Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 57-64 recibido 27-03-2015 revisado 28-04-2015 aceptado 19-05-2015

P. S. Post Scriptum: Dos corpus diacrónicos de escritura cotidiana*

P. S. Post Scriptum: Two Diachronic Corpora of Ordinary Writing

Gael Vaamonde
Centro de Linguística da Universidade de Lisboa (CLUL)
Av. Prof. Gama Pinto, 2. 1649-003 Lisboa - Portugal
gaelvmnd@gmail.com

Resumen: En este trabajo se da a conocer el proyecto de investigación *P. S. Post Scriptum*, que tiene por objeto la búsqueda sistemática, edición y estudio histórico-lingüístico de cartas privadas escritas en España y Portugal durante la Edad Moderna. Estas cartas constituyen manuscritos inéditos escritos por personas de muy diferente condición social y suelen presentar una retórica cercana a la oralidad, tematizando asuntos de lo cotidiano. Son, por tanto, de gran interés para la investigación en lingüística diacrónica. La finalidad del proyecto es publicar y estudiar 7000 de estas cartas, ofreciendo una edición crítica digital del manuscrito y, simultáneamente, convirtiendo el contenido de las cartas en dos corpus anotados de un millón de palabras cada uno: uno para el español y otro para el portugués.
Palabras clave: Lingüística de corpus, lingüística histórica, español, portugués, cartas.

Abstract: In this paper, we present an overall description of *P. S. Post Scriptum*. Within this research project, systematic research will be developed, along with the publishing and historical-linguistic study of private letters written in Portugal and Spain along the Modern Ages. The letters included in *P. S. Post Scriptum* are unpublished manuscripts, written by authors from different social backgrounds. In addition, these textual resources often present an (almost) oral rhetoric, treating everyday issues of past centuries. They are, therefore, of great interest for research in Diachronic Linguistics. We aim to publish and study 7,000 of those letters. For this purpose, we are preparing a scholarly digital edition of the manuscripts and, simultaneously, converting the content of the letters into two annotated corpora of a million words each, one containing the Portuguese letters, the other the Spanish.
Keywords: Corpus Linguistics, Diachronic Linguistics, Spanish, Portuguese, Letters.

1 Introducción

La investigación en lingüística histórica no ha sido ajena al desarrollo de las nuevas tecnologías informáticas, beneficiándose en las últimas décadas –como no podía ser de otro modo– de las enormes ventajas que ofrecen los corpus en formato electrónico, que permiten almacenar y procesar grandes cantidades de datos lingüísticos de manera rápida y eficaz.

En este sentido, se puede afirmar que el español es una lengua privilegiada, ya que cuenta con dos grandes corpus históricos de acceso libre en red: el *Corpus Diacrónico del Español* (CORDE) y el *Corpus del Español* de Mark Davies (CdE). El gran volumen de texto recopilado –250 millones de palabras y 100 millones de palabras, respectivamente– los convierte en verdaderas herramientas de referencia para la investigación diacrónica en esta lengua.

* El proyecto de investigación *P. S. Post Scriptum* está siendo financiado por el Consejo Europeo de Investigación (7FP/ERC Advanced Grant – GA 295562)

ISSN 1135-5948

Por otro lado, en el ámbito hispánico han ido apareciendo recientemente otros corpus diacrónicos más especializados que, a expensas de reducir el tamaño de la muestra, permiten mejorar algunos aspectos, como son transcripciones paleográficas uniformes o la posibilidad de acceso a los facsímiles. Entre ellos cabe citar el proyecto *Biblia Medieval*[1] (Enrique-Arias, 2010), un corpus paralelo de cinco millones de palabras con las traducciones de la Biblia al castellano producidas durante la Edad Media, y el corpus CODEA[2] (Sánchez-Prieto et al., 2009), que consta de 1500 documentos anteriores al siglo XVIII editados según la triple presentación propuesta por la red CHARTA[3]: transcripción paleográfica, presentación crítica y facsímil.

