Diseño y construcción de un corpus de referencia de latín

Mercedes García Ferrer

IES J. B. Porcar, Castellón

xedes.garcia@gmail.com

Recepción: 4/11/2013

Resumen

La incorporación de los avances de la lingüística de corpus a la didáctica del latín es ya una realidad, que ha permitido mejorar las herramientas para trabajar los textos latinos y las posibilidades de explotación para su estudio. Por ello proponemos el diseño y la creación de un gran Corpus LAtino de REferencia (CLARE) como un monitor corpus, un corpus colaborativo de carácter abierto, que se ajuste a las normas de codificación de textos electrónicos, que pueda seguir creciendo de forma ordenada, estructurada y supervisada, que sirva como herramienta de enseñanza y aprendizaje del latín, y con el cual poder contrastar los diferentes trabajos e investigaciones sobre textos latinos.

Palabras clave: corpus, latín, didáctica, lingüística, lexicografía.

La incorporación de los avances de la lingüística de corpus a la didáctica del latín es ya una realidad. Desde los primeros corpus dedicados al estudio de obras concretas, como el Corpus Thomisticum de Busa, hasta los proyectos colaborativos como el LASLA, el CETEDOC, el consorcio ADHO o los diversos proyectos que integran la asociación del CHLT, las herramientas para trabajar los textos latinos y las posibilidades de su estudio se han multiplicado.

La lingüística de corpus ha evolucionado desde la creación de corpus *ad hoc* hasta la generación de grandes corpus de referencia de millones de palabras como el BNC, el BoE en el ámbito anglosajón, o el CREA, el CUMBRE y el CORDE para el español, el CTILC para el catalán, el CORGA para el gallego, o el EHHA para el euskera.

Por ello proponemos el diseño y la creación de un gran Corpus LAtino de REferencia (en adelante CLARE) como un *monitor corpus*, un corpus colaborativo de carácter abierto, que se ajuste a las normas de codificación de textos electrónicos, que pueda seguir creciendo de forma ordenada, estructurada y supervisada, que sirva como herramienta de enseñanza y aprendizaje del latín, y con el cual poder contrastar los diferentes trabajos e investigaciones sobre textos latinos.

1. Descripción y taxonomía de los corpus electrónicos

Según Bowker y Pearson (2002, 9) los corpus han de reunir cuatro criterios fundamentales:

- que sean auténticos (*authentic*), es decir, que no hayan sido creados para la ocasión;
- que estén recogidos en formato electrónico *(electronic form*) para poder ser procesados por un ordenador;

- que sean lo suficientemente cuantiosos (*large collection*) para poder ser realmente significativos y cumplir los objetivos deseados;
- que los criterios de selección sean rigurosos (*specific set of criteria*) y no se trate de una selección aleatoria de textos.

En el presente trabajo, pues, entendemos «corpus» como un conjunto de textos recogidos según unos criterios determinados (los anteriormente citados) para ser utilizados con unos propósitos específicos y en un formato legible por el ordenador. Una vez determinado qué entendemos por corpus, conviene detenerse en su tipología, para así luego describir, con mayor precisión y rigor, el corpus propio.

Los tipos de corpus que se pueden crear son tan variados como los propósitos para los cuales se compilan. El primer nivel de los estudios de corpus lo establece Roberts (2006) al distinguir entre corpus generales (CG) y corpus lingüísticos (CLG). Los CG son conjuntos de textos recopilados para funciones diversas, mientras que los CLG se destinan especialmente al estudio de facetas lingüísticas.

En el segundo nivel de descripción, investigadores como Rabadán y Fernández Nistal (2002) dividen los corpus en monolingües, entre los que diferencian los referenciales de los especializados, y los bilingües o multilingües, que a su vez Bowker y Pearson (2002, 92-93), Granger (2003) y Olohan (2004) dividen en paralelos y comparables. Los corpus comparables ponen en relación conjuntos de textos que, siendo del mismo idioma, pertenecen a situaciones diversas pero contrastables entre sí. Comparable sería, por ejemplo, el contraste de un corpus electrónico de *La guerra de las Galias* con una producción en latín de otro autor latino coetáneo a César. Los corpus paralelos, por su parte, ponen en relación unos textos originales y otros traducidos. Así un corpus de *La guerra de las Galias* en latín y otro en español serían paralelos o podrían utilizarse de forma paralela.

De entre las diferentes taxonomías de corpus existentes, nos parecen especialmente pertinentes los trabajos de Laviosa, por su exhaustiva clasificación teórica, los de Granger, por su precisa aplicación de los estudios de corpus al ámbito de la enseñanza, y los de Cabré, por sus investigaciones en el ámbito de la lexicología y sus aplicaciones en el campo de la lexicografía.

Laviosa (2002) dividió en cuatro niveles todos los tipos de corpus electrónicos que existen en los Estudios de Traducción de Corpus (Corpus-based Translation Studies, CTS). En el primer nivel, Laviosa (2002, 34-5) establece seis apartados:

1. en el primero recoge los corpus de textos completos, de extractos, los mixtos (con mezcla de textos completos y extractos) y los corpus monitorizados o *monitor corpus*, (también llamados «corpus abiertos»);
2. en el segundo Laviosa identifica corpus diacrónicos y sincrónicos;
3. en el tercer apartado sitúa los generales y los terminológicos;
4. en el cuarto, los monolingües, bilingües y multilingües;
5. en el quinto se centra en la lengua (o lenguas) del corpus;
6. en el sexto apartado, distingue entre corpus escritos, orales o mixtos (mezcla de las dos modalidades anteriores).

