

Aportaciones al estudio de la variación diacrónica del español con un corpus paralelo de traducciones bíblicas medievales y renacentistas

Andrés Enrique-Arias

Universitat de les Illes Balears

Biblia Medieval is a free-access computer tool on the web that features all the Spanish biblical translations composed during the Middle Ages in paleographic transcription. This online resource is being expanded into the new *Biblias Hispánicas* corpus, which will include two sixteenth century translations as well as upgraded texts with normalized spelling, lemmatization and POS tagging. This article presents an overview of the main technical and philological issues that arised in the preparation of this corpus and it includes examples of some of its applications for the study of Spanish language and culture in the medieval and Renaissance periods.

Keywords: Bible, historical linguistics, parallel corpora, digital humanities, language variation

1. Introducción: las biblias castellanas como corpus

Las traducciones de la Biblia al castellano hechas durante la Edad Media constituyen un capítulo extraordinario en la historia cultural europea tanto por su amplio número, sin parangón entre los vernáculos premodernos europeos, como por el hecho de ser en su mayoría traducciones hechas directamente desde el hebreo. Esta rica tradición tiene su continuidad en el Renacimiento, con las biblias producidas en el siglo XVI por exiliados y huidos españoles vinculados bien al protestantismo o al judaísmo.

Por su importancia en la historia cultural española y europea el amplio corpus de las traducciones bíblicas castellanas medievales y renacentistas ha suscitado el interés de especialistas provenientes de variados campos humanísticos: lengua, literatura, religión, estudios judíos, traductología, filosofía, historia general, historia del arte. Un aspecto que ha empezado a explotarse de manera más intensa en

los últimos años es su aprovechamiento como corpus para el estudio de la variación y el cambio lingüísticos en la historia de la lengua española. Esta posibilidad ha sido facilitada por la creación, por parte del equipo de la Universitat de les Illes Balears coordinado por el que suscribe, de nuevos recursos informáticos accesibles de manera libre en la página *Biblia Medieval* (www.bibliamedieval.es). El principal recurso es el corpus *Biblia Medieval*, que contiene transcripciones paleográficas de los romanceamientos bíblicos medievales dispuestos en paralelo junto a sus fuentes latinas y hebreas así como más de 17000 imágenes digitales de los facsímiles de los originales. Asimismo los usuarios pueden descargarse sin restricciones los textos del corpus y los resultados de las búsquedas para trabajar con los materiales en su propio ordenador. La página contiene otros recursos que incluyen una base de datos bibliográfica exhaustiva y enlaces a concordancias de los textos.

En las páginas que siguen presentamos de manera resumida los avances conseguidos a través del corpus y sus herramientas asociadas y avanzamos los desarrollos y mejoras que estamos llevando a cabo en la nueva fase del proyecto.

2. El corpus *Biblia medieval*: avances y necesidades

La proliferación de ediciones electrónicas y el desarrollo de herramientas informáticas para el análisis automático de extensas colecciones de textos históricos han supuesto una auténtica revolución metodológica en el campo de la lingüística diacrónica de los últimos 25 años (Joseph 2008). En lo que respecta al caso del español, la temprana creación desde finales de la década de 1990 de extensas bases de datos textuales de uso libre en la red ha abierto grandes posibilidades para incorporar el uso de estas metodologías al estudio de la historia de la lengua española.

La edición integral del corpus de las traducciones bíblicas al castellano en transcripción electrónica ha permitido por primera vez hacer estudios de conjunto de las versiones medievales y avanzar en el conocimiento de los rasgos que caracterizan las diferentes traducciones y su distribución en los testimonios existentes (cf. Pueyo y Enrique-Arias 2013). Asimismo la disponibilidad del corpus paralelo ha permitido nuevas posibilidades de estudio de un amplio número de fenómenos de variación de la historia del español (Enrique-Arias 2012, 2016).

Está claro, por otro lado, que los textos paleográficos sin anotación con los que contamos en la actualidad no pueden ser sustitutos de ediciones críticas con texto normalizado: los errores de copista y la falta de puntuación hace que muchos

pasajes sean incomprensible incluso para especialistas. Al mismo tiempo, la considerable variación gráfica del castellano medieval y la falta de lematización limita y entorpece las búsquedas. El problema es que eliminar esta variación gráfica sin más supone excluir información que puede ser relevante para los estudiosos que se interesan por el componente gráfico-fónico de los manuscritos. Afortunadamente, la búsqueda de soluciones para mantener este delicado equilibrio entre normalización y paleografismo se ha visto beneficiada por la aparición de propuestas renovadas de criterios de presentación gráfica. En el caso concreto de nuestro proyecto de edición suscribimos las propuestas de Pedro Sánchez-Prieto (2011) desarrolladas en el ámbito de CHARTA (www.charta.es), una nutrida red internacional de grupos de investigación dedicados a la edición de documentos antiguos hispánicos a la que pertenece el grupo que desarrolla el corpus bíblico que nos ocupa. La solución consiste en ofrecer al lector varios niveles de acceso al texto, algo que con la disponibilidad de los medios electrónicos actuales cuenta con muchas menos limitaciones que en el caso de los formatos en papel. En consecuencia, nos proponemos añadir a las transcripciones paleográficas y facsímiles ya existentes en la página del proyecto un tercer nivel de acceso al texto con una normalización sistemática de grañas, puntuación, acentuación y uso de mayúsculas.

Además, los textos paleográficos sin anotación que forman el corpus *Biblia Medieval* en la actualidad presentan dificultades evidentes para realizar búsquedas lingüísticas complejas. Por ello, los textos del nuevo corpus que estamos creando tendrá la ventaja y versatilidad de la triple presentación; además, el texto estandarizado que proponemos contará con lematización y anotación gramatical exhaustiva que permitirá búsquedas complejas con posibilidad de discriminación de categorías gramaticales. De este modo la edición que proponemos permitirá hacer estudios innovadores de historia de la lengua pues será posible hacer búsquedas con condicionamientos gramaticales que discriminen formas homónimas.

Otra limitación del corpus *Biblia Medieval* es que, dado que incluye exclusivamente versiones compuestas durante la Edad Media, su arco cronológico, que abarca aproximadamente desde 1200 hasta 1450, es insuficiente para estudiar muchos de los procesos evolutivos de la historia del español. Hay por ejemplo todo un elenco de fenómenos de cambio morfosintáctico de gran relevancia teórica que se dan en la transición del español medieval al español clásico (Sánchez Lancis 2009: 163). Por tanto, extender el corpus incorporando las versiones del XVI, amplía las posibilidades del corpus para la investigación del cambio lingüístico. La disciplina de Humanidades Digitales pretende acercar los objetos culturales producidos por las diversas civilizaciones tanto a la comunidad investigadora como al público general. Este último, gracias a los avances tecnológicos y a la

creciente apertura del conocimiento científico (*Open Access*) de los últimos años, puede tener acceso de forma asequible y sencilla a los recursos generados por los investigadores en el campo de las Humanidades. Un componente esencial de nuestro proyecto desde sus inicios es la apuesta por los espacios en la red aprovechando todas las nuevas funcionalidades que ofrecen los medios electrónicos. La creación de *Biblia Medieval* ha supuesto, además de su aprovechamiento para nuestras tareas de edición y de estudios filológicos y lingüísticos de las biblias medievales en castellano, la habilitación de un espacio abierto a la comunidad científica en el que es posible acceder a todos los elementos necesarios para la constitución de estos textos: imágenes de los manuscritos, transcripciones de las fuentes primarias, estudios relacionados, concordancias, enlaces a otras webs, etc, todo ello accesible mediante una base de datos con una herramienta de búsqueda. Se trata en definitiva de la propuesta que los teóricos de la ecdótica angloamericana han denominado “edición social” (Siemens *et al.* 2010). El nuevo corpus y recursos asociados aspiran a mantener y continuar esta función de referente internacional de todo lo relacionado con la traducción bíblica al español en la Edad Media y, ahora, también en el Renacimiento.

3. El nuevo corpus *Biblias Hispánicas*

El nuevo corpus que proponemos tiene dos novedades principales. En lo que se refiere a los textos que lo componen destacamos la inclusión de las dos traducciones renacentistas de biblias completas que existen. La primera de ellas es la Biblia de Ferrara, publicada por primera vez en 1553 y destinada a conversos y criptojudíos españoles que regresaban a la práctica abierta del judaísmo; la segunda es la Biblia de Casiodoro de la Reina, publicada en Basilea en 1569 para el uso de los luteranos españoles exiliados en Europa. Ya hemos comentado el interés que supone aumentar un siglo el ámbito cronológico del corpus bíblico medieval para incluir textos del XVI y facilitar así el estudio de la evolución del español en el tránsito del español medieval al clásico. Otro aspecto valioso de la inclusión de estos textos es que vienen a remediar la falta de versiones en línea de estas importantes biblias: nuestra edición electrónica permitirá por fin hacer búsquedas en estas traducciones en su conjunto y aportar estudios novedosos. Por último, el poder consultar estas biblias del XVI en paralelo con las versiones medievales permitirá conocer mejor las interrelaciones entre estos textos y sus antecedentes premodernos.

