Post Scriptum: Archivo Digital de Escritura Cotidiana

Gael Vaamonde
Ana Luísa Costa
Rita Marquilhas
Clara Pinto
Fernanda Pratas
CLUL, Universidade de Lisboa
gaelvmnd@gmail.com

1. INTRODUCCIÓN

El proyecto de investigación *P.S. Post Scriptum: Archivo Digital de Escritura Cotidiana en Portugal y España en la Edad Moderna*[1] (en adelante, Post Scriptum) tiene como finalidad recoger y publicar cartas particulares escritas en portugués y en español durante la Edad Moderna (desde el s. XVI hasta el primer tercio del s. XIX) por personas pertenecientes a diferentes estratos sociales. Estas cartas, en su mayoría inéditas, sobrevivieron excepcionalmente al cruzarse las vidas de sus autores con los medios de persecución tanto de la Inquisición como de distintos tribunales civiles y eclesiásticos, instituciones que solían hacer uso de la correspondencia privada como prueba de los delitos que estaban siendo juzgados. Los documentos, que forman parte de procesos judiciales, vienen generalmente acompañados de 'interrogatorios sociológicos' llevados a cabo por inquisidores y jueces varios, lo que permite a los investigadores una contextualización adecuada de las relaciones interpersonales en las sociedades tradicionales. Estas fuentes escritas suelen presentar una retórica casi oral, tematizando asuntos cotidianos que hasta ahora no se habían estudiado más que a partir de casos aislados.

A partir de estas premisas, el objetivo fundamental de Post Scriptum es la creación de una base de datos electrónica formada por 7.000 de estas cartas (3.500 para cada lengua) que sirva como herramienta de trabajo para diferentes estudios humanísticos, principalmente para aquellos relacionados más

[1] El proyecto Post Scriptum está siendo financiado por el Consejo Europeo de Investigación (7FP/ERC Advanced Grant – GA 295562) y se desarrolla actualmente en el Centro de Lingüística da Universidade de Lisboa (CLUL). Además de los firmantes de este texto, en este momento forman parte del equipo investigador Guadalupe Adámez, Sandra Antunes, Catarina Carvalheiro, Tiago Castro, Elisa García, Raïssa Gillier, Mariana Gomes, Ana Leitão, Laura Martínez, Víctor Pampliega, Liliana Romão, Carmen Serrano y Leonor Tavares.

directamente con disciplinas como la historia moderna, la historia cultural, la crítica textual, la lingüística diacrónica y la lingüística de corpus. Por lo tanto, la edición digital en línea de *Post Scriptum* ofrece un conjunto heterogéneo de cartas escritas en diferentes contextos sociales y que obedecen a situaciones comunicativas variadas. En el ámbito de la lingüística diacrónica, la naturaleza dialógica de estos documentos privados permite compensar, en su justa medida, la carencia de fuentes orales. Por un lado, la espontaneidad en las interacciones que se generan a través de la correspondencia privada puede servir como una ventana al discurso cotidiano (Nevalainen y Tanskanen, 2007); por otro lado, un repertorio de cartas propias de contextos informales, producidas por manos poco instruidas y escritas casi como si fuesen habladas, constituye un recurso extraordinario para el estudio fonológico, morfológico y sintáctico de un determinado período histórico. Finalmente, los datos biográficos de individuos anónimos, así como sus formas de vida y sus interrelaciones sociales, suponen un valor de interés indiscutible, tanto desde la perspectiva histórica y cultural como desde la historiografía moderna.

En el presente trabajo, se explicará la metodología utilizada en Post Scriptum para la edición digital de los documentos y su disponibilidad en línea; además, se discutirán cuestiones relacionadas con la modernización de los textos y con su posterior etiquetación morfológica y sintáctica. Este artículo está estructurado de la siguiente manera: el siguiente apartado se centra en el proceso de localización y selección de las cartas; en el tercer apartado, se explica el procedimiento para la edición digital de los documentos, mediante la transcripción en formato XML; el cuarto apartado está dedicado al proceso de modernización de los textos y a su posterior etiquetación lingüística; finalmente, en el apartado quinto se alude a la relación entre diferentes tareas de edición; el trabajo se cierra con unas conclusiones generales sobre el proyecto.