En el segundo nivel taxonómico, Laviosa (2002, 36) divide los corpus monolingües en simples (recopilación de textos en una única lengua) y comparables. Los bilingües quedan clasificados en: paralelos (textos originales en lengua A y sus traducciones en lengua B) y comparables (textos originales en lengua A y textos originales en lengua B). Los multilingües se agrupan en: paralelos (textos originales en lenguas diversas con sus respectivas traducciones) y comparables.

En el tercer nivel de la taxonomía de Laviosa (2002) los corpus simples se bifurcan en traductores y no traductores. Los corpus bilingües paralelos son unidireccionales y bidireccionales. Los corpus multilingües paralelos se componen de una única lengua origen, de dos lenguas origen y de varias lenguas origen. En el cuarto y último nivel, Laviosa (2002) profundiza en los corpus traductores y también encuentra corpus de una, dos o más lenguas de partida.

Por su parte, Granger (2003, 21) estructura los corpus en multilingües y monolingües. En esta primera fase se observa ya una diferencia respecto a Laviosa, pues Granger comienza por los multilingües, en los que distingue entre corpus traductores/paralelos frente a los comparables.

Los primeros se bifurcan en unidireccionales (en los que los documentos traducidos son de la misma lengua), y bidireccionales (en los que conviven traducciones y no traducciones de distintas lenguas). Los corpus comparables pueden ser de textos originales o de textos traducidos. En el bloque de los corpus monolingües, Granger sólo señala corpus comparables subdivididos en dos grupos, los que enfrentan textos originales a traducciones, y los que oponen textos nativos y no nativos.

Por último, Cabré (citada por Gelpí, 1997) propone una tipología funcionalista de corpus lingüísticos, tomando como criterio la combinación de tres parámetros:

1. La finalidad para la que se han construido (corpus generales frente a corpus específicos). Mientras que los corpus con finalidades generales constituyen una fuente de información textual para diversos usos, los corpus con finalidades específicas intentan dar respuesta a propósitos claramente definidos, como el estudio de aspectos concretos de la gramática o el léxico, la extracción de datos estadísticos, el estudio del comportamiento lingüístico de una determinada población de hablantes, el análisis comparativo de diversas variedades lingüísticas, el desarrollo y evolución de sistemas de procesamiento del lenguaje o la elaboración de modelos estadísticos.
2. El canal de producción de los textos recogidos (orales frente a escritos). La complejidad de tratamiento y recogida de los textos del corpus está directamente relacionada con el canal por el cual estos se han producido.
3. El contenido de los corpus (de lengua general frente a sublenguajes determinados). Un corpus que abarque la lengua general también puede ser utilizado para usos específicos y un corpus de un determinado sublenguaje puede usarse para usos generales.

Así pues, de los principales trabajos para abordar una taxonomía de corpus, nos parecen destacables las aportaciones de Laviosa, por su exhaustiva concreción del marco teórico, las de Granger, por su idoneidad para el trabajo con corpus en la enseñanza y aprendizaje de una lengua, y las de Cabré por su orientación funcionalista y sus grandes aportaciones en el ámbito de la lexicología y la lexicografía.

De esta manera, según los criterios establecidos por Laviosa, Granger y Cabré, el corpus de latín que nos proponemos construir ha de ser un corpus general monolingüe y comparable, compuesto por textos escritos y con una triple finalidad: que sirva para la mejora de la investigación, para la enseñanza y aprendizaje del latín y para el perfeccionamiento de las herramientas lexicográficas.

2. Especificaciones y diseño del CLARE

Como afirman los estudiosos especializados en la explotación del corpus, entre los que cabe destacar a Partington (1998) y, sobre todo, McEnery, Xiao y Tono (2006), en

la etapa de planificación del corpus hay que definir toda una serie de especificaciones divididas en dos ejes principales: el diseño lingüístico del corpus y la planificación del proyecto en su totalidad. Siguiendo pues estas premisas, las etapas o fases previstas para elaborar un corpus electrónico son las siguientes:

- Una primera fase de compilación del material, en la que distinguiremos dos tareas: la elección y el etiquetado de los textos.
- Una segunda fase que McEnery, Xiao y Tono (2006) denominan de análisis de los datos obtenidos.

En el presente trabajo abordaremos las dos tareas iniciales de la fase de compilación y dejaremos el análisis de los datos para posteriores entregas, dada la envergadura de su exposición.

2.1. Elección del corpus

El trabajo de recopilación de textos para su explotación posterior en formato electrónico conlleva una serie de dificultades entre las que Rabadán y Fernández Nistal (2002, 56) señalan la «inversión de tiempo, esfuerzo y recursos humanos», el hecho de que «la disciplina ha evolucionado vertiginosamente a lo largo de los últimos años» y «la diversidad de parámetros» que determinan el éxito de un proyecto basado en corpus. Por ello, y como advierten Rabadán y Fernández Nistal (2002, 57), «con el objeto de evitar fracasos innecesarios, antes de iniciar la construcción del corpus es imprescindible abordar toda una serie de cuestiones».

En primer lugar hay que reflexionar sobre el tamaño del corpus. Inicialmente se pensaba que cuanto más grandes fueran los corpus más información proporcionarían a los usuarios. No obstante, el propio Sinclair (1991, vii) afirma que un mayor tamaño no garantiza el éxito de la investigación pues «[quantity] has nothing to do with the quality».

Flowerdew (2004, 11) y Olohan (2004, 45) aconsejan ajustar el tamaño de los corpus a las necesidades de los usuarios y vinculan el tamaño del corpus a su representatividad. Es decir, los corpus han de adquirir el tamaño que los haga representativos para aquello que pretenden confirmar o refutar, y por tanto, como comentan Rabadán y Fernández Nistal (2002, 58), «tenemos que saber muy claramente qué es lo que buscamos».

Según Biber (1993, 243) la representatividad hay que entenderla como «the extent to which a sample includes the full range of variability in a population». Y aunque la mayor parte de los investigadores reconocen que es muy complicado lograr este propósito, Bowker y Pearson (2002) proponen la consideración de los siguientes aspectos para reforzar la representatividad de un corpus: el tamaño, el número de textos incluidos, el medio (preferentemente escrito), los temas objeto de estudio, una tipología textual homogénea y ponderada, la autoría, la lengua objeto de estudio y la fecha de publicación o emisión de los textos.

Para la descripción de las etapas de compilación del corpus, usaremos el trabajo de Flowerdew, quien propone una serie de preguntas para orientar a los creadores de corpus especializados y evitarles, de esta manera, los fracasos innecesarios ante los que nos previenen Rabadán y Nistal (2002). Según Flowerdew (2004, 25-27), para compilar un corpus el investigador debe realizarse una serie de preguntas previas que se concretan en las siguientes:

1) What is the purpose for building a corpus?

La primera cuestión es preguntarse por la finalidad del corpus. En nuestro caso y, siguiendo también los planteamientos de Cabré (1997) y de Torruella y Llisterri (1999, 15), comenzaremos diciendo que nuestro objetivo principal es confeccionar un corpus de referencia del latín. Éste se concibe a imagen y semejanza de los grandes corpus de referencia (BNC, BoE, CREA, CORDE, etc.) y con él se persiguen varias finalidades:

1. Que se convierta en una herramienta más para la investigación y el estudio del latín junto a las ya existentes.
2. Que constituya el embrión de un corpus general de carácter abierto, o *monitor corpus*, con el que poder contrastar y refrendar los diferentes corpus especializados que ya existen en la actualidad.
3. Que sirva también para la enseñanza y el aprendizaje del latín.

2) What genre is to be investigated?

La adecuación textual del corpus a un proyecto concreto, como señala Sánchez (1995, 41), implica tener en cuenta que el tipo de textos que contenga el corpus esté bien delimitado. El CLARE nace como un *monitor corpus* con vocación de convertirse en un corpus de referencia de latín. Para este propósito, como ya hemos comentado, la recopilación textual no se limita a un género o época, sino que por razones de representatividad y equilibrio, el corpus dispondrá de la mayor cantidad y variedad de textos electrónicos en latín que puedan ser capturados desde las bibliotecas y bases de datos seleccionados.

La historia de la lengua latina abarca 2000 años de producción textual, aunque la tradición ha identificado el latín fundamentalmente con las obras de los autores del llamado periodo clásico (ss. I a.n.e. - I n.e.). Nuestra investigación propone superar esta visión reduccionista ampliando los límites del corpus a todo tipo de producción textual en lengua latina, sin jerarquizar las muestras desde una perspectiva de género literario y sin desestimar la información lingüística que nos pueden brindar textos como el recetario de Apicio, el libro de arquitectura de Vitrubio o el manual de enseñanza de Comenius.

3) How large is the corpus supposed to be?

Respecto al tamaño del corpus Bowker y Pearson (2002, 54) consideran que los corpus comprendidos entre «a few thousand to a few hundred thousand words have proved useful for LSP purposes», y Flowerdew (2004, 19) opina que los corpus comprendidos entre 20.000 y 200.000 palabras ya arrojan resultados interesantes sobre la lengua estudiada.

Evidentemente el CLARE, como tantos otros, será un corpus abierto a las aportaciones que, sin duda, seguirán produciéndose en la medida que los equipos de investigación dispongan de aplicaciones más avanzadas y fiables para la edición electrónica de textos. Será una tarea de constante crecimiento con la incorporación progresiva de muchas obras que, hasta hoy, solo están disponibles en formatos de difícil manipulación electrónica. En la actualidad el CLARE está formado por unos 2.471 ficheros de texto y un total de unos 14 millones de palabras.

4) How will data be collected?

Trabajar con textos en formato electrónico permite menor esfuerzo en la recopilación de textos que si los estuviéramos digitalizando a partir de un formato en papel. Esta metodología de compilación facilita una mayor concentración en el tratamiento posterior de los mismos. Para nuestra investigación hemos utilizado el conjunto de textos disponibles en formato electrónico HTML o TXT. Existe en la red una gran cantidad de documentos escaneados en formato JPG, que podrían ser tratados con aplicaciones OCR. Pero tras tantos siglos de transmisión textual, las diferentes caligrafías existentes para el latín (carolingia, itálica, gótica, etc.) plantean problemas de reconocimiento de caracteres, lo cual origina innumerables problemas de postedición.