Siguiendo esta línea de corpus diacrónicos especializados se sitúa el proyecto que presentamos en este trabajo: *P. S. Post Scriptum. Archivo digital de escritura cotidiana en la Edad Moderna*[4]. El objetivo de este proyecto es la creación de dos corpus compuestos por cartas privadas, uno para el español y otro para el portugués, junto con su edición crítica digital. El marco cronológico estudiado comprende desde el siglo XVI hasta el primer tercio del siglo XIX y el tamaño del corpus alcanza un total de 3500 cartas (un millón de palabras, aproximadamente) para cada lengua.

La idea que motivó la creación de *P. S. Post Scriptum* partió de una posibilidad excepcional para recuperar este tipo de material epistolar. Los tribunales de la Edad Moderna, tanto civiles como inquisitoriales, utilizaban la correspondencia privada como una prueba instrumental para condenar o exonerar a sus autores, a sus destinatarios o a otras personas relacionadas o mencionadas en el contenido de las misivas. Por tanto, buena parte de esta documentación se conservó hasta nuestros días archivada en el interior de procesos judiciales de la época.

Las cartas, en su mayoría inéditas, fueron escritas por gente de muy diversa índole, generalmente manos poco instruidas, y suelen reflejar una retórica cercana a la oralidad, ofreciendo así una ventana a variedades lingüísticas del español y del portugués que no suelen tener cabida en los corpus de corte diacrónico, compuestos predominantemente por textos de carácter literario o notarial. En otras palabras, la naturaleza dialógica y coloquial de estas misivas permite compensar, en su justa medida, la carencia de fuentes orales.

P. S. Post Scriptum es un proyecto interdisciplinar formado por lingüistas e historiadores españoles y portugueses. En este trabajo explicamos la metodología de trabajo, desde la búsqueda de los manuscritos hasta la publicación en línea de los textos, y ofrecemos el estado actual del proyecto, que finalizará en 2017.

2 Antecedentes

P. S. Post Scriptum constituye una continuación de un proyecto anterior, llamado *CARDS. Cartas Desconhecidas*. Este proyecto se centró en la recopilación y edición electrónica de cartas privadas portuguesas anteriores a 1900. El corpus pretendido en CARDS ascendía a 2000 cartas. En términos cuantitativos, por tanto, el objetivo de *P. S. Post Scriptum* es completar el corpus portugués con 1500 cartas y crear desde el inicio el corpus epistolar español.

3 Búsqueda en archivos

El primer paso en *P. S. Post scriptum* consistió en la localización, recopilación y digitalización de los manuscritos. Esta tarea fue central en los primeros años del proyecto (2012-2014) y está prácticamente concluida en el momento de redactar estas líneas. Para la localización de las cartas, se han consultado –y se están consultando– fondos judiciales (civiles y criminales), eclesiásticos e inquisitoriales a lo largo de toda la Península Ibérica.

En el caso del español, se han examinado fondos en el Archivo Histórico de Asturias, el Archivo de la Real Chancillería de Valladolid, el Archivo General de Simancas, el Archivo de la Real Chancillería de Granada, el Archivo Histórico Nacional, el Archivo General de la Corona de Aragón, el Archivo Histórico del Reino de Galicia y el Archivo General de Indias, además de varios archivos provinciales y diocesanos (Murcia, Pontevedra, Orense, Toledo, Barcelona, Guadalajara, Cuenca, Sevilla y Zaragoza).

En el caso del portugués, la documentación inquisitorial está concentrada en el Arquivo Nacional Torre do Tombo. Además del trabajo

[1] http://www.bibliamedieval.es/
[2] http://demos.bitext.com/codea/
[3] http://www.charta.es/
[4] http://ps.clul.ul.pt/index.php

continuado en este archivo, también se han consultado fondos en el Arquivo Distrital do Porto, el Arquivo Histórico Militar, El Arquivo Distrital de Braga o el Arquivo Histórico Ultramarino, entre otros.

La tarea de archivo no solo consistió en identificar las cartas, sino también en extraer toda una serie de metadatos relacionados con la producción del texto (fecha, lugar de origen y destino, descripción física del manuscrito, etc). Generalmente, la lectura atenta del proceso permite contextualizar la situación comunicativa de la carta, así como trazar un perfil biográfico de autores y destinatarios. Estas fichas biográficas están siendo almacenadas en una base de datos independiente, cuya información es posible cruzar con los datos del corpus.

El número de archivos y fondos que se han consultado es amplio y variado. En términos históricos y culturales, esta variedad permite obtener un panorama más completo de las sociedades tradicionales y de las relaciones interpersonales en la Edad Moderna, reflejadas en los contextos históricos que acompañan a cada carta o conjunto de cartas relacionadas. En términos lingüísticos, supone el control de un espacio más ampio y, por tanto, la posibilidad de incluir autores de diversa procedencia geográfica, lo que se traduce en un corpus dialectalemente más rico y representativo.

4 Transcripción en XML-TEI

Una vez localizadas las cartas, el siguiente paso es transcribirlas con el objeto de ofrecer una edición crítica digital del manuscrito, esto es, una transcripción paleográfica del texto en edición electrónica que conserve rigor filológico. Para tal fin, se han tomado algunas decisiones de carácter técnico.

Se ha utilizado el lenguaje de marcación XML (eXtensible Markup Language). Los ficheros XML son legibles, sin pérdida de información, por todos los procesadores de texto, lo que facilita su conversión para otros formatos y evita problemas de procesamiento electrónico. Por otro lado, y en consonancia con las prácticas actuales en el campo de las Humanidades Digitales, se han adoptado los estándares de codificación propuestos por el consorcio TEI (*Text Encoding Initiative*) para la edición de textos en formato digital[5]. El consorcio TEI es una convención ya consolidada en la edición virtual de fuentes primarias, lo que garantiza la integración con otros corpus electrónicos de naturaleza similar.

Conviene apuntar que al inicio del proyecto, en 2012, el consorcio TEI todavía no había proporcionado estándares de codificación para la publicación digital de material epistolar. Por este motivo, en un primer momento se adoptó la propuesta de codificación del proyecto DALF (*Digital Archive of Letters in Flanders*), que está a su vez basada en una versión no estándar del consorcio TEI. Actualmente, el modelo XML-TEI que se ofrece en *P. S. Post Scriptum* está basado en dos fuentes: la propuesta de la Red CHARTA (*Corpus Hispánico y Americano en la Red: Textos Antiguos*) y la propuesta del módulo TEI-CORRESP-SIG para material epistolar creada por Peter Stadler, Marcel illetschko y Sabine Seifert. Ambas fuentes toman como referencia la versión más actual y estandarizada del consorcio TEI.

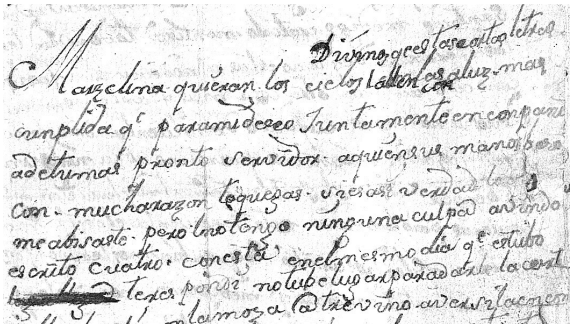

Figura 1: Facsímile de un fragmento de carta

```
<salute>Marzelina</salute> quieran los cielos <add hand="VF3"
place="supralinear">Divinos <abbr>q<expan>u</expan>e</abbr> estas cortas
letras</add> t allen <add hand="VF3" place="underlinear">con</add> la saluz mas
<lb/> cunplida <abbr>q<expan>u</expan>e</abbr> para mi deseo juntamente en
conpañi<lb n="false"/>a de tu mas pronto servidor a quien sus manos besa <lb/>
con mucha razon tu quegas si es asi verdad lo <abbr>q<expan>u</expan>e</abbr>
<lb/> me abisaste pero no tengo ninguna culpa avindo <lb/> escrito cuatro con
esta en el mesmo dia <abbr>q<expan>u</expan>e</abbr> estubo <lb/> <del
hand="VF3">la gallega</del> te respondi no tube lugar para darle la carta
```