El proyecto supone además el empleo de herramientas novedosas dentro del campo de la lingüística de corpus. El uso de técnicas innovadoras y de programas

para el Procesamiento del Lenguaje Natural (PLN) resulta fundamental para el objetivo final de recuperar los textos paleográficos de los que se parte y transmitirlos en forma normalizada y crítica para facilitar la consulta del contenido cultural de los textos a los estudiosos de las diferentes disciplinas del área de la Humanidades, así como al público general que quiera hacer uso de este recurso digital. La metodología que vamos a emplear para realizar los procesos de normalización, segmentación, lematización y etiquetación lingüística resulta innovadora en el sentido de que se aplicarán los numerosos avances y recursos de la disciplina de PLN desarrollados para las lenguas modernas romances y se adaptarán al tratamiento del español de los periodos históricos representados en el corpus. Para llevar a cabo esta adaptación nuestro grupo de investigación hará uso de la *suite* de programas y recursos lingüísticos *Freeling*, licenciado en su mayor parte para el uso, distribución y adaptación de forma libre y abierta. El grupo contará con un experto en lingüística computacional y de corpus que desarrollará las reglas, diccionarios y corpus de entrenamiento particulares al español bíblico medieval y hará la integración con el código de *Freeling*. Todos los recursos lingüísticos generados se ofrecerán a los desarrolladores de *Freeling* para su integración en la suite, con el propósito de que otros grupos de investigación puedan aprovechar el trabajo realizado en este proyecto. Igualmente, siguiendo el espíritu del *Open Access* y nuestro respeto por los estándares internacionales, todo el conjunto textual generado, incluyendo la lematización y la etiquetación del corpus completo, estará disponible para su descarga en el lenguaje de intercambio XML, siguiendo la propuesta de marcación establecida por el consorcio *Text Encoding Initiative* (TEI) para la representación e intercambio de textos humanísticos en formato digital.

En nuestro proyecto es prioritaria la creación de una interfaz vistosa e innovadora que facilite el acceso de los usuarios al corpus. El objetivo final es facilitar y generalizar lo más posible los procesos de consulta y lectura de los textos. Estamos trabajando en diseños y prototipos para el desarrollo de dicha interfaz y ya hemos conseguido el objetivo de ofrecer, en una página única y nítida, el acceso dinámico a la compleja triple presentación que se propone para el corpus. El uso de ventanas emergentes asociadas a los diferentes elementos de la página permite tener a la vista de forma paralela cualquier fragmento textual en sus tres variantes, paleográfica, crítica y facsimilar (y en este último caso, el aumento dinámico del tamaño del texto a cualquier nivel de zoom necesario para poder leer cómodamente la escritura medieval o impresos quinientistas originales). En el apartado de los resultados de las búsquedas, se generarán, también de forma interactiva, una serie de gráficos y tablas estadísticos que explicarán de forma instantánea la distribución de dicha consulta teniendo en cuenta numerosos parámetros cronológicos, geográficos y tipológicos. Finalmente se crearán formularios de consulta

guiados e interactivos que pondrán a disposición del usuario de forma sencilla todas las herramientas lingüísticas necesarias para explorar el corpus de forma no solo cómoda sino también exhaustiva.

4. Conclusión

Biblias Hispánicas permitirá la consulta en paralelo de las diferentes versiones bíblicas castellanas que abarcan desde la época de aparición de los primeros textos en prosa plenamente castellanos en la primera mitad del siglo XIII hasta el siglo XVI. Los textos, que estarán lematizados y etiquetados gramaticalmente, permitirán apreciar con particular claridad la evolución de la lengua en todos sus niveles de análisis y asimismo facilitarán nuevas vías de investigación en el estudio de esta importante faceta de la historia cultural de España y Europa.

Agradecimientos

Esta investigación se ha llevado a cabo en el marco del proyecto “Edición electrónica integral de las Biblias españolas medievales y renacentistas” financiado con una ayuda de la Fundación BBVA a Equipos de Investigación Científica 2016/Humanidades Digitales.

Referencias

- Enrique-Arias, A. 2012. Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad, *Scriptum digital* 1: 85-106.
- Enrique-Arias, A. 2016. Sobre la noción de perspectiva en lingüística de corpus: algunas ventajas de los corpus paralelos. En J. Kabatek (ed), *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín: De Gruyter, 21-39.
- Joseph, B. 2008. Historical Linguistics in 2008: The state of the art. En P. van Sterkenburg (ed), *Unity and Diversity of Languages*. Amsterdam: John Benjamins, 175-188.
- Pueyo Mena, F. J., & Enrique-Arias, A. 2013. Los romanceamientos castellanos de la Biblia Hebrea compuestos en la Edad Media: manuscritos y traducciones. *Sefarad* 73, 1: 165-224.
- Sánchez Lancis, C. 2009. Corpus diacrónicos y periodización del español. *Cahiers d'études hispaniques medievales* 32: 159-180
- Sánchez-Prieto Borja, P. 2011. *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*. San Millán de la Cogolla: Cilengua.
- Siemens, R., Elkind, M., McColl, A. et al. 2010. Underpinnings of the social edition? A narrative, 2004-9, for the Renaissance English Knowledgebase (REKn) and Professional Reading Environment (PRE) projects. En J. McGann, con A. Stauffer, D. Wheelles &

Michael P. (eds), *Online humanities scholarship: The shape of things to come*. Houston: Rice University Press.