Por ello, de momento, la disponibilidad de textos ya editados por reconocidos proyectos de investigación y puestos a disposición de la comunidad científica por medio de bibliotecas digitales, nos ha llevado a confeccionar nuestro corpus con este tipo de textos.

5) How will the (specialized) corpus be tagged / marked up?

Existe un consenso generalizado acerca de las ventajas del etiquetado contextual en SGML (Standard Generalized Markup Language) y, sobre todo en la actualidad, en XML (eXtensible Markup Language). Como es bien sabido, el XML es un metalenguaje creado por el World Wide Web Consortium (W3C) que permite el intercambio de información estructurada entre diversas plataformas. En la práctica, los creadores de corpus lo utilizan, entre otras cosas, para enriquecer sus textos con información metatextual que permita incrementar la dificultad de las consultas de los corpus.

En nuestro caso, consideramos como Leech (2002) que un corpus de referencia debe estar lo menos etiquetado posible con el fin de manchar lo menos posible el corpus, de manera que facilite futuras modificaciones, incorporaciones o explotaciones textuales. Por esta razón, elaboraremos dos corpus: uno sin etiquetar, a disposición de futuras investigaciones y de programas informáticos aún hoy desconocidos, y otro con etiquetas metatextuales que permitan explotar el corpus y generar subcorpus para la investigación y la enseñanza.

6) What kind of reference corpus would be suitable to contrast with the specialized corpus?

Confrontar los resultados de un corpus específico con otro general o referencial no sólo es conveniente sino totalmente imprescindible para calibrar la repercusión estadística de los datos que arroje el análisis de los corpus. Y es este argumento la idea capital para el desarrollo de nuestro proyecto, al no existir en la actualidad ningún corpus de referencia del latín con el que poder contrastar los datos obtenidos con los diferentes corpus específicos de latín.

Así pues, como ya hemos avanzado, según los criterios establecidos por Laviosa, Granger y Cabré, el corpus de latín que nos proponemos construir ha de ser un corpus general monolingüe y comparable, compuesto por textos escritos y con una triple finalidad: que sirva para la mejora de la investigación, para la enseñanza y aprendizaje del latín y para el perfeccionamiento de las herramientas lexicográficas.

No obstante, dejaremos para futuras investigaciones, como ya hemos comentado, el completar las restantes etapas previstas por el método de Calzada (2007) y, como

propone Cabré, no hay que descartar que pueda ser explotado con fines específicos o generales, en función de las distintas necesidades de los potenciales usuarios.

2.2. Procesamiento del CLARE

En la etapa de procesamiento del corpus hay que considerar tres fases bien diferenciadas: un primer momento de captura de los textos, una segunda fase de edición y una tercera fase de etiquetado. Pasamos, a continuación, a describir cómo han sido estas etapas en la construcción del CLARE.

Para la captura de los textos hemos utilizado HTTrack en su versión 3.47-21 (7/05/2013), una aplicación informática de Software libre desarrollada por Xavier Roche, con licencia GPL, multilenguaje y multiplataforma, que sirve para la captura de sitios web, es decir, para la descarga a un sistema de almacenamiento de todo o parte de un sitio web, para poder navegar en cualquier momento sin necesidad de estar conectado a internet. Permite copiar todas las páginas de un mismo sitio en pocos minutos. Finalmente el programa descarga los ficheros seleccionados en una carpeta del PC o disco duro, dentro de la cual los archivos se organizan en carpetas que contienen las obras de cada autor, si se organizan en diferentes archivos, o bien fuera de carpetas, cuando la obra comprende un solo archivo.

Una vez capturados los textos en formato electrónico, hemos trabajado con un editor de textos, en nuestro caso jEdit versión 5.1.0 (28/07/2013), que se puede descargar gratuitamente desde su página <http://www.jedit.org/>. jEdit es un editor de texto libre, creado por Slava Pestov en 1998 y distribuido bajo los términos de la licencia pública general de GNU. Está escrito en Java y se ejecuta en Windows, Linux, Mac y cualquier otro sistema operativo que disponga de la máquina virtual Java. Dispone de docenas de *plugins* para diferentes áreas de aplicaciones. Soporta de forma nativa el resaltado coloreado de la sintaxis para más de 200 formatos de fichero. También se pueden incluir nuevos formatos de forma manual utilizando ficheros XML. jEdit trabaja tanto con UTF-8 como con otros formatos de codificación del texto.