2. BÚSQUEDA Y SELECCIÓN DE CARTAS

En un proyecto como el de Post Scriptum, donde lo que se pretende es construir y publicar una base de datos, la primera tarea que cobra importancia es la localización y selección de los datos. En la planificación de tareas que se ha trazado a lo largo de este proyecto, se han reservado los dos primeros años casi exclusivamente a la búsqueda de los textos junto a su transcripción en formato digital (ver el apartado siguiente), lo que deja ya entrever la dificultad de esta labor[2].

Por lo que se refiere a la localización de datos, la recuperación de correspondencia privada de los siglos XVI al XIX se lleva a cabo mediante la consulta *in situ* de archivos históricos, preferentemente, y como ya se apuntó más arriba, de aquellos que cuenten con un repertorio más o menos amplio de fondos judiciales y/o inquisitoriales. Así, para la búsqueda de cartas españolas se han consultado o se están consultando fondos documentales en el Archivo Histórico General (Madrid), el Archivo General de Simancas (Valladolid) y el Archivo del Reino de Galicia (La Coruña), además de otros archivos en diversos lugares de

[2] Por lo que se refiere a la búsqueda en archivos españoles, la media aproximada es de una carta encontrada por cada 10 procesos consultados; en archivos portugueses, la media es de 1 carta por cada 11 procesos.

la geografía española (Asturias, Cuenca, Guadalajara, Murcia, Orense, Pontevedra y Toledo)[3]. En cuanto al caso portugués, la mayor parte de documentación judicial e inquisitorial se encuentra centralizada en el Arquivo Nacional da Torre do Tombo (Lisboa); no obstante, también se están teniendo en consideración otros archivos no nacionales, como son los de Évora, Braga y Oporto, así como el archivo de Goa (India).

La consulta de diferentes archivos a lo largo y ancho de toda la Península Ibérica no está pensada sólo como una forma de aumentar las probabilidades de éxito en el hallazgo de cartas, sino que obedece también a la necesidad de controlar una cierta representatividad geográfica y lingüística de los datos. Junto a la tarea de localización, por tanto, se ha adoptado una estrategia de selección de las cartas encontradas. Además, este proceso selectivo se aplica tanto a la variabilidad espacial como a la variabilidad temporal de los textos: para asegurar el equilibrio cronológico del corpus, la previsión de 3.500 cartas por cada lengua se distribuye del siguiente modo: 500 para el siglo XVI –toda vez que la documentación judicial quinientista es difícil de encontrar en archivos–, 1.250 para el siglo XVII, 1.250 para el siglo XVIII, y 500 para el siglo XIX. Las cartas del Post Scriptum acaban en 1834, fecha representativa de los procesos de reforma de la administración que se siguieron al final del Antiguo Régimen.

Finalmente, también se aplica un control de selección en función del tipo de delito al que se vincula la carta. Experiencias previas en la consulta de fondos documentales nos han permitido comprobar que los procesos judiciales asociados a determinados delitos son más propensos a incluir misivas como prueba instrumental (por ejemplo, las acusaciones de bigamia o las de solicitación[4]). Para evitar un desequilibrio en este sentido, en aquellos archivos que albergan una cantidad considerable de documentación se ha procedido de manera selectiva, consultando siempre un porcentaje mínimo de procesos por cada delito existente.

En definitiva, los trabajos de búsqueda y selección de cartas están encaminados a la construcción de un corpus equilibrado y que sea representativo de los intereses de investigación en Post Scriptum: la escritura cotidiana epistolar en español y portugués durante la Edad Moderna.

3. TAREA PALEOGRÁFICA Y EDICIÓN DIGITAL

Una vez que se localiza la carta, el siguiente paso es transcribir el manuscrito, convirtiéndolo a un formato que sea legible por un ordenador. Y esto conlleva una serie de decisiones técnicas y metodológicas.

La codificación de los datos, tanto textuales como extratextuales, se realiza mediante el uso de lenguaje XML. 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.

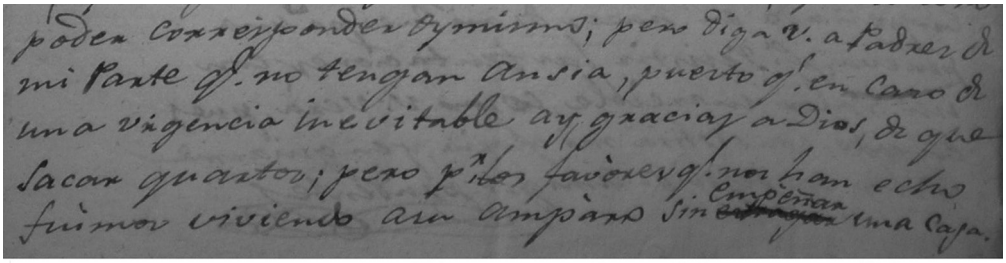
Buscando practicidad y garantías de integración con otros corpus electrónicos de naturaleza similar, se han adoptado los estándares de anotación propuestos

[3] En un futuro próximo, se contempla la consulta de fondos en archivos de Barcelona, Granada, Sevilla y Zaragoza.

[4] El delito de solicitación incluye cualquier tipo de provocación, insinuación o seducción del confesor hacia el penitente.

por el consorcio TEI (*Text Encoding Initiative*), por ser una convención internacional ya consolidada en la edición virtual de fuentes primarias. De todos los proyectos que siguen las directrices TEI, destaca el DALF (*Digital Archive of Letters in Flanders*), dedicado específicamente a la edición epistolográfica. Por consiguiente, la Definición del Tipo de Documento utilizada en Post Scriptum es un instrumento proporcionado por el proyecto DALF (DALF.dtd).

Para la transcripción del manuscrito se ha adoptado una actitud bastante conservadora, lo que da lugar a una edición semipaleográfica del texto original. Tan sólo 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 escritura, entre otros aspectos, se han respetado en la edición digital, mediante el uso de las etiquetas XML definidas por los proyectos TEI-P4 y DALF. Esto permite ofrecer una edición electrónica del texto manuscrito sin perder rigor filológico. Como muestra de lo dicho, sirva el ejemplo recogido a continuación, en el que las etiquetas "</lb>", "<del>", "<add>" y "<abbr>" permiten marcar cambios de línea, tachones, añadidos autoriales fuera de línea y abreviaturas, respectivamente:



```
poder corresponder oy mismo; pero diga <abbr>u<expan>sted</expan></abbr> a Padres de<lb/> mi Parte
<abbr>q<expan>ue</expan></abbr> no tengan ansia, puesto <abbr>q<expan>ue</expan></abbr> en caso
de<lb/> una vigencia inevitable ay, gracias a Dios, de que<lb/> sacar quartos; pero
<abbr>p<expan>o</expan>r</abbr> los favores <abbr>q<expan>ue</expan></abbr> nos han echo<lb/>
fuimos viviendo a su ampàro sin <del hand="JBC4">entregar</del> <add hand="JBC4"
place="supralinear">empeñar</add> una casa.<pb n="30v"/>
```

Fig. 1. Fragmento de una carta transcrita con etiquetas TEI-XML.

La información que ofrece Post Scriptum de cada manuscrito no se limita a la transcripción semipaleográfica del texto en sí, sino que incluye también diferentes datos de carácter extratextual, desde las características del soporte físico (disposición gráfica del texto, medidas del papel, estado de conservación) hasta la descripción histórica y contextual de la carta. Además, siempre que es posible se facilitan los datos biográficos de los autores y destinatarios de las cartas editadas (nombre, origen, ocupación, religión, estado civil…), los cuales se recopilan en una base de datos independiente.

4. INVESTIGACIÓN LINGÜÍSTICA

El proyecto Post Scriptum se propone enriquecer la investigación en lingüística diacrónica, beneficiándose para ello de los avances que proporcionan actual-

mente disciplinas como la lingüística de corpus o la lingüística computacional. Así, entre los objetivos de Post Scriptum, ocupa un lugar importante la anotación de tipo lingüístico. En principio, se ha contemplado la anotación morfosintáctica de todo el corpus (p. ej., etiquetado de clases de palabras) y la anotación sintáctica de, al menos, una parte importante de los datos.

Hoy en día, existen a disposición del investigador diferentes programas de etiquetación automática que agilizan este tipo de trabajo; no obstante, para obtener de ellos el máximo rendimiento y fiabilidad es deseable que los datos de entrada presenten una ortografía normalizada. Por eso, en Post Scriptum la anotación lingüística está supeditada a la normalización ortográfica previa de los textos que van a ser etiquetados. En el presente apartado se explicará la metodología adoptada para llevar a cabo ambas tareas, normalización y anotación, tanto para el portugués como para el español.

4.1. Normalización ortográfica

Los manuscritos originales de las cartas presentan una gran variabilidad ortográfica. Así, una misma palabra (p. ej., *vergüenza*) puede aparecer escrita de muy diversas formas (p. ej., *berguensa*, *verguensa*, *berguenza*, *vergüenza*, *berguença*, *verguença*…). Esta diversidad, que tiene un interés filológico y lingüístico en sí misma, es respetada escrupulosamente en la edición digital semipaleográfica de Post Scriptum. Sin embargo, a efectos de anotación automática resulta contraproducente, lo que hace necesario un proceso de modernización ortográfica de los textos.

Frente a la edición semipaleográfica, en la edición modernizada se ha normalizado la grafía y la acentuación de todas las formas, se han revisado los signos de puntuación y se ha divido el texto en párrafos. Además, se han expandido todas las abreviaturas[5], a excepción de: etc, PD, AD, XPTO y similares. 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[6]. Tampoco se ha intervenido sobre el nivel léxico: se han conservado los regionalismos y los arcaísmos léxicos, aunque se han señalado con una marca [*sic*] para posibilitar su recuperación.

El proceso de normalización de los textos se realiza mediante la herramienta eDictor (Faria, Kepler y Sousa, 2010), desarrollada por el grupo de investigación del proyecto Tycho Brahe. Se trata de un programa de interfaz amigable que permite seleccionar la palabra original y editarla de acuerdo con la grafía estándar actual. Además, una ventaja importante que presenta esta herramienta es su compatibilidad con las extensiones TXT, XML y HTML, lo que permite crear archivos de salida en cualquiera de estos formatos. Una vez normalizado el texto, eDictor también es capaz de organizar el contenido en tablas de datos, lo que facilita el visionado de las modificaciones que fueron realizadas en cada nivel de edición (p. ej., formas normalizadas, abreviaturas expandidas y variedades léxicas señaladas).

---

[5] En la edición semipaleográfica, la extensión de las abreviaturas aparece marcada con la etiqueta <expan> (ver Fig. 1).

[6] El texto correspondiente al sobrescrito de la carta, si es que lo hay, queda fuera de la edición normalizada.

El inconveniente de este proceso de edición ortográfica es que ha de realizarse de manera manual, es decir, seleccionando y modificando palabra por palabra todas aquellas formas que sean objeto de normalización. Por eso, en Post Scriptum se está trabajando en un procesamiento semiautomático que agilice esta tarea. 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.

4.2. Anotación lingüística

Los textos electrónicos que constituyen un corpus se pueden presentar de dos formas posibles: no anotados (en su estado original) y anotados (enriquecidos con varios tipos de información lingüística). Como ya se ha mencionado, uno de los objetivos de Post Scriptum es la creación de un corpus anotado de escritura cotidiana, que facilite al usuario la recuperación y el análisis de toda aquella información lingüística contenida en los textos. Actualmente, en Post Scriptum esta tarea incluye al menos la consideración de dos niveles de información: el etiquetado por clases de palabras y la anotación sintáctica. En ambos casos, la anotación se realiza sobre la edición modernizada de los textos y mediante el uso de diferentes programas informáticos, según la lengua de entrada. El proceso de anotación es siempre semiautomático, esto es, el programa devuelve automáticamente el texto etiquetado y un equipo de lingüistas revisa manualmente posibles errores de anotación.

Por lo que se refiere al primer nivel, o anotación morfosintáctica, la parte portuguesa del corpus está siendo anotada con el propio programa eDictor, que incluye un etiquetador morfológico. Para los textos en español, se está utilizando la herramienta FreeLing 3.0 (Padró y Stalinovsky, 2012). A continuación, se muestra un ejemplo de cada tipo de anotación morfológica:

> Isto/DEM sei-o/VB-P+CL eu/PRO hoje/ADV ,/, com/P toda/Q-F a/D-F certeza/N ,/, porque/C o/D que/WPRO há/HV-P são/SR-P inimigos/N-P e/CONJ aqueles/D-P home/N do/P+D Regimento/NPR ,/, bem/ADV entendido/VB-PP ,/, soldados/N-P e/CONJ alguns/Q-P oficiais/N-P inferiores/ADJ-G-P ,/, que/WPRO tu/PRO mesmo/FP os/CL conheces/VB-P ,/, estes/D-P são/SR-P aqueles/D-P que/WPRO te/CL fizeram/VB-D o/D mal/N ,/, mas/CONJ Deus/NPR sempre/ADV acode/VB-P a/D-F inocência/N ./.
>
> Fragmento de un texto en portugués anotado con eDictor
> CARDS0020, carta familiar de 1827

> En_en/SPS00 nuestro_nuestro/DP1MSP país_país/NCMS000 ha_haber/VAIP3S0 habido_haber/VMP00SM muchas_mucho/DI0FP0 tempestades_tempestad/NCFP000 ;_;/Fx ha_haber/VAIP30S quitado_quitar/VMP00SM mucho_mucho/DI0MS0 fruto_fruto/NCMS000 ;_;/Fx estamos_estar/VAIP1P0 trabajosos_trabajoso/AQ0MP0 ._./Fp Luego_luego/RG ,_,/Fc muchos_mucho/DI0MP0 soldados_soldado/NCMP000 ;_;/Fx un_uno/DI0MS0 regimiento_regimiento/NCMS000 se_se/P00CN000 va_ir/VMIP3S0 ,_,/Fc otro_otro/PI0MS000 entra_entrar/VMIP3S0 ;_;/Fx
>
> Fragmento de un texto en español anotado con FreeLing 3.0
> PS6001, carta familiar de 1736

Como se puede ver, eDictor asocia a cada palabra del corpus una etiqueta representativa de la clase de palabras a la que pertenece: p. ej., *hoje*, ADV(erbio). Por su parte, el analizador de FreeLing permite obtener, además de la categoría gramatical, el lema de cada palabra: p. ej., *tempestades*, lema *tempestad*,

N(ombre) C(omún) F(emenino) P(lural). 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, Santorini y Diertani, 2010), ligeramente revisado para adecuarse a las características de la gramática portuguesa[7]. 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[8].

Por lo que se refiere a la anotación sintáctica, la metodología adoptada en Post Scriptum varía de nuevo en función de la lengua que es objeto de análisis. Los textos escritos en portugués son etiquetados mediante el sistema de anotación Penn-Helsinki, siguiendo así un estándar de anotación compartido por otros proyectos de sintaxis diacrónica del portugués y desarrollados en cooperación con Post Scriptum: WochWel (URL: <http://alfclul.clul.ul.pt/wochwel/index.html>) y Tycho Brahe (Galves y Faria, 2010). Por otro parte, para el conjunto de cartas en español se ha optado de nuevo por la herramienta FreeLing 3.0, que permite tanto un análisis superficial del texto, es decir, la segmentación en constituyentes y la identificación de categorías sintácticas (sintagma nominal, sintagma verbal, sintagma preposicional…), como un análisis de dependencias, más detallado y con información sobre funciones sintácticas (sujeto, objeto directo, objeto preposicional…)[9].

La etiquetación por separado y mediante diferentes parsers del corpus español y el portugués responde a cuestiones prácticas. Es la única solución que permite hacer uso de la mejor tecnología inmediatamente disponible para el apoyo a los estudios de lingüística diacrónica en estas dos lenguas, al tiempo que permite la cooperación con proyectos paralelos ya en curso: en las Universidades de Lisboa y Campinas para el caso del portugués (proyectos WOChWEL y Tycho Brahe arriba mencionados), y en diversas universidades catalanas para el caso del español (Sánchez-Marco *et al*., 2010). Somos conscientes de que una decisión como esta dificulta la obtención y el análisis de datos comparativos entre ambas lenguas; no obstante, el sistema de etiquetas utilizado por el etiquetador POS de FreeLing resulta más completo y detallado que el utilizado por eDictor, por lo que siempre es posible convertir automáticamente las etiquetas del primer formato al segundo y anotar sintácticamente una muestra del corpus español siguiendo el sistema de anotación Penn-Helsinki.

5. RELACIÓN ENTRE NIVELES

En los apartados anteriores, se ha ofrecido una visión general sobre el proyecto Post Scriptum, incidiendo especialmente en el proceso de edición de las cartas y su posterior anotación lingüística. A la luz de lo expuesto, se constata

[7] En <http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/tags.html> se puede acceder a la lista completa de etiquetas utilizadas por el analizador morfológico de eDictor [comprobado el 13/09/2013].

[8] En <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html#verbos> se puede acceder a la lista completa de etiquetas EAGLES utilizadas por el analizador morfológico de FreeLing [comprobado el 13/09/2013].

[9] En <http://devel.cpl.upc.edu/freeling/svn/trunk/doc/grammars/esCHUNKtags> se puede acceder a la lista completa de etiquetas utilizadas por el analizador sintáctico de FreeLing [comprobado el 13/09/2013].

que la metodología adoptada en este proyecto de investigación implica diversas tareas o niveles de trabajo que siguen un orden secuencial: búsqueda de cartas en archivos históricos, transcripción del manuscrito en lenguaje XML, normalización ortográfica del texto y anotación lingüística. Con todo, es importante matizar que este conjunto de pasos conlleva, más que un proceso ascendente (de la transcripción semipaleográfica a la anotación sintáctica), una relación dinámica entre diferentes niveles de actuación. De hecho, se da el caso de que algunas normas o decisiones que son tomadas en un determinado nivel obligan a precisar o reformular otras decisiones tomadas en un nivel anterior. Así sucede, por ejemplo, con el tratamiento de las abreviaturas, con las contracciones no estandarizadas o con la puntuación de los textos.

La presencia de abreviaturas, que son de uso frecuente en la escritura cotidiana epistolar, representan un problema para los anotadores automáticos, no sólo por la incapacidad para reconocer la forma correspondiente sino porque una misma abreviatura puede referirse a más de una palabra (por ejemplo, *no.* puede referirse a *nuestro* o a *número*; *pa.* puede referirse a *para* o a *padre*). Para solventar este inconveniente, todas las fórmulas abreviadas aparecen ya expandidas en la transcripción XML mediante el uso de la etiqueta <expan>. De esta forma, la abreviatura se conserva como tal en la edición semipaleográfica, pues el contenido ausente en el original aparece entre paréntesis (*n[úmer]o*, *pa[dre]*), pero se ofrece ya expandida y sin ninguna marca adicional en la edición modernizada (*número*, *padre*), precisamente para evitar errores al aplicar el anotador automático.

Respecto a las contracciones no estandarizadas (español: *la venta destas caxas*, *en lamable compania*, *ya tescrito*; portugués: *deu pedir*, *que lle poço dever*, *pelamor de Deus*), que también aparecen con bastante frecuencia en las cartas, estas pasan a constituir palabras gráficas diferentes en la edición normalizada del texto (español: *la venta d´ estas cajas*, *en l´ amable compañía*, *ya t´ he escrito*; portugués: *d´ eu pedir*, *qu´ eu lhe posso dever*, *pel´ amor de Deus*). Sin embargo, casos como estos impiden mantener el mismo número de tokens en las dos ediciones digitales de la carta, semipaleográfica y modernizada, algo que es necesario respetar para poder sobreponer ambas ediciones en la web y que el usuario pueda cotejar, palabra por palabra, las dos versiones del texto. Por eso, se ha optado por dividir las contracciones no estandarizadas en la edición semipaleográfica (español: *la venta d estas caxas*, *en l amable compania*, *ya t e escrito*; portugués: *d eu pedir*, *qu eu lle poço dever*, *pel amor de Deus*), de tal forma que se conserve la alineación de tokens con respecto a la edición modernizada. Una etiqueta XML (<note n = ”contraction”>) permite dejar constancia en la edición semipaleográfica de la existencia de la contracción (*la venta <note n = ”contraction”>d</note> estas caxas*).

Finalmente, un tercer aspecto cuyo tratamiento está condicionado por tareas posteriores es el de la puntuación de los textos. Ya se ha comentado que la corrección de los signos de puntuación se realiza en el proceso de normalización ortográfica y que es esta versión ya normalizada la que se utiliza como base para el etiquetado automático. En este sentido, la puntuación que presente el texto de entrada es relevante, puesto que puede facilitar el análisis sintáctico de los datