especificar el vocabulario que puede utilizarse para los elementos y los atributos. Eso es principalmente lo que las guías de TEI ofrecen: una especificación de XML para la descripción de textos. El anterior ejemplo podría codificarse en TEI utilizando XML de la siguiente manera:

```
<sp who="#semp">
   <speaker>Sempronio</speaker>
   <p><stage rend="italic">(Ap.)</stage> No me engaño yo, que
loco está este mi amo.</p>
</sp>
```

Sin embargo, hipotéticamente TEI podría expresarse en otros formatos digitales como JSON, aunque en la prácticamente la totalidad de los proyectos trabajan con TEI en XML.

Hay numerosas introducciones y manuales para aprender sobre TEI. Se puede acceder a los materiales que la asociación de TEI pone a disposición o al contenido de *TEI by Example* (Terras, Vanhoutte y Van den Branden). Si se prefieren materiales en español, recomiendo el portal TTHUB (Allés-Torrent, Rio Riande y Calarco, 2019).

Uno de los aspectos más interesantes de TEI es el hecho de que permite e incluso fomenta que, en caso de que sea necesario, los proyectos puedan modificar los elementos, atributos o relaciones (Cummings, 2019). El lector puede pensar que esto va en contra de la idea de crear un formato que diferentes proyectos puedan utilizar y compartir. Sin embargo, la comunidad TEI acepta que nunca se podría ofrecer un vocabulario que incluyese todos los casos posibles. Si se intentase, la cantidad de elementos que poblarían las guías de referencias explotarían hasta los millones de elementos. Esto haría que TEI resultase prácticamente inutilizable. Por otro lado, si TEI no ofreciese la posibilidad de que los proyectos lo modificasen para sus características específicas, los usuarios se verían obligados a abusar de otros elementos que ya están contenidos en TEI, pero que tienen otras funciones.

Al margen de las modificaciones locales que los proyectos pueden realizar para sí mismos, TEI abre la posibilidad de que los investigadores propongan modificaciones para las guías en general. Más adelante explicaré mi experiencia sobre cómo el proyecto XML-TEI Bible modificó en primer lugar un atributo de manera local, atributo que posteriormente fue propuesto para que se añadiese a las guías de TEI. De esta manera, las guías de TEI se van actualizando: se modifican características de elementos ya existentes, se sustituyen algunos, o se añaden otros nuevos.

Uno de los argumentos más fuertes actualmente para utilizar TEI es su afinidad con los criterios FAIR (Wilkinson *et al.*, 2016). Estos proponen que los datos de investigación deberían cumplir una serie de criterios: que se puedan encontrar (*Findable*), sean accesibles (*Accesible*), interoperables (*Interoperable*) y reusables (*Reusable*). Esos cuatro criterios se desglosan a su vez en subpuntos más concretos y mesurables. Muchos de ellos prevén aspectos concretos sobre cómo los datos deben ser descritos detalladamen-

te mediante metadatos utilizando vocabularios controlados. Además, prevén que los formatos sean abiertos y conocidos en la comunidad de investigación. Sin embargo, algunos de los criterios FAIR caen en decisiones posteriores sobre la manera de realizar el archivado de los datos o su indexación en línea. No todos los proyectos que utilizan TEI satisfacen la misma cantidad de estos criterios: dependiendo de los elementos y atributos que hayamos utilizado y de qué manera los hemos rellenado, nuestros conjuntos de datos serán más o menos FAIR. De cualquier manera, seguir estos criterios resulta más sencillo si utilizamos TEI en comparación a si utilizamos otros formatos como bases de datos MySQL, nuestros propios esquemas de JSON o XML, formatos tabulares (CSV, TSV, Excel, etc.) o texto plano.

2. Luces y sombras sobre el uso de TEI en español

Algunos países cuentan con decenas de proyectos que han codificado grandes cantidades de textos de una manera relativamente sencilla o una cantidad más reducida de textos pero con mayor detalle en su descripción. Algunos ejemplos de esto son *TextGrid* (Horstmann, 2006) o *Deutsches Textarchiv* (Grötschel, 2007) para el alemán, *Théâtre classique* (Fièvre, 2007) para el francés o *Shakespeare Folger Library* (Mowat y Werstine, 2010) para el inglés, por mencionar algunos proyectos en algunas lenguas. Cualquier lector o investigador puede descargarse de estos proyectos de manera sencilla los textos, el etiquetado TEI y sus metadatos.

¿Y para el español? Durante muchos años el español fue una de las lenguas pujantes en la aplicación de TEI. Grandes proyectos se pusieron en marcha durante la década de 1990 con el objetivo de digitalizar y anotar en TEI miles de textos. Entre ellos hay que mencionar la Biblioteca Virtual Miguel de Cervantes (1999, a la que me referiré como Cervantes Virtual) o los proyectos de corpus CREA y CORDE de la Real Academia Española (Sánchez Sánchez y Domínguez Cintas, 2007). Además, aunque no exactamente en TEI, el proyecto TESO (Simón Palmer, 1997) codificó miles de textos del teatro del Siglo de Oro en un formato muy similar. Sin embargo, este comienzo dorado se vio truncado por las decisiones sobre cómo ofrecer los datos *online*. En lugar de mostrar y ofrecer la codificación original en TEI, cada archivo en XML se convirtió en otro archivo web en formato HTML. En esta transformación se perdían muchas de las ventajas del uso de TEI, al eliminarse prácticamente todos los metadatos y en buena medida el vocabulario controlado. En muchos casos incluso se producía HTML que no estaba bien formado, lo que impedía que muchas herramientas afines a XML como XSLT o Xquery (Allés-Torrent, 2015) pudiesen procesar los datos. Los proyectos habían tejido codificaciones con los mejores

materiales, pero a los usuarios se les daba versiones degradadas. La Biblioteca Cervantes Virtual incluso troceaba los textos en diferentes páginas, lo que hacía más complicado al usuario su lectura y al investigador acceder al texto completo.

Esta práctica puede ser resumida en el lema «edito TEI pero publico HTML». En mi opinión, esto lastró a la comunidad en español de diferentes maneras. En primer lugar, el esfuerzo de codificación en TEI se perdió, desperdiciando los presupuestos que se habían invertido en ello. En segundo lugar, la comunidad en español no dispuso durante varias décadas de ejemplos de TEI. En tercer lugar, otros proyectos (grandes, medianos y pequeños) continuaron la senda ya trazada del «edito TEI pero publico HTML», asumiendo de manera más o menos consciente las malas prácticas que estos grandes proyectos habían implementado. En cuarto lugar, esto produjo una falta de datos para el español en formatos aceptables. Para cuando las HD se quisieron centrar en la aplicación de nuevas metodologías como estilometría y aprendizaje automático (técnicas como clasificación, reducción de dimensionalidad como *topic modeling* o PCA, clusterización), el español tenía un déficit en comparación con otras lenguas: no había datos en formatos aceptables como TEI. Esta automutilización de la comunidad HD en español explica que muchos trabajos en los que se analiza una metodología en varios idiomas (como los trabajos de Rybicki y Eder, 2011, o Evert *et al.*, 2017) no pudiesen utilizar datos en español, generando más perjuicios décadas después.

En buena medida motivado por esta falta de datos, a partir de mediados de 2010 la situación cambia. Proyectos como DISCO (Ruiz Fabo, Martínez Cantón y Calvo Tello, 2017) o ADSO (Navarro-Colorado, 2015) pusieron a disposición miles de sonetos en formato TEI. Estos proyectos han continuado analizando y anotando posteriormente sus conjuntos de datos (Navarro-Colorado, 2018; Ruiz Fabo *et al.*, 2018).