Nuestro corpus no refleja el latín de un autor, una obra o una etapa, sino que, como ya hemos explicado, contiene textos de todas las épocas y lugares en los que se realizaron producciones textuales en lengua latina. Para su procesamiento hemos generado ocho etiquetas metatextuales que responden a los siguientes parámetros:

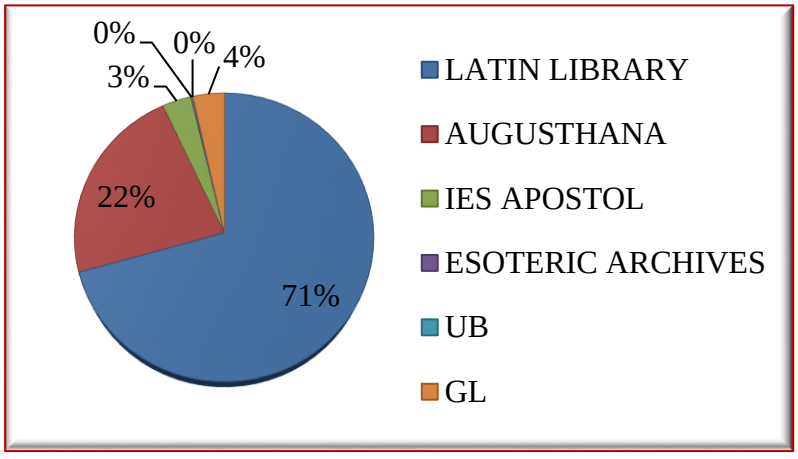
<bibliotheca> <genre> <latin> <time> <place> <gender> <author> <title>

La primera etiqueta, <BIBLIOTHECA>, hace referencia a las diferentes bibliotecas electrónicas de las que hemos extraído los textos del CLARE. Éstas aparecen reseñadas con las siguientes siglas:

- TLL, The Latin Library. Esta es la biblioteca que más textos aporta a nuestro corpus. Entre las ventajas más destacables podemos citar que se trata de una biblioteca exclusivamente de textos latinos, a diferencia de otras que presentan también griegos, germanos, etc. Es muy útil por disponer de un grafismo minimalista que imita la página del libro (negro sobre blanco) y por carecer de imágenes, que complican normalmente la descarga de los textos. Entre los inconvenientes, cabe decir que no siempre informa de la edición utilizada y que los estándares informáticos no son universales, sino que presenta diferentes formatos de edición, lo cual crea numerosos problemas de ruido en el corpus.

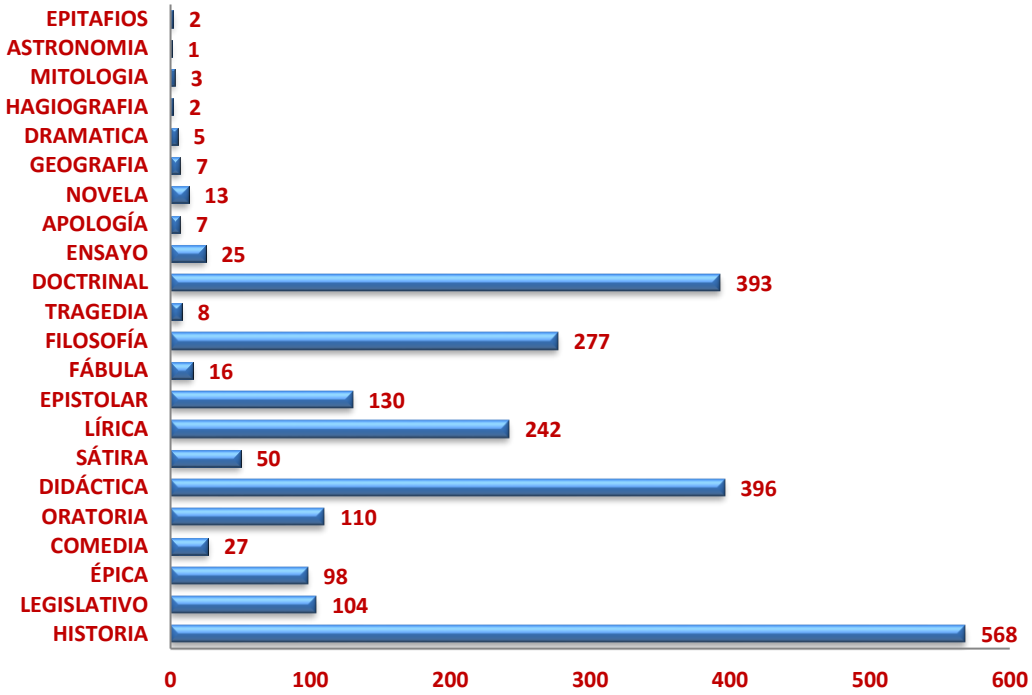
Esta circunstancia impone un largo trabajo de postedición para unificar los formatos en HTML.

- AUG, la Bibliotheca Augustana, de Ulrich Harsch. Esta es la segunda biblioteca que más textos aporta al CLARE. Sus principales ventajas son que ofrece obras de autores poco frecuentes en los demás bancos textuales, especialmente los de origen nórdico y germano. Entre los inconvenientes, hay que destacar que la biblioteca latina es una más entre las otras de este sitio web, y que presenta múltiples imágenes y un formato muy elaborado que provoca problemas de codificación en la descarga de los textos y, consecuentemente, obliga también a largos trabajos de postedición.
- IA, IES Apostol. De ella hemos obtenido la *Minerva* del Brocense, gracias al trabajo de Carlos Cabanillas. Hemos tenido que editar el texto con jEdit, pues presentaba muchos problemas de ruido añadido, pero aun así, nos resultaba altamente interesante, pues viene a cubrir una laguna en el déficit de autores de procedencia hispana que hemos detectado en la red en formato electrónico HTML.
- EA, Esoteric Archives. No se trata de una biblioteca general sino de un sitio web que recoge textos de literatura de tema esotérico. Lo hemos utilizado para obtener las obras de Giordano Bruno, que no estaban disponibles en las anteriores bibliotecas. Presenta ciertos problemas de conversión de su UTF y abundante ruido, pero hemos decidido utilizarla por ser de las pocas que contienen el texto de Bruno en formato HTML.
- UB, se trata de la base de datos Ramon Llull de la Universidad de Barcelona: <http://orbita.bib.ub.es/llull/velec.asp>. Es, como reza la propia página, un instrumento bibliográfico electrónico concebido para ordenar, sistematizar y facilitar la consulta exhaustiva de toda la información referente a la obra adjudicada a Llull. Ha sido creada por Anthony Bonner.
- GL, Grexlat es una biblioteca de textos latinos de la Universidad de Kentucky: <http://www.grexlat.com/biblio/index.asp>. De esta web hemos tomado los textos del *Orbis pictus*, de Comenius, y las *Exercitationes* de Luis Vives. El diseño de la página nos ha obligado a limpiar el ruido con el editor de textos jEdit.

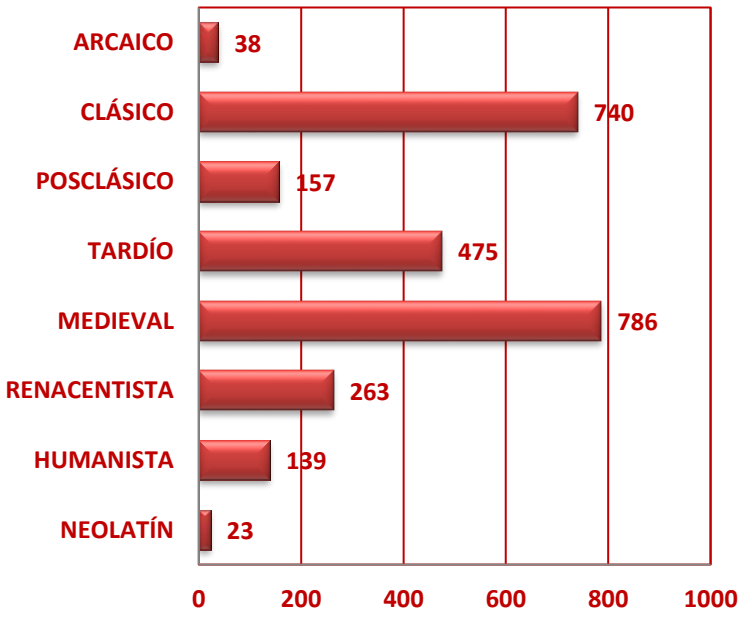


La segunda etiqueta, <GENRE>, hace referencia al contenido del texto. Como ya hemos comentado, nuestro corpus no refleja solamente textos de carácter literario sino

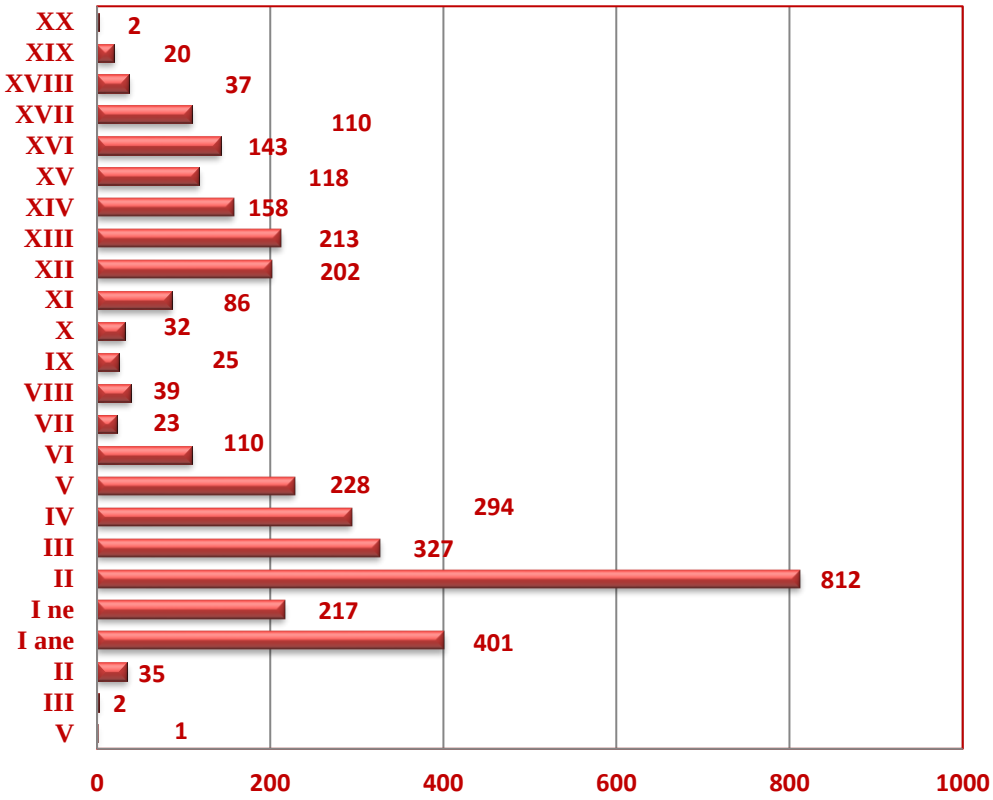
también todo el conjunto de escritos en formato electrónico disponible en las bibliotecas de referencia.



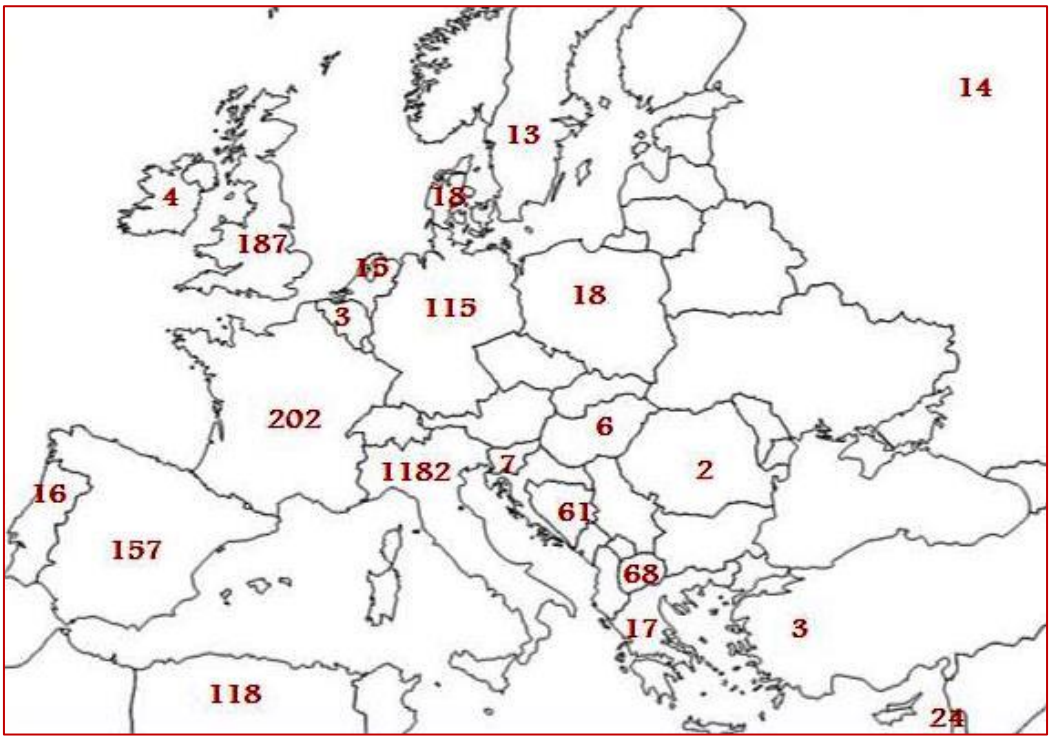
La tercera etiqueta, <LATIN>, hace referencia a los diferentes registros lingüísticos. Los subapartados que recogerá esta etiqueta serán, por tanto, los siguientes: arcaico, clásico, posclásico, medieval, humanístico, renacentista y neolatín.



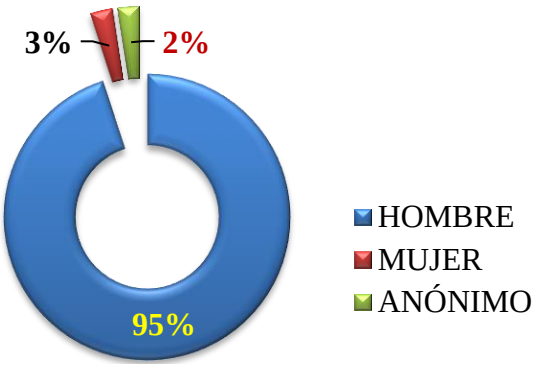
La cuarta etiqueta, <TIME>, hace referencia a la información cronológica, siempre que la fidelidad de las fuentes filológicas lo permiten. Puesto que para muchas obras la referencia temporal es aproximativa, hemos preferido reseñar el siglo y no el año.



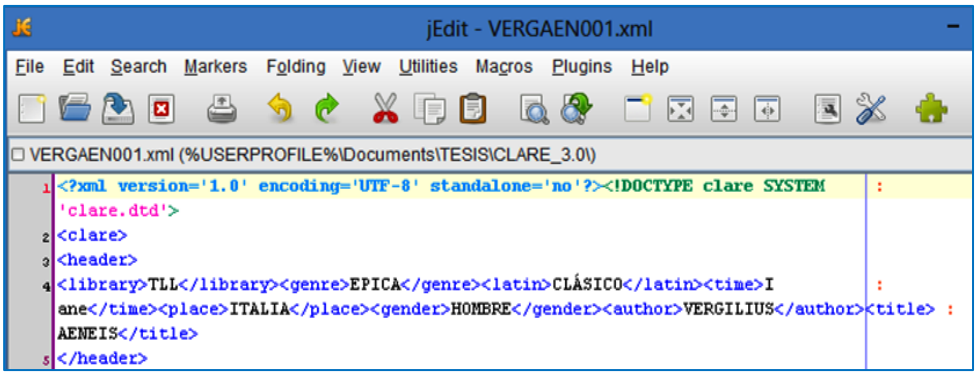
La quinta etiqueta <PLACE>, hace referencia al lugar geográfico al que se asocia la obra y/o el autor. Cuando esta circunstancia no está suficientemente acreditada, o cuando se desconoce la autoría, hemos preferido utilizar la etiqueta «unknown» (208). La distribución actual de textos por países se muestra en la siguiente ilustración:



La sexta etiqueta, <GENDER>, hace referencia al autor o autora de la obra. En este apartado sólo se dan tres opciones: «hombre», «mujer» y «anónimo». Dada la tradición literaria latina, en este apartado el corpus está desequilibrado, pues en la actualidad sólo hemos podido acreditar textos de autoría femenina en 9 casos.



La séptima etiqueta, <AUTHOR>, reseña el nombre original del autor o autora. Cuando el texto es anónimo o se desconoce la autoría, hemos preferido utilizar la etiqueta «unknown». La octava etiqueta, <TITLE>, reproduce el título original de la obra en latín. A continuación vemos un ejemplo que ilustra las etiquetas comentadas:



Las etiquetas metatextuales descritas nos permitirán efectuar búsquedas de textos por cualquiera de los parámetros expuestos: época, lugar de producción, tema del escrito, autoría, género, tipo de latín, etc. Con ellas se podrán realizar múltiples tareas, entre otras:

- elaborar listas de frecuencia;
- contrastar los resultados de corpus ad hoc con un corpus general de latín;
- elaborar subcorpus;
- explotar de forma paralela los textos del CLARE por épocas, lugar, género, etc.;
- rastrear variantes diatópicas, diafásicas, diacrónicas o diastráticas en los textos de autoría anónima;
- verificar hipótesis gramaticales.

3. Conclusiones

En el presente artículo hemos expuesto los criterios con los que hemos diseñado la construcción del CLARE y los parámetros con los que proponemos implementar las posibilidades de crecimiento de este *monitor corpus*, de carácter abierto y colaborativo,

con la incorporación de nuevos textos en formato electrónico que se ajusten a los criterios técnicos y filológicos propuestos.

Para ello nos hemos servido de las propuestas metodológicas de la Lingüística de Corpus y, en especial de las aportaciones de Flowerdew, y las orientaciones funcionalistas de Laviosa, Granger y Cabré.

Dejamos, pues, para próximas comunicaciones la presentación y descripción de los resultados que se deriven de su aplicación en el campo de la lexicografía, de la investigación y de la enseñanza y aprendizaje del latín.

Bibliografía

D. BIBER (1993), «Representativeness in Corpus Design», *Literary and Linguistic Computing*, vol. 8 (4), pp. 243-257.

L. BOWKER; J. PEARSON (2002), *Working with Specialized Language. A Practical Guide to Using Corpora*, London, Routledge.

M. CALZADA (2007), *Proyecto investigador para la habilitación de cátedras*, Castellón de la Plana, inédito.

L. FLOWERDEW (2004), «The argument for using English specialized corpora to understand academic and professional language», en U. Connor - T. A. Upton (eds.), *Discourse in the Professions. Perspectivas from Corpus Linguistics*, Amsterdam-Filadelfia, John Benjamins, pp. 11-36.

C. GELPÍ (1997), *Mesures d'avaluació lexicogràfica de diccionaris bilingües*, Barcelona, Universitat de Barcelona.

S. GRANGER (2003), «The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies», en S. Granger, J. Lerot J. y S. Petch-Tyson (eds), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Amsterdam - New York, Rodopi, pp. 17-30.

S. LAVIOSA (2002), *Corpus-based Translation Studies: Theory, Findings, Applications*, Amsterdam - New York, Rodopi.

G. LEECH (2002), «The importance of reference Corpora», en el congreso de *Corpus Lingüísticos. Presente y futuro*, Donostia, UZEI. <https://www.yumpu.com/en/document/view/36035492/the-importance-of-reference-corpora-uzei> [fecha de consulta: 21/2/2016]

A. MCENERY; Z. XIAO; Y. TONO (2006), *Corpus-based Language Studies*, London, Routledge.

M. OLOHAN (2004), *Introducing Corpora in Translation Studies*, London - New York, Routledge.

A. PARTINGTON (1998), *Patterns and Meanings: Using Corpora for English Language Research and Teaching*, Amsterdam-Philadelphia, John Benjamins.

R. RABADÁN; P. FERNÁNDEZ NISTAL (2002), *La traducción inglés-español: fundamentos, herramientas y aplicaciones*, León, Universidad de León.

R. ROBERTS (2006), «Corpora and Translation», en L. Bowker (eds.), *Lexicography, Terminology and Translation*, Ottawa, University of Ottawa Press, pp. 201-214.

A. SÁNCHEZ *et alii* (1995), *Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*, Madrid, SGEL.

J. SINCLAIR (1991), *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

J. TORRUELLA; J. LLISTERRI (1999), «Diseño de corpus textuales y orales», en AA.VV. *Filología e informática. Nuevas tecnologías en los estudios filológicos*, pp. 45-77, Barcelona, Milenio.

Curriculum

Mercedes García Ferrer es licenciada en Filología Clásica por la Universidad de Salamanca y DEA en traducción por la UJI. Como docente, ha enseñado latín en IES de Castellón, así como en la Universidad del País Vasco y en la Universidad Jaume I. Como investigadora, ha trabajado en el ámbito de la lexicografía didáctica con la editorial SM, y en la elaboración de numerosos materiales para la enseñanza del latín.