Figura 2. Transcripción XML-TEI

Para la transcripción del manuscrito se ha adoptado una actitud conservadora. Tan solo se ha normalizado la segmentación de palabras y el uso de las grafías «i», «j», «u» y «v». Los cambios de línea, la ortografía, las abreviaturas, los tachones, las correcciones del autor, los accidentes del soporte o la orientación de la

[5] http://www.tei-c.org/index.xml

escritura, entre otros aspectos, se han respetado en la edición digital. Como ejemplo de los dicho, sirva el ejemplo recogido en las figuras 1 y 2, que permite comparar el facsímile con la transcripción en XML. En el ejemplo de la figura 2, los elementos XML </lb>, <del>, <add> y <abbr> permiten marcar cambios de línea, tachones, añadidos autoriales fuera de línea y abreviaturas, respectivamente.

5 Tratamiento lingüístico del corpus

Las tareas de transcripión y edición crítica digital forman parte del objetivo filológico del proyecto *P. S. Post Scriptum*. El otro objetivo fundamental es de carácter lingüístico: se trata de ofrecer dos corpus anotados, uno para el español y otro para el portugués.

Este segundo objetivo consta de tres tareas fundamentales: la tokenización del texto, la normalización de la grafía y la anotación lingüística de cada token normalizado. En principio, se ha contemplado para la finalización del proyecto la estandarización y anotación morfosintáctica de todo el corpus (i.e. etiquetado de clases de palabras) y la anotación sintáctica de, al menos, un subconjunto de los datos.

Desde finales de 2014, todas las tareas de tratamiento lingüístico del corpus están centralizadas en TEITOK, un sistema en línea creada por Maarten Janssen[6]. TEITOK fue diseñado para poder compatibilizar en un mismo conjunto de datos XML tanto la transcripción paleográfica como la anotación lingüística, respondiendo así a las demandas de *P. S. Post Scriptum*; cumple además con un doble objetivo: para los miembros del proyecto, funciona como ambiente de trabajo, permitiendo insertar o modificar cualquier información en los diferentes niveles de edición del texto; para el usuario, funciona como interfaz de consulta, facilitando la búsqueda cruzada de los datos que ya hallan sido almacenados[7]. A continuación, se explican brevemente cada una de las tareas de edición del texto para su procesamiento lingüístico.

5.1 Tokenización

Una vez que las cartas son transcritas mediante XML, se importan a la interfaz de TEITOK, en donde se procede al tratamiento lingüístico del texto. El primer paso es la tokenización, que se realiza de manera automática. Durante el proceso de tokenización, cada forma original de la palabra es marcada dentro de un elemento <tok>, al que se le asigna una identificación única también de manera automática.

Esta estructura inicial permite separar cada token para su posterior edición lingüística y permite salvaguardar además los diferentes niveles de edición, que se van almacenando en forma de atributos dentro de cada elemento <tok>. Por ejemplo, la forma *Otbre* como abreviatura de *octubre* en el manuscrito original sería procesada en TEITOK del modo siguiente:

```
<tok id="w-144" form="Otbre" fform="Otubre" nform="octubre">Otbre</tok>
```

Figura 3. Ejemplo de token en TEITOK

Los atributos "form", "fform" y "nform" señalan la forma original, la forma expandida y la forma normalizada de la palabra, respectivamente. Otros niveles de edición, como pueden ser variantes dialectales, información metalingüística, lemas o etiquetas morfosintácticas, también son añadidos de forma correlativa mediante atributos dentro de <tok>. Esta estrategia permite mantener siempre una vinculación entre los diferentes niveles para su posterior recuperación a través del motor de búsqueda de la interfaz.

5.2 Normalización ortográfica

Es obvio que los manuscritos originales de las cartas presentan una gran variedad ortográfica. Así, una misma palabra (p. ej. *vergüenza*) puede aparecer escrita de muy diversas formas (p. ej. *berguensa*, *verguensa*, *berguenza*, *vergüenza*, *berguença*, *verguença*, etc.). Esta diversidad tiene un interés filológico y lingüístico, principalmente para llevar a cabo estudios de carácter fonético o gráfico. Por eso, la forma original es respetada escrupulosamente y conservada en uno de los niveles de edición, como se explicó anteriormente. Esta diversidad gráfica, no obstante, constituye un problema central para la anotación automática de textos históricos (Sánchez-Marco et al., 2010). Esa es la razón principal por la que se decidió realizar una normalización ortográfica de los textos en *P. S. Post Scriptum*, que sirva como archivo de entrada para el anotador automático y maximice

[6] http://maarten.janssenweb.net

[7] Véase el apartado "Búsqueda" en la dirección electrónica del proyecto (nota 4).

su porcentaje de acierto; otra razón secundaria, además, es la posibilidad de ofrecer al público lego una edición que facilite la lectura de los textos.

En este nivel de edición, se ha normalizado la grafía y la acentuación de todas las formas originales y se ha introducido la puntuación propia de la lengua contemporánea, aunque la separación de párrafos se ha mantenido fiel al original. Conviene precisar que las modificaciones realizadas sobre el texto primario se ciñen únicamente al nivel ortográfico, por lo que no se eliminó ni se añadió ninguna palabra respecto del contenido original de la carta. Tampoco se ha intervenido sobre el nivel léxico: se han conservado los regionalismos y los arcaísmos léxicos, así como cualquier otra forma no estándar, si bien se han tratado en un nivel independiente para facilitar su recuperación.

A modo de ejemplo de normalización ortográfica, recurrimos de nuevo al fragmento ofrecido en la figura 1:

> *Marcelina, quieran los cielos divinos que estas cortas letras t' hallen con la salud más cumplida que para mí deseo, juntamente en compañía de tu más pronto servidor a quien sus manos beso. Con mucha razón te quejas, si es verdad lo que me avisaste, pero no tengo ninguna culpa habiendo escrito cuatro con esta. En el mismo día que estuvo te respondí. No tuve lugar para darle la cart' a ella.*

La edición ortográfica se está realizando de manera manual es decir, seleccionando y modificando palabra por palabra todas aquellas formas que son objeto de normalización. Aunque la interfaz de TEITOK ofrece algunas posibilidades para agilizar este proceso, se trata de una tarea que consume bastante tiempo. Por eso, en *P. S. Post Scriptum* se está trabajando actualmente en un procesamiento semiautomático de normalización. De momento, se están haciendo pruebas con la herramienta VARD 2 para el portugués (Hendrickx y Marquilhas, 2011), aunque su aplicación al proyecto todavía se encuentra en fase experimental.

5.3 Anotación lingüística

La tarea de anotación morfosintáctica ha sido objeto de un cambio de estrategia. En un primer momento, se llevó a cabo recurriendo a herramientas diferentes en función de la lengua tratada. Para el español se hizo uso del analizador automático de Freeling 3.0 (Padró y Stalinovsky, 2012) y para el portugués se utilizó la herramienta eDictor (Faria et al., 2010). En el caso de eDictor, el código de etiquetas está basado en el sistema de anotación manual utilizado por los *Penn Corpora of Historical English* (Kroch et al., 2010), ligeramente revisado para adecuarse a las características de la gramática portuguesa. En cuanto al analizador de FreeLing, el etiquetario se basa en la propuesta del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. Siguiendo esta metodología, se llegaron a anotar y revisar manualmente unas 90 cartas del corpus español y unas 890 cartas del corpus portugués. Todo ese conjunto está disponible y puede ser descargado en formato TXT desde la sección “Descargas” de la página electrónica del proyecto.

Actualmente, se está utilizando FreeLing para todo el conjunto de datos, tanto españoles como portugueses, debido a sus posibilidades de configuración, a los buenos resultados que ofrece su adaptación al portugués (García Marcos y Gamallo, 2010) y a las ventajas que conlleva el empleo de una única herramienta para las dos lenguas. Además, esta decisión coincide en el tiempo con la nueva metodología de trabajo de *P. S. Post Scriptum* a través del sistema TEITOK. Así, una vez anotado cada texto con FreeLing, el resultado es importado a esta plataforma para proceder a su revisión manual. Esta importación, que se realiza automáticamente, consiste en la adición para cada elemento <tok> de dos nuevos atributos, uno para el lema y otro para la etiqueta lingüística que sugiere FreeLing.