En el marco del proyecto de CLiGS de la Universidad de Würzburg, en el que participé, se publicaron diferentes corpus y conjuntos de datos en formato TEI. Entre ellos cabe mencionar el caso de Textbox, un conjunto de corpus en diferentes lenguas romances: francés, español, italiano y portugués (Schöch *et al.*, 2019). El principal objetivo del proyecto era el análisis de géneros literarios en diferentes épocas mediante técnicas de clasificación, aunque también se realizaron análisis contrastivos entre novelas de España y Latinoamérica (Calvo Tello, Henny-Krahmer y Schöch, 2018), clasificación (Calvo Tello, 2018), *topic modeling* (Schöch *et al.*, 2016), *sentiment analysis* (Henny-Krahmer, 2018) o grafos (Calvo Tello, 2020). En el caso del español se encuentra una serie de corpus: novelas latinoamericanas, colecciones de cuentos y novelas españolas. Este último es una sección del corpus CoNSSA, que es la base de los análisis de mi tesis (Calvo Tello, 2021). Un conjunto de datos originado en este proyecto que se diferencia de los corpus hasta

ahora mencionados es Bib-ACMé, una bibliografía de novelas publicadas en México, Cuba y Argentina (Henny-Krahmer, 2017).

La publicación de datos, especialmente en TEI, puede generar efectos sinérgicos. Por ejemplo, junto a investigadoras de la UNIR hemos publicado un corpus de teatro de la Edad de Plata (Jiménez Fernández *et al.*, 2017), que fue analizado de diferentes maneras (Santa María, Calvo Tello y Jiménez Fernández, 2020; Jiménez Fernández y Calvo Tello, 2020). Tras estos primeros pasos, el proyecto DraCor mostró interés en integrar el corpus dentro de su plataforma, que ofrece acceso a una docena de corpus de teatro en diferentes lenguas europeas (Fischer *et al.*, 2018). La integración de los textos resultó sencilla debido a su codificación en TEI. Otro proyecto a nivel europeo iniciado en los últimos años es ELTeC (Odebrecht *et al.*, 2019), que ofrecerá cien novelas por cada lengua, entre ellas el español.

Hasta ahora, muchos de los proyectos mencionados tienen una tendencia cuantitativa que ha ganado peso en las HD en los últimos años. Sin embargo, es necesario recalcar que proyectos con objetivos más cualitativos pueden realizarse cómodamente en TEI. A nivel internacional, probablemente la mayor comunidad de uso de TEI se encuentra entre las actividades de edición filológica de textos históricos y literarios. Una iniciativa que centra buena parte de esta actividad es *RIDE*, una revista de investigación que exclusivamente publica reseñas, tradicionalmente de ediciones críticas digitales, aunque en los últimos años se han abierto a otros objetos como corpus y herramientas. Lamentablemente, solo hay un puñado de proyectos en español de ediciones filológicas utilizando TEI y que hayan publicado sus datos. Algunos ejemplos son la edición de las *Soledades* de Góngora (Rojas Castro, 2016), la edición de *La dama boba* de Lope de Vega (Presotto *et al.*, 2015), el corpus de Poesía Medieval dirigido por Rio Riande (2020) y las *Siete Partidas* (Fradejas Rueda, 2018).

Un proyecto similar, aunque con algunas particularidades, es la colección Clásicos Hispánicos, que fue iniciada en la Universidad Autónoma de Madrid por Pablo Jauralde Pou y algunos estudiantes y colaboradores de su grupo de investigación, entre los que me encontraba (Jauralde Pou, 2013). El objetivo de esta colección fue y sigue siendo poner a disposición de los lectores ediciones filológicas de calidad de los clásicos de la literatura en formato digital. Para ello, se creó un flujo de trabajo en el que los especialistas podían trabajar en la herramienta que ellos conocían: editores de texto como Word. La conversión del texto generaba finalmente ediciones del texto en los formatos de libros electrónicos: ePUB y mobi. El formato ePUB tiene una serie de ventajas frente al formato PDF: utilización de tecnologías web (CSS, HTML, XML), superación del modelo de libro impreso como referente, apertura de código, validación, etc. Poco después de sus comienzos, el proyecto asumió TEI como su princi-

pal formato. Desde su última fase, el proyecto ofrece diferentes formatos (TEI, ePUB, mobi) de casi cien obras publicadas de manera abierta y gratuita a través del portal Zenodo (<https://zenodo.org/communities/clasicos_hispanicos/> [fecha de consulta: 28-3-2022]). Entre los textos publicados cabe mencionar el *Quijote*, *El Buscón*, *La Celestina*, *Libro de los gorriones*, etc.

3. Un caso concreto: XML-TEI Bible

3.1. La Biblia como conjunto de datos para las Humanidades Digitales

Me gustaría en esta sección prestar especial atención a otro proyecto en el que he estado trabajando desde 2015. Se trata de XML-TEI Bible, un proyecto de anotación en TEI que toma como base una edición de la traducción al español de Reina y Valera. Independientemente del interés que la Biblia pueda despertar, considero que ofrece a las HD una serie de ventajas en comparación con otros textos o géneros.

Las HD es una categoría que se propone durante los años 2000 como un hiperónimo que aúna investigadores de diferentes ramas de las Humanidades que se interesan por nuevos formatos y metodologías computacionales. Sin embargo, esto causa que el intercambio sobre el objeto de estudio suele ser limitado entre investigadores de HD. Disciplinas como Teología, Filosofía, Historia, Historia del Arte, Lingüística o Literatura comparten pocos objetos de estudio. Sin embargo, la Biblia despierta interés en sectores concretos de la mayoría de disciplinas humanísticas, ya sea por temas y escenas (Literatura e Historia del Arte), su influencia (Teología, Filosofía, Historia) o sus traducciones (Lingüística).

Además de su carácter interdisciplinar, las HD tienen un fuerte componente internacional y multilingüístico. Se observa al visitar secciones de conferencias sobre metodologías aplicadas a la literatura, en las que cada ponencia trata una lengua diferente. El enriquecimiento que esto conlleva no oculta la dificultad en la comunicación: nuestros conocimientos de las literaturas haitianas, canadienses, holandesas, rusas, paquistaníes, japonesas o filipinas, por mencionar algunos casos, son tremendamente limitados. De la misma manera, cabe esperar que otros investigadores solo puedan nombrar a Cervantes y Lorca entre los literatos españoles. Esto dificulta el intercambio de conocimiento y la discusión: los ponentes hablan de autores y textos que el público desconoce.

Incluso acarrea problemas metodológicos: si una herramienta se aplica solo a textos en una lengua, sus resultados no pueden extrapolarse a otras literaturas. Por ejemplo, si un algoritmo fuese capaz de anotar automáticamen-

te la ironía de manera correcta y este se hubiese desarrollado y evaluado con textos literarios en francés, no se sabría si solo funciona para textos literarios en francés, si es extrapolable (y, si lo es, hasta qué punto) y a qué lenguas.

Las HD se beneficiarían de un objeto de estudio que sea conocido por la mayoría de los investigadores y que pueda ser representado en diferentes lenguas. De esta manera, los investigadores podrían discutir casos concretos de los datos, y se podrían aplicar los algoritmos a las diferentes versiones lingüísticas. Eso es precisamente lo que la Biblia ofrece. La mayoría de investigadores tienen un conocimiento básico sobre las personas más mencionadas de la Biblia (Abraham, José, Moisés, David, Jesús, Pablo, por citar algunos). Al menos estos conocimientos son mayores que los que poseemos del resto de literaturas del mundo. Además, la Biblia es uno de los libros más traducidos de la historia. De hecho, muchas lenguas solo disponen de este texto. La Biblia cuenta además con un sistema de identificación de sus secciones mediante libros (Génesis, Salmos, Evangelio de Mateo), capítulo y versículos. Este sistema de identificación permite que un identificador como «GEN.001.001» pueda ser fácilmente interpretado como el primero versículo del primer capítulo del libro de Génesis, y esto *mapearse* en cualquier traducción en cualquier lengua.

Otra de las ventajas de la Biblia como objeto cultural para las Humanidades es que rompe con los actuales sesgos de los datos. El inglés hoy en día aglutina la mayoría de los recursos digitales, tanto en términos de datos como en términos de herramientas. Esto no solo dificulta que otras lenguas estén presentes en la investigación, sino que también sesga las herramientas que se desarrollan a favor de las características lingüísticas del inglés o de sus conceptos culturales (Gutiérrez de la Torre, 2020). Una de las tendencias actuales de las HD es llamar la atención sobre este sesgo angloparlante y reaccionar frente él. La Biblia no fue escrita en inglés sino principalmente en hebreo y en griego. Los lugares representados cubren principalmente el área entre el nordeste de África (actual Egipto), sureste de Europa (actual Grecia) y oeste de Asia (principalmente, las actuales Israel, Siria, Turquía, Líbano e Irak). Diferentes lenguas y países en todos los continentes se han visto influidos durante muchos siglos por la Biblia, creándose instituciones lingüísticas como las traducciones de Lutero en Alemania, la King James en el Reino Unido o la Vulgata para el latín.

Por último, la Biblia representa un punto intermedio en cuanto a la cantidad de los datos. Más que un libro, la Biblia como tal es una colección de libros, por lo que bien se le podría designar como corpus. Su extensión total depende de la traducción que se siga: el canon judío (seguido también por las iglesias protestantes) suele contener 39 libros en el Antiguo Testamento; el canon católico le añade siete libros más. A estos se les suman los 27 libros del Nuevo Testamento. En total, contiene más de 31 000 versículos y alrede-

dor de un millón y medio de *tokens*. Esto la hace bastante más larga que una novela muy amplia, y tiene un tamaño comparable a muchos otros corpus de investigación. De este modo, la Biblia supone un punto intermedio entre un corpus y un libro de notable extensión. La cantidad de textos es suficientemente grande como para realizar análisis estadísticos, pero su extensión limitada permite también que una persona lo lea por completo. Así, el investigador puede combinar lecturas distantes y cercanas sobre el mismo objeto de estudio. Diferentes trabajos procedentes de la lingüística ya han utilizado la Biblia como objeto de estudio (Resnik, Olsen y Diab, 1999; Hatav, 2000; Grandjean, 2013; Christodouloupoulos y Steedman, 2015).

3.2. Anotación en XML-TEI Bible

La anotación actual del proyecto XML-TEI Bible contiene una serie de fenómenos que iré presentando. Todos los datos, visualizaciones y programas que aquí menciono están disponibles en el repositorio GitHub (<https://github.com/morethanbooks/XML-TEI-Bible> [fecha de consulta: 6-2-2021]). Además, diferentes versiones han sido archivadas en Zenodo (DOI: <https://doi.org/10.5281/zenodo.3873336>) y en DARIAH Repository (DOI: <http://dx.doi.org/10.20375/0000-000C-D90A-5>). Texto, metadatos y anotación se encuentran en un único archivo XML que contiene todos los metadatos generales del proyecto, así como todos los libros. Dentro de este, cada libro está contenido en un elemento TEI. Así, cada libro tiene sus propios metadatos asociados, seguidos del texto y su anotación. A su vez, cada libro contiene capítulos, codificados en elementos *div*, los cuales agrupan una o más perícopas, también como elementos *div*. Los versículos están anotados mediante el elemento *ab*, asociados con identificadores en sus atributos. En el siguiente ejemplo se pueden observar los versículos:

```
<div type="pericope">
   <head type="pericope">Nacimiento de Jesucristo</head>
   <ab xml:id="b.MAT.001.018" type="verse" n="18">El nacimiento de
Jesucristo fue Así: Su madre María estaba desposada con José; y
antes de que se unieran, se Halló que ella Había concebido del
Espíritu Santo.</ab>
   <ab xml:id="b.MAT.001.019" type="verse" n="19">José, su marido,
como era justo y no Quería difamarla, se propuso dejarla
secretamente.</ab>
   <ab xml:id="b.MAT.001.020" type="verse" n="20">Mientras él pensaba
en esto, he Aquí un ángel del Señor se le Apareció en sueños y
le dijo: José, hijo de David, no temas recibir a María tu mujer,
porque lo que ha sido engendrado en ella es del Espíritu Santo.</
ab>
```

```
   <ab xml:id="b.MAT.001.021" type="verse" n="21">Ella Dará a luz un
hijo; y Llamarás su nombre Jesús, porque él Salvará a su pueblo
de sus pecados.</ab>
</div>
```

Extrayendo la cantidad de elementos *ab* del tipo «verse» que hay en cada libro, se puede visualizar la cantidad de versículos por cada libro. El siguiente gráfico muestra cada libro de la Biblia con los nombres en inglés codificados con tres caracteres:

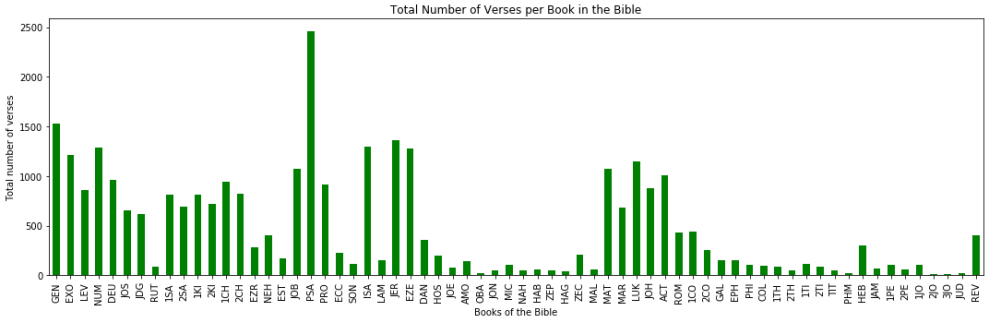

Figura 1. Diagrama de barras con la cantidad de versículos por libro

En una primera fase del proyecto, anoté de manera manual dos fenómenos: la identificación de entidades y la comunicación directa. Mientras leía el texto, fui marcando en TEI (mediante el elemento *rs*) todos aquellos sintagmas nominales (ya estuviesen compuestos de nombres comunes o nombres propios) que refiriesen una persona, un lugar o un grupo. En un siguiente paso, se marca cada persona a la que se le asignaba un identificador con un número creciente. Cualquier referencia a Jesús está marcada con *#per1*, independientemente de si en el texto aparecen las palabras *Jesús*, *Jesucristo*, *hijo de Dios*, *niño* o *maestro* (por dar algunos ejemplos). Lo mismo se realizó con lugares (ciudades, regiones, países, ríos, montañas y lugares celestiales) y grupos, principalmente aquellos con un carácter étnico (hebreos, filisteos, etc.), profesionales (soldados, siervos del faraón, etc.) o religiosos (fariseos, sacerdotes, etc.). Así, los anteriores versículos quedarían marcados de la siguiente manera:

```
<div type="pericope">
   <head type="pericope">Nacimiento de Jesucristo</head>
   <ab xml:id="b.MAT.001.018" type="verse" n="18">El nacimiento de
<rs key="#per1">Jesucristo</rs> fue Así: Su <rs key="#per11">madre</
rs> <rs key="#per11">María</rs> estaba desposada con <rs
key="#per12">José</rs>; y antes de que se unieran, se Halló que
ella Había concebido del <rs key="#per17">Espíritu Santo</rs>.</ab>
```

```
    <ab xml:id="b.MAT.001.019" type="verse" n="19"><rs key="#per12">José</
rs>, su <rs key="#per12">marido</rs>, como era justo y no Quería
difamarla, se propuso dejarla secretamente.</ab>
    <ab xml:id="b.MAT.001.020" type="verse" n="20">Mientras él pensaba
en esto, he Aquí un <rs key="#org4">ángel del <rs key="#per14">Señor</
rs></rs> se le Apareció en sueños y le dijo: <rs key="#per12">José</
rs>, <rs key="#per12">hijo de <rs key="#per35">David</rs></
rs>, no temas recibir a <rs key="#per11">María</rs> tu <rs
key="#per11">mujer</rs>, porque lo que ha sido engendrado en ella
es del <rs key="#per17">Espíritu Santo</rs>.</ab>
    <ab xml:id="b.MAT.001.021" type="verse" n="21">Ella Dará
a luz un <rs key="#per1">hijo</rs>; y Llamarás su nombre <rs
key="#per1">Jesús</rs>, porque él Salvará a su pueblo de sus
pecados.</ab>
</div>
```