Finalmente, una pequeña parte del corpus portugués ya ha sido anotada sintácticamente, siguiendo el sistema de anotación de los *Penn Parsed Corpora of Historical English* (Kroch et al., 2004). Se trata de un subconjunto de 260 cartas, lo que equivale a unos 90000 tokens. Este subconjunto también está disponible para descargar desde la página electrónica de *P. S. Post Scriptum*.

Téngase en cuenta que tanto la anotación morfosintáctica como sintáctica se aplican únicamente sobre las partes no formulares del texto. Es decir, en el análisis se excluye el contenido de las abertura y cierres de las cartas así como el de los segmentos formulares que aparezcan en el cuerpo del texto (arengas y peroraciones). El objetivo es conservar únicamente aquel contenido lingüístico que haya sido lo más espontáneo posible.

6 *La base de datos biográfica*

Además del tratamiento textual en sus diferentes niveles de edición, *P. S. Post Scriptum* ofrece información extratextual de diferente naturaleza. Fundamentalmente, se está recogiendo información sobre los aspectos siguientes:

- Datos contextuales de la carta: fecha, lugar de origen y destino, resumen del contenido, contextualización.
- Datos físicos del manuscrito: descripción del soporte, medidas, grafismo, estado de conservación.
- Datos biográficos de los participantes: fecha y lugar de nacimiento, ocupación, parentesco, estado civil, religión, categoría social, etc.

Los detalles biográficos de los participantes son organizados y almacenados en una base de datos XML-TEI. Esta base de datos es en principio independiente del contenido XML de las cartas; no obstante, todos los datos almacenados se pueden relacionar entre sí a través del sistema de consulta incluido en la página electrónica del proyecto *P. S. Post Scriptum*.

Respecto al autor del manuscrito de la figura 1, sabemos que se llamaba Vicente Fernández, que era vecino de Asturias, que era labrador y que fue acusado de estupro en 1789 por el padre de la destinataria, a quien había dejado embarazada. Toda esta información, obtenida a partir del proceso o de la propia carta y debidamente catalogada, puede ser usada a voluntad del usuario, ya sea con un interés histórico y cultural, ya sea para cruzarla con los datos lingüísticos del corpus. Variables como el sexo, la edad, la categoría social o la procedencia geográfica resultan de indiscutible interés para estudios sobre dialectología o sociolingüística históricas.

7 *Resultados*

Se ofrecen a continuación los resultados alcanzados en *P. S. Post Scriptum* desde el inico del proyecto en 2012 hasta el momento actual. Por lo que respecta a la localización de los manuscritos, se decidió establecer una distribución temporal que tuviese en cuenta la realidad demográfica de cada época. Esa distribución es la que sigue:

- siglo XVI: 500 cartas
- siglo XVII: 1000 cartas
- siglo XVIII: 1500 cartas
- siglo XIX: 500 cartas[8]

Teniendo en cuenta esta referencia, los resultados obtenidos hasta la fecha son los que se muestran en la tabla 1:

	español	portugués
XVI	452	283
XVII	1172	770
XVIII	1519	1016
XIX	526	784
Total	3668	2853

Tabla 1: Cartas encontradas

Como se puede apreciar, la tarea de recopilación de las misivas está ya prácticamente rematada para el corpus español. De hecho, el número total de cartas sobrepasa el límite pretendido, aunque todavía es preciso realizar una revisión general del material para descartar documentos no originales[9]. Sin lugar a dudas, la mayor dificultad para obtener fuentes se sitúa en siglo XVI. Basándonos en nuestra experiencia en archivos históricos, podemos constatar que la documentación judicial quinientista que ha sobrevivido hasta el presente es bastante inferior a la producida en siglos posteriores, lo que reduce considerablemente la posibilidad de encontrar material epistolar.

Por lo que se refiere a la transcripción en XML-TEI y a la normalización ortográfica, los datos alcanzados hasta la fecha son los recogidos en la tabla 2:

[8] La fecha extrema con la que se trabaja es 1830, de ahí que el siglo XIX se limite a 500 cartas.

[9] Algunas copias también se transcriben si se consideran especialmente interesantes como fuentes históricas, pero nunca integran el corpus anotado.

	español	portugués
transcritas	1832	1677
normalizadas	893	1042

Tabla 2. Cartas transcritas y normalizadas

Son estas dos tareas, transcripción y normalización, las que están recibiendo mayor atención actualmente, y se espera poder finalizarlas a finales del presente año.

8 *Conclusiones y trabajo futuro*

P. S. Post Scriptum tiene como objetivo crear un recurso digital para el estudio de la escritura cotidiana en España y Portugal durante la Edad Moderna (1500-1830) que responda a los intereses de varias disciplinas: la crítica textual, la lingüística histórica y la historia cultural. Con esa pretensión, se propone reunir una amplia colección de cartas privadas, ofreciéndolas en dos formatos preparados para la búsqueda: edición crítica digital y corpus anotado lingüísticamente.

El recurso es de libre acceso y ofrece al usuario toda una serie de información textual y extratextual. Actualmente, están disponibles para su consulta los siguientes aspectos:

- Digitalización del facsímil
- Edición crítica digital
- Edición normalizada del texto
- Contextualización de las cartas
- Descripción del manuscrito
- Fichas biográficas de autores y destinatarios

Toda esta información se integra en una interfaz que facilita no solo la consulta de cualquiera de los aspectos mencionados, sino también la búsqueda cruzada de los datos. Además de la opción de consulta, el usuario puede descargar libremente los archivos XML con la transcripción (<body>) y el extratexto (<header)> de cada carta, así como los achivos TXT con la parte del corpus que ya ha sido anotada.

Entre las tareas de futuro más inmediatas hay que señalar en primer lugar la necesidad de avanzar en la anotación morfosintáctica de las dos lenguas. Esto nos permitirá contar con un conjunto de datos cada vez más amplio que pueda ser reutilizado como corpus de entrenamiento. Además, téngase en cuenta que el analizador de FreeLing fue entrenado con corpus de época contemporána, por lo que progresar en la revisión manual de nuestros datos será de gran utilidad para poder evaluar el comportamiento de esta herramienta en textos históricos normalizados. Otras tareas que merecerán nuestra atención en el futuro serán el desarrollo semiautomático de la normalización ortográfica, la traducción al inglés de al menos una parte de los textos recopilados y la anotación sintáctica de un subconjunto de los datos en ambos corpus.

Bibliografía

Davies, M. 2002. *Corpus del Español: 100 million words, 1200s-1900s*. Disponible en línea en <http://www.corpusdelespanol.org>

Enrique-Arias, A. 2010. Una nueva herramienta para la investigación de fuentes bíblicas en la Edad Media: el corpus Biblia medieval. En *Actas del XII Congreso Internacional de la Asociación Hispánica de Literatura Medieva*, páginas 85-94, Cáceres, septiembre de 2007.

Faria, P., F. Kepler y M. C. de Sousa. 2010. An Integrated Tool for Annotating Historical Corpora. En *Proceedings of the Fourth Linguistic Annotation Workshop*, páginas 217-221.

García, M. y P. Gamallo. 2010. Análise Morfossintáctica para o Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2): 59-67.

Hendrickx, I. y R. Marquilhas. 2012. From old texts to modern spellings: an experiment in automatic normalisation. *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(2). 65-76.

Kroch, A., B. Santorini y L. Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition.

Kroch, A., B. Santorini y A. Diertani. 2010. *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition.

Padró Ll. y E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. En

Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA, páginas 2473-1479, Estambul (Turquía), mayo de 2012.

Real Academia Española. Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.* <http://www.rae.es>

Sánchez-Marco, C., G. Boleda, J. M. Fontana y J. Domingo. 2010. Annotation and Representation of a Diachronic Corpus of Spanish. En *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, páginas 2713-2718, Malta.

Sánchez-Prieto B., F. Paredes García, R. Martínez Sánchez, R. Miguel Franco, M. Simón Parra e I. Vicente Miguel. 2009. El Corpus de Documentos Españoles Anteriores a 1700 (CODEA). En A. Enrique-Arias (ed.). *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid. Frankfurt am Main: Iberoamericana-Vervuert, páginas 25-38.