Una vez que se han anotado las entidades y se extraen sus frecuencias por cada versículo, se puede observar su frecuencia relativa por la cantidad de versículos en cada libro:

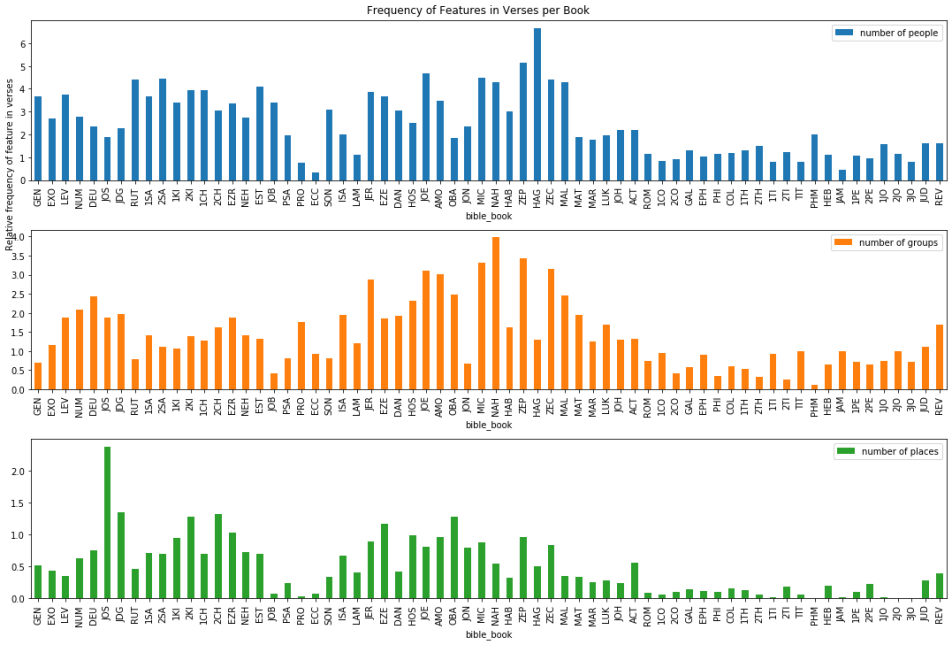

Figura 2. Frecuencia relativa de número de personas, grupos y lugares por libro

Esto permite, por un lado, comparar un fenómeno concreto en diferentes libros de la Biblia. Por ejemplo, la frecuencia de menciones a lugares (en verde) en el libro de Josué (JOS) es mucho más alta que en cualquier otro libro.

Utilizando el ejemplo de Mateo, observamos que en comparación con el resto de evangelios, la frecuencia de grupos (en naranja) es notablemente mayor. Por otro lado, se pueden observar tendencias generales de uno o a varios rasgos a lo largo de toda la Biblia. Por ejemplo, la frecuencia de aparición de estas entidades es en general inferior en las epístolas (últimos libros de la Biblia, desde ROM, Romanos, exceptuando el último libro, REV, Apocalipsis) que en la mayoría de libros.

El siguiente fenómeno que se anotó en la primera fase del proyecto fue la comunicación directa. En concreto se anotó quién comunica con quién y de qué manera. Para marcar la persona que comunica, utilizo el atributo *@who* dentro del elemento *q* (*quote*). Sin embargo, cuando comencé el proyecto, TEI no ofrecía ninguna posibilidad unívoca de señalar quién es el receptor de la información. Por eso propuse que se aceptase un atributo *@toWhom* en el que quedase recogido quién es el principal receptor de esa información. Tras un proceso de discusión, este atributo fue aceptado y ahora es parte de las guías generales de TEI, facilitando que otros proyectos lo utilicen. Por último, se anotó también si la comunicación entre dos personajes es por medio oral o escrito, además de otras modalidades mucho menos frecuentes (por sueños, juramento, canto, etc.). Dos de los versículos anteriormente vistos recogen cómo un ángel se comunica con José mediante sueños:

```
   <ab xml:id="b.MAT.001.020" type="verse" n="20">Mientras él
pensaba en esto, he Aquí un <rs key="#org4">ángel del <rs
key="#per14">Señor</rs></rs> se le Apareció en sueños y le dijo: <q
who="#org4" type="dream" toWhom="#per12"><rs key="#per12">José</rs>,
<rs key="#per12">hijo de <rs key="#per35">David</rs></rs>, no temas
recibir a <rs key="#per11">María</rs> tu <rs key="#per11">mujer</
rs>, porque lo que ha sido engendrado en ella es del <rs
key="#per17">Espíritu Santo</rs>.</q></ab>
   <ab xml:id="b.MAT.001.021" type="verse" n="21"><q who="#org4"
type="dream" toWhom="#per12">Ella Dará a luz un <rs key="#per1">hijo</
rs>; y Llamarás su nombre <rs key="#per1">Jesús</rs>, porque él
Salvará a su pueblo de sus pecados.</q></ab>
```

Si se realiza una extracción similar a la realizada con las entidades, se puede observar por ejemplo la cantidad relativa de comunicación directa. Además, esta se puede desglosar con la información de si la comunicación se encuentra a su vez dentro de otro proceso de comunicación (alguien dice que alguien le dijo algo...). Esta comunicación directa anidada es muy frecuente en la Biblia, especialmente en los libros proféticos donde el profeta cuenta que Dios le dijo que él debía decir a Israel que Dios les decía algo concreto. En algunos libros llega hasta cinco niveles de anidación, como muestra el siguiente gráfico:

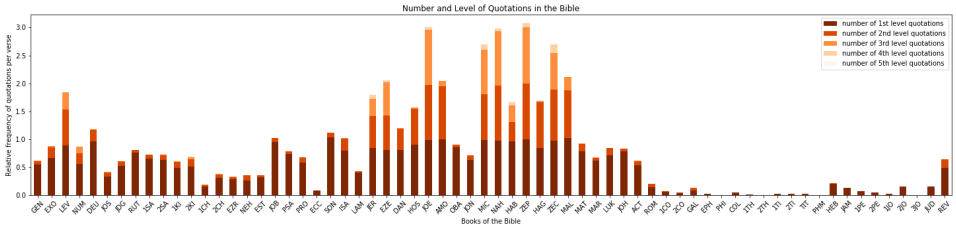

Figura 3. Frecuencia comunicación y nivel de anidación por libro

Además de los gráficos de barras hasta ahora mostrados, esta información de comunicación se puede modelar y visualizar como grafos, utilizando un libro concreto o la Biblia al completo. La mayoría de los trabajos que han aplicado grafos para representar interacción entre personajes se basan en la coaparición de estos en ciertas unidades textuales (Grandjean, 2013; Hettinger *et al.*, 2015; Trilcke *et al.*, 2016; Isasi, 2017; Santa María, Calvo Tello y Jiménez Fernández, 2020). Estos trabajos han recibido en los últimos años fuertes críticas que ponen en duda la validez de los datos de coaparición (Krautter *et al.*, 2018; Santa María, Calvo Tello y Jiménez Fernández, 2020; Jiménez Fernández y Calvo Tello, 2020). Por el contrario, la siguiente visualización muestra la comunicación entre personajes, que ha sido objeto de una anotación cuidada que se ha realizado manualmente. Así, se obtiene de manera resumida qué persona o grupo se comunica con quién basándonos en datos cualitativos. Junto a la información de la comunicación, se pueden añadir a la ilustración características de las entidades que la forman. Por ejemplo, en la siguiente ilustración se muestra la comunicación en el Evangelio de Mateo, con los hombres como nodos verdes, las mujeres de color lila, los grupos de color amarillo, y Jesús y Dios en rosa:

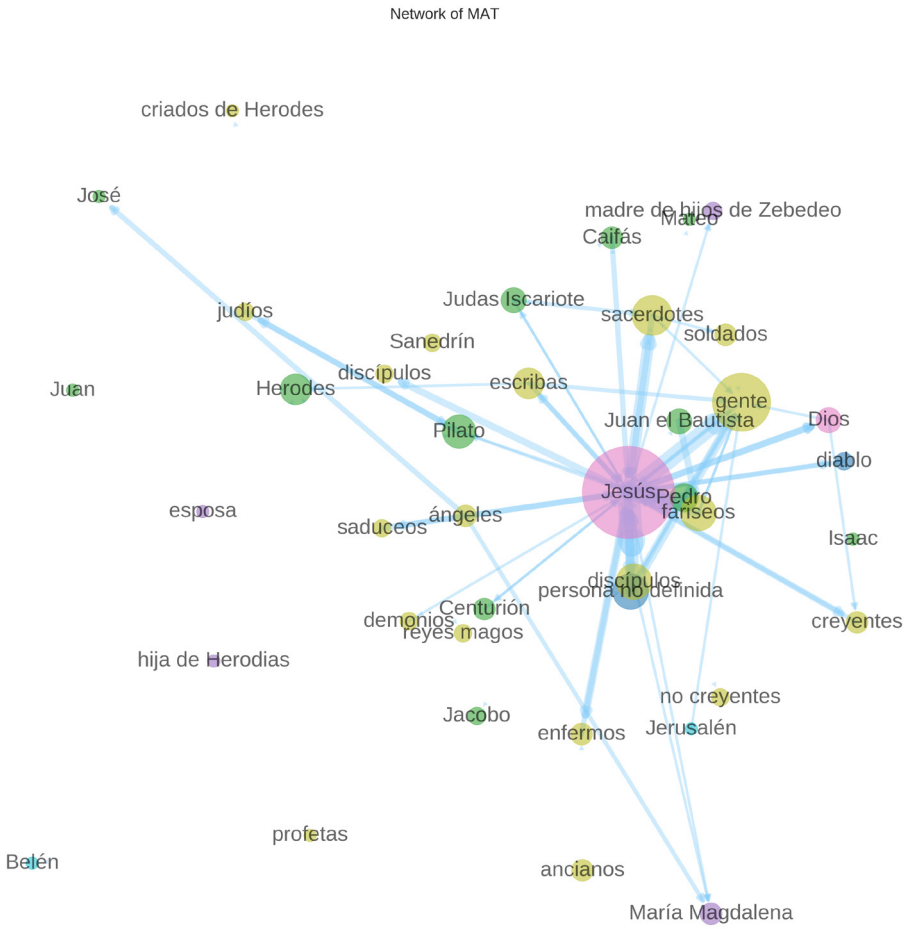

Figura 4. Grafo del libro de Mateo mostrando la comunicación entre entidades

Como se puede ver, el grupo de ángeles (en el medio, a la izquierda de Jesús) se encuentra conectado con José (arriba a la izquierda), debido a la comunicación que tiene lugar entre ellos en los versículos anteriormente mostrados.

Esta anotación de entidades y comunicación es suficiente para entregar algoritmos de clasificación que intentarían aprender qué rasgos lingüísticos tienden a aparecer en asociación con estos fenómenos. Este algoritmo podría aplicarse posteriormente a otros textos similares en español. Pero su potencial no se termina aquí. Los identificadores de las entidades, unidos a los identificadores de los versículos, permiten al investigador abstraer los

fenómenos de la superficie lingüística que fue anotada e incluso sobreponerla a otras lenguas. Si *#per1* es el valor del atributo *@who* en el versículo MAT.001.005, el investigador puede expresar esto sin necesidad de un idioma concreto. Así, lo aprendido en la Biblia en español en una traducción relativamente moderna, podría utilizarse para entrenar algoritmos en cualquier otra lengua traducida en cualquier época. Estos algoritmos podrían utilizarse posteriormente sobre otros tipos de textos. Es decir, la anotación de XML-TEI Bible permite potencialmente entrenar algoritmos que localicen personas en textos medievales, o en idiomas que tengan pocos recursos lingüísticos. El único requisito necesario es disponer de una traducción de la Biblia y que sus versículos estén mapeados con los de la versión utilizada aquí.

En una nueva y reciente fase del proyecto, he comenzado a anotar menciones a temas sexuales. El objetivo es tratar de tener una base empírica de qué temas sexuales son tratados en la Biblia y de qué manera. A comienzos de 2021 he anotado de esta manera ocho libros de la Biblia. Para ello he desarrollado una taxonomía de temas, también en TEI (para ver una discusión sobre taxonomías de temas sexuales, consúltese Pizarro Pedraza, 2013). Hasta ahora contiene 260 temas relacionados de alguna manera con el sexo, incluyendo relaciones, actos, estados, partes del cuerpo, objetos, lugares y grupos. Se entiende la sexualidad como una categoría en términos prototípicos. Es decir, algunos actos, temas o partes están más claramente relacionadas con el sexo que otros. Un acto sexual con penetración vaginal es un hecho claramente más sexual que un beso en la boca, y este es más sexual que, por ejemplo, lanzar una piedra a un lago. De manera similar, pene o vagina son órganos claramente sexuales; los pechos o los muslos pueden serlo o no, pero resultan más sexuales que, por ejemplo, los codos. Este grado de sexualidad ha sido formalmente marcado en la taxonomía mediante tres valores ordinales: explícitamente sexual, implícitamente sexual y no sexual. Este último valor cubre aspectos que no tienen *per se* un cariz sexual, pero que, por cultura o tradiciones literarias, se suelen asociar a temas sentimentales o eróticos, como la comparación de los amantes con gacelas o cabras, o las viñas y la noche como marco de encuentros eróticos. Esto me permite como investigador un mayor grado de flexibilidad sin que me obligue a decir, por ejemplo, que un beso siempre es parte de la categoría sexual.

Cada versículo puede contener diferentes referencias a temas sexuales en la Biblia. Esto lo he formalizado como elementos *spanGrp* dentro de un bloque *standOff*, detrás del elemento *text*, pendiendo del elemento raíz *TEI*. Como ejemplo, podemos observar que he anotado que en el versículo 20 del capítulo 1 de Mateo se menciona un embarazo:

```
<spanGrp type="theme" inst="#b.MAT.001.020">
    <span ana="pregnancy">
        <certainty match="@ana" locus="value" cert="high" given="text"/>
    </span>
</spanGrp>
```

El objetivo de esta anotación es obtener datos exactos sobre qué tema relacionado con el sexo se encuentra en qué versículo de la Biblia. Esto, por un lado, permite sencillas cuantificaciones como observar cuáles son los temas más frecuentemente anotados en, por ejemplo, el Evangelio de Mateo.

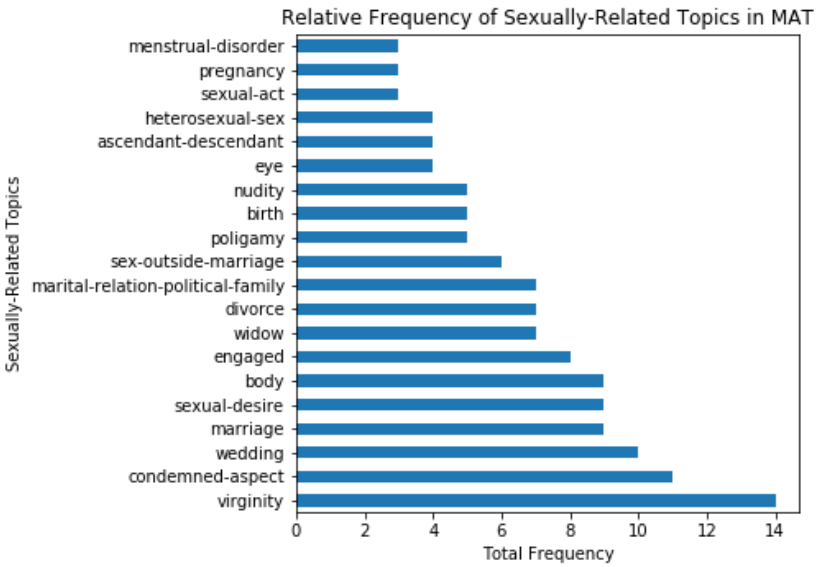

Figura 5. Frecuencia de temas sexuales en el libro de Mateo

Este gráfico muestra que en este libro bíblico el tema de la virginidad es el más frecuente, seguido por aspectos condenatorios, bodas, matrimonio, deseo sexual y el cuerpo. Sin embargo, esta información resulta insuficiente para saber quién trata esos temas o de qué manera. Sería interesante observar, por ejemplo, qué personas o grupos tienden a lanzar mensajes condenatorios, o con qué personas o grupo se relacionan los temas de la virginidad o el deseo sexual, por ejemplo. Los diferentes niveles de anotación explicados hasta ahora pueden ponerse en relación y calcularse qué entidades coaparecen en los mismos versículos junto a los temas sexuales. En el siguiente gráfico se muestra una tabla como mapa de calor (*heat map*). Las entidades del eje horizontal son las más frecuentes en el libro de Mateo, a las que se ha añadido a María, José y el grupo de los ángeles que aparecen en los versículos anteriormente mencionados. Los temas

sexuales del eje vertical son las 15 categorías más frecuentemente anotadas en este libro. Se muestra una columna y una fila de sumatorio de esta selección para tener información de conjunto de ambos tipos de anotaciones:

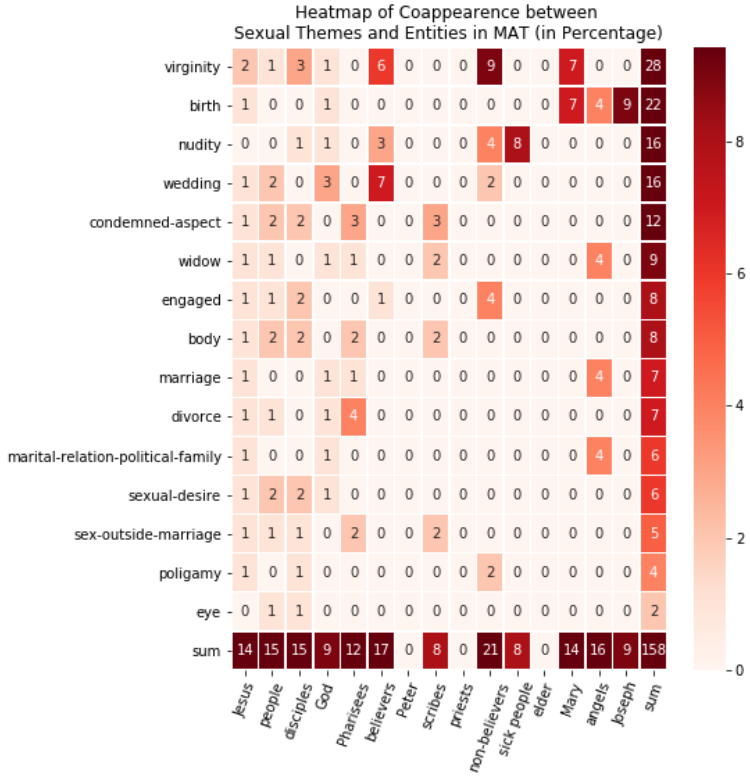

Figura 6. Mapa de calor con la coaparición entre temas sexuales y entidades en el libro de Mateo

El gráfico muestra claramente que una buena cantidad de referencias sobre la virginidad están relacionadas con María, para quien el tema del parto es dominante. La desnudez está principalmente relacionada con grupos sociales de estratos socioeconómicos bajos, por lo que se puede extraer que no es una desnudez erótica, sino falta de ropa causada por bajos recursos económicos. Los aspectos condenatorios no coaparecen con Jesús o Dios, sino más bien con grupos religiosos, principalmente con fariseos y escribas. También es interesante observar que no hay ningún tema sexual que predomine cuando Jesús o Dios aparecen mencionados en los versículos, a diferencia de otras entidades.

Como se aprecia en el ejemplo de la anotación de temas sexuales en TEI, he añadido explícitamente la certeza de la anotación. En el ejemplo anterior,

esta aparece como alta (*cert="high"*). Este valor puede ser medio o bajo en aquellos casos donde la interpretación no sea tan clara, o incluso en aquellos puntos en los que se sabe que diferentes tradiciones suelen interpretarlo de manera diferente. Por ejemplo, las iglesias protestantes difieren de la doctrina católica en cuanto a la vida sexual de María y José tras la concepción de Jesús. Mientras que la iglesia católica defiende que María nunca tuvo relaciones sexuales, las iglesias protestantes asumen que, como cualquier otra pareja, sí tuvieron relaciones sexuales y que además tuvieron hijos. Esta diferencia tiene repercusiones concretas al anotar textos como el capítulo 12 de Mateo, en el que aparecen mencionados los «hermanos de Jesús». En el caso de los versículos del primer capítulo de Mateo anteriormente visto, hay diferentes interpretaciones posibles sobre aspectos sexuales. En concreto, en el versículo 18 del capítulo 1 de Mateo, el verbo *unieran* se puede interpretar de dos maneras diferentes: como unión sexual o como unión en matrimonio. Estas dos posibles interpretaciones se han marcado explícitamente de la siguiente manera:

```
<spanGrp type="theme" inst="#b.MAT.001.018">
   <span ana="sexual-act">
      <certainty    match="@ana"    locus="value"    cert="medium"
given="protestant"/>
   </span>
   <span ana="wedding">
      <certainty    match="@ana"    locus="value"    cert="medium"
given="catholic" />
   </span>
</spanGrp>
```

Como se observa, la anotación recoge que la certeza es solo media en ambos casos, y que depende de la tradición (protestante o católica) a la que el investigador o anotador se acoja.

Estos no son los únicos sesgos relacionados con cualquier proceso de anotación, sino que hay otros aspectos lingüísticos, culturales y personales. En ningún caso quiero presentar este trabajo como la única anotación posible, la única interpretación objetiva. Otras muchas anotaciones similares podrían realizarse. Lo interesante es que, si estas anotaciones también se realizasen en TEI, nos sería relativamente sencillo contrastar mi anotación con la anotación de otras personas, o incluso cómo mi interpretación va modificándose con el paso del tiempo en caso de que volviese a anotarlo dentro de algunos años. Así, se podría ver con exactitud en qué temas o qué libros hay consenso y qué categorías son anotadas de manera dispar. De esta manera, TEI y medios informáticos posibilitan discusiones subjetivas sobre categorías abstractas utilizando datos muy concretos.

4. Cambio de paradigma en las Humanidades: datos cualitativos a escala cuantitativa

La principal motivación histórica de las HD no ha sido el interés de los humanistas por nuevos métodos computacionales. Más bien ha sido el acceso a datos digitales, ya fuesen procedentes de una digitalización masiva o generados como objetos digitales de manera nativa (como blogs poéticos, *fan fiction* o comunicación en Twitter). Una vez que los datos estaban en formatos como texto plano, PDF o web, investigadores y usuarios se preguntaron en su día sobre las posibilidades de visualización y análisis. Muchas veces ni siquiera eran humanistas quienes realizaban esas primeras preguntas, sino informáticos. Con ese desarrollo, resulta comprensible que en las HD se hayan establecido predominantemente metodologías cuantitativas que buscan tendencias generales en datos que no han sido explícitamente anotados. Me refiero a técnicas como el *clustering* de datos (en estilometría), *topic modeling*, grafos de coocurrencias o *word embeddings*. Para aplicar todas estas metodologías, lo único que se requiere son los textos en formato digital, sin necesidad de demasiados metadatos ni anotación en oraciones o párrafos.

Sin embargo, el recorrido de esas metodologías muchas veces es corto. Tras explorarse en diferentes conjuntos de datos y mostrar visualizaciones espectaculares, en seguida el mismo investigador percibe sus limitaciones: selección de resultados (*cherry picking*), inexactitud, falta de evaluación, sobreinterpretación de visualizaciones, etc. (Da, 2019).

Esta fase de las HD que analiza materiales digitalizados con metodologías no supervisadas es a largo plazo peligrosa por dos razones. En primer lugar, porque restringe fuertemente las posibilidades de análisis de materiales humanísticos utilizando métodos computacionales. Estas metodologías no hubiesen sido suficientes como para analizar el tema del sexo en la Biblia, o me habrían obligado a simplificarlo hasta puntos cuestionables.

El segundo peligro que estas metodologías no supervisadas suponen para las HD es el cuestionamiento de una parte importante de nuestro trabajo como humanistas. El argumento más frecuente para utilizar metodologías no supervisadas como *clustering* o reducciones dimensionales como *topic modeling* es ahorrarse horas de trabajo de especialistas en el área. Argumentaciones similares pueden encontrarse en cualquier introducción al aprendizaje automático (Alpaydin, 2010; Müller y Guido, 2016; Géron, 2019). Los humanistas son los especialistas de las Humanidades, por lo que ese argumento en realidad puede resultar nocivo a la larga para nuestra comunidad.

Los Humanistas Digitales harían bien en oponerse al enfoque de que el aprendizaje automático es una manera de crear mayores beneficios eliminando las horas de trabajo de especialistas en el área. En su lugar, propongo que las HD acentúen que nuestro trabajo como especialistas en combinación con metodologías computacionales como el aprendizaje automático nos permitirá afrontar retos que hasta ahora estaban fuera de nuestro alcance. Este nuevo paradigma puede verse facilitado mediante tres marcos concretos:

1. Métodos supervisados: Favorecer métodos supervisados en los que anotaciones cualitativas sean parte del proceso.
2. *Open Access*: Publicar los datos de manera abierta para visibilizar nuestro trabajo y que otros proyectos se beneficien también de él.
3. TEI y LOD: La anotación cuidada por parte de especialistas debe realizarse en formatos digitales apropiados. Aunque para ciertos datos como información biográfica o bibliográfica los datos enlazados en formatos Linked Open Data son buenos candidatos, TEI continúa siendo la mejor solución para textos.

De esta manera, no estoy de acuerdo con que las HD aporten un nuevo enfoque cuantitativo frente al enfoque cualitativo de las Humanidades tradicionales. Está en el interés de las Humanidades (Digitales o no) anotar y desarrollar cada vez más información, no solo una digitalización en masa. Datos en formato digital, subjetivos, marcados por múltiples realidades históricas, culturales y lingüísticas complejas. Datos de certeza explícitamente cuestionable que puedan ser discutidos también en formatos digitales. Todos esos matices pueden marcarse en TEI, así como compartirse con el resto de la comunidad. De esta manera superaremos la dicotomía actual y estaremos en una senda que aúna lo cuantitativo con lo cualitativo.

Referencias bibliográficas

Allés-Torrent, Susanna (2015): «Edición digital y algunas tecnologías aliadas», *Ínsula. Revista de Letras y Ciencias Humanas*, 822, pp. 18-21.

Allés-Torrent, Susanna, Gimena del Rio Riande y Gabriel Calarco (2019): «TTHUB: Text Technologies Hub for Extending TEI Training in Spanish». DOI: <https://doi.org/10.5281/zenodo.3514441>.

Alpaydin, Ethem (2010): *Introduction to Machine Learning*, Cambridge, MIT Press.

Biblioteca Virtual Miguel de Cervantes, Alicante, Universitat d'Alacant, <www.cervantesvirtual.com> [fecha de consulta: 6-2-2021].

Burnard, Lou (2014): *What is the Text Encoding Initiative?*, Marsella, OpenEdition Press, <http://books.openedition.org/oep/679> [fecha de consulta: 6-2-2021].

Calvo Tello, José (2018): «Genre Classification in Spanish Novels: A Hard Task for Humans and Machines?», en *Data in Digital Humanities*, Galway, EADH, <https://eadh2018.exordo.com/programme/presentation/82> [fecha de consulta: 6-2-2021].

Calvo Tello, José (2020): «What is a Genre? A Graph Unified Model of Categories, Texts, and Features», Ottawa, ADHO, <https://hcommons.org/deposits/item/hc:31713/> [fecha de consulta: 6-2-2021].

Calvo Tello, José (2021): *The Novel in the Spanish Silver Age. A Digital Analysis of Genre Using Machine Learning*, Bielefeld, transcript, <http://www.transcript-verlag.de/978-3-8376-5925-2> [fecha de consulta: 28-3-2022].

Calvo Tello, José, Ulrike Henny-Krahmer y Christof Schöch (2018): «Textbox: análisis del léxico mediante corpus literarios», en Dolores Corbella, Alejandro Fajardo y Jutta Langenbacher-Liebgott (eds.), *Historia del léxico español y Humanidades digitales*, Berlín, Peter Lang, pp. 223-251.

Canet, José Luis (ed.) (2017): *La Celestina (Tragicomedia de Calisto y Melibea)*, Clásicos Hispánicos. DOI: <https://doi.org/10.5281/zenodo.3753556>.

Christodouloupoulos, Christos y Mark Steedman (2015): «A massively parallel corpus: the Bible in 100 languages», *Language Resources and Evaluation*, 49, 2, pp. 375-395. DOI: <https://doi.org/10.1007/s10579-014-9287-y>.

Cummings, James (2019): «A world of difference: Myths and misconceptions about the TEI», *Digital Scholarship in the Humanities*, 34, número suplementario 1, pp. i58-i79. DOI: <https://doi.org/10.1093/llc/fqy071>.

Da, Nan Z. (2019): «The Computational Case against Computational Literary Studies», *Critical Inquiry*, 45, 3, pp. 601-639. DOI: <https://doi.org/10.1086/702594>.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch y Thorsten Vitt (2017): «Understanding and explaining Delta measures for authorship attribution», *Digital Scholarship in the Humanities*, 32, número suplementario 2, pp. ii4-ii16. DOI: <https://doi.org/10.1093/llc/fqx023>.

Fièvre, Paul (2007): *Théâtre classique*, París, Université Paris-IV Sorbonne, <http://www.theatre-classique.fr> [fecha de consulta: 6-2-2021].

Fischer, Frank, Peer Trilcke, Julia Jennifer Beine y Boris Orekhov (2018): *DraCor*, Moscú, <https://dracor.org/> [fecha de consulta: 6-2-2021].

Fradejas Rueda, José Manuel (2018): *7 Partidas Digital XML-TEI*, Valladolid, Universidad de Valladolid. DOI: <http://doi.org/10.5281/zenodo.1195642>.

Géron, Aurélien (2019): *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems*, Beijing, O'Reilly.

Grandjean, Martin (2013): «Comparing the relational structure of the Gospels. Network Analysis as a tool for biblical sciences», en *Conference of the Society of Biblical Literature*, Saint Andrews, University of St Andrews, <https://hal.archives-ouvertes.fr/hal-01525574/document> [fecha de consulta: 6-2-2021].

Grötschel, Martin (2007): *Deutsches Textarchiv*, Berlín, Berlin-Brandenburgische Akademie der Wissenschaften, <http://www.deutschestextarchiv.de/> [fecha de consulta: 6-2-2021].

Gutiérrez de la Torre, Silvia Eunice (2020): «Bibliotecas y Humanidades Digitales en América Latina», *Revista de Humanidades Digitales*, 5, pp. 113-131. DOI: <https://doi.org/10.5944/rhd.vol.5.2020.27826>.

Harold, Elliotte Rusty y W. Scott Means (2004): *XML in a Nutshell*, Beijing, O'Reilly, <http://shop.oreilly.com/product/9780596007645.do> [fecha de consulta: 6-2-2021].

Hatav, Galia (2000): «(Free) Direct Discourse in Biblical Hebrew», *Hebrew Studies*, 41, 1, pp. 7-30. DOI: <https://doi.org/10.1353/hbr.2000.0063>.

Henny-Krahmer, Ulrike (2017): *Bib-ACMé. Bibliografía digital de novelas argentinas, cubanas y mexicanas (1830-1910)*, Würzburg, CLiGS, <http://bibacme.cligs.digital-humanities.de/> [fecha de consulta: 6-2-2021].

Henny-Krahmer, Ulrike (2018): «Exploration of Sentiments and Genre in Spanish American Novels», en *Puentes/Bridges*, México, ADHO, <https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/> [fecha de consulta: 6-2-2021].

Hettinger, Lena, Martin Becker, Isabella Reger, Fotis Jannidis y Andreas Hotho (2015): «Genre Classification on German Novels», en *Proceedings of the 12th International Workshop on Text-based Information Retrieval*, Weimar.

Hockey, Susan M. (2000): *Electronic Texts in the Humanities. Principles and Practice*, Oxford, Oxford University Press.

Horstmann, Wolfram (2006): *TextGrid. Virtuelle Forschungsumgebung für die Geisteswissenschaften*, Göttingen, TextGrid Konsortium, <https://textgrid.de> [fecha de consulta: 6-2-2021].

Isasi, Jennifer (2017): «Acercamiento al análisis del sistema de los personajes en la narrativa escrita en español: el caso de *Zumalacárregui y Mendizábal* de Pérez Galdós», *Caracteres. Estudios Culturales y Críticos de la Esfera Digital*, 6, 2, pp. 107-137, <http://revistacaracteres.net/wp-content/uploads/2017/12/Caracteresvol6n2noviembre2017.pdf> [fecha de consulta: 6-2-2021].

Jauralde Pou, Pablo (2013): *Clásicos Hispánicos*, Madrid-Würzburg, More Than Books, <http://www.clasicoshispanicos.com/> [fecha de consulta: 6-2-2021].

Jiménez Fernández, Concepción María, Elena Martínez Carro, María Teresa Santa María, José Calvo Tello, María Simón Parra, Roxana Beatriz Martínez Nieto y María García Sánchez (2017): «BETTE: Biblioteca Electrónica Textual del Teatro en Español de la Edad de Plata», en *Sociedad, políticas, saberes*, Málaga, HDH, pp. 88-91, <http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf> [fecha de consulta: 6-2-2021].

Jiménez Fernández, Concepción María y José Calvo Tello (2020): «Grafos de escenas y estudios literarios digitales: una propuesta computacional crítica», *452ºF. Revista de Teoría de la Literatura y Literatura Comparada*, 23, pp. 78-101. DOI: <https://doi.org/10.1344/452f.2020.23.4>.

Krautter, Benjamin, Janis Pagel, Nils Reiter y Marcus Willand (2018): *Titelhelden und Protagonisten - Interpretierbare Figurenklassifikation in deutschsprachigen Dramen*, Digital Humanities Cooperation,

<https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf> [fecha de consulta: 6-2-2021].

Mowat, Barbara y Paul Werstine (2010): *Shakespeare Folger Library*, Washington, Folger, <https://www.folgerdigitaltexts.org> [fecha de consulta: 6-2-2021].

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

Navarro-Colorado, Borja (2015): «A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects», en *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, <http://www.dlsi.ua.es/~borja/navarro2015_GoldenAgeSonnets.pdf> [fecha de consulta: 6-2-2021].

Navarro-Colorado, Borja (2018): «A metrical scansion system for fixed-metre Spanish poetry», *Digital Scholarship in the Humanities*, 33, 1, pp. 112-127. DOI: <https://doi.org/10.1093/llc/fqx009>.

Odebrecht, Carolin, Lou Burnard, Borja Navarro-Colorado y Christof Schöch (2019): «European Literary Text Collection (ELTeC): Release with 10 collections of at least 50 novels», Zenodo. DOI: <https://doi.org/10.5281/ZENODO.4274954>.

Pizarro Pedraza, Andrea (2013): «Tabú y eufemismo en la ciudad de Madrid: estudio sociolíngüístico-cognitivo [*sic*] de los conceptos sexuales», tesis doctoral, Universidad Complutense de Madrid.

Presotto, Marco, Sònia Boadas, Eugenio Maggi, Aurelia Pessarrodona, Francesca Tomasi, Marilena Daquino y Raffaele Messuti (2015): *«La dama boba»: edición crítica y archivo digital*, Barcelona-Bolonia, PROLOPE-Alma Mater Studiorum-Università di Bologna, CRR-MM, <http://damaboba.unibo.it> [fecha de consulta: 6-2-2021].

Resnik, Philip, Mari Broman Olsen y Mona Diab (1999): «The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues"», *Computers and the Humanities*, 33, 1, pp. 129-153. DOI: <https://doi.org/10.1023/A:1001798929185>.

Rio Riande, Gimena del (2020): *Poesía Medieval*, Buenos Aires, Universidad de Buenos Aires-CONICET, <http://hdlab.space/Poesia-Medieval/> [fecha de consulta: 6-2-2021].

Rojas Castro, Antonio (2016): *«Soledades» de Luis de Góngora*, Barcelona, Universitat Pompeu Fabra, <http://www.soledadesediciondigital.com/> [fecha de consulta: 6-2-2021].

Ruiz Fabo, Pablo, Helena Bermúdez Sabel, Clara Martínez Cantón, Elena González-Blanco y Borja Navarro-Colorado (2018): «The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings-DH 2018», en *Puentes/Bridges*, México, ADHO, <https://dh2018.adho.org/the-diachronic-spanish-sonnet-corpus-disco-tei-and-linked-open-data-encoding-data-distribution-and-metrical-findings/> [fecha de consulta: 6-2-2021].

Ruiz Fabo, Pablo, Clara Martínez Cantón y José Calvo Tello (2017): *DISCO: Diachronic Spanish Sonnet Corpus*, Madrid, UNED, <https://github.com/pruizf/disco> [fecha de consulta: 6-2-2021].

Rybicki, Jan y Maciej Eder (2011): «Deeper Delta across genres and languages: do we really need the most frequent words?», *Literary and Linguistic Computing*, 26, 3, pp. 315-321. DOI: <https://doi.org/10.1093/llc/fqr031>.

Sánchez Sánchez, Mercedes y Carlos Domínguez Cintas (2007): «El banco de datos de la RAE: CREA y CORDE», *Per Abbat. Boletín Filológico de Actualización Académica y Didáctica*, 2, pp. 137-148, <https://dialnet.unirioja.es/descarga/articulo/2210249.pdf> [fecha de consulta: 6-2-2021].

Santa María, María Teresa, José Calvo Tello y Concepción María Jiménez Fernández (2020): «¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata», *Digital Scholarship in the Humanities*, 36, número suplementario 1, pp. i81-i88. DOI: <https://doi.org/10.1093/llc/fqaa015>.

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer y Stefanie Popp (2019): «The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML», *Journal of the Text Encoding Initiative*. DOI: <https://doi.org/10.4000/jtei.2085>.

Schöch, Christof, Ulrike Henny, José Calvo Tello, Daniel Schlör y Stefanie Popp (2016): «Topic, Genre, Text. Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1880-1930)», Leipzig, nisaba verlag, pp. 235-238, <http://dhd2016.de/boa.pdf> [fecha de consulta: 6-2-2021].

Simón Palmer, María del Carmen (1997): *Teatro Español del Siglo de Oro*, Ann Arbor, ProQuest, <teso.chadwyck.com> [fecha de consulta: 6-2-2021].

Terras, Melissa, Edward Vanhoutte y Ron Van den Branden: *TEI by Example*, <https://teibyexample.org> [fecha de consulta: 6-2-2021].

Trilcke, Peer, Frank Fischer, Mathias Göbel y Dario Kampkaspar (2016): «Dramen als small worlds? Netzwerkdaten zur Geschichte und Typologie deutschsprachiger Dramen 1730-1930», en Elisabeth Burr (ed.), *DHd 2016 Modellierung, Vernetzung, Visualisierung*, Leipzig, DHd-nisaba, pp. 254-257.

Wilkinson, Mark D. *et al.* (2016): «The FAIR Guiding Principles for scientific data management and stewardship», *Scientific Data*, 3. DOI: <https://doi.org/10.1038/sdata.2016.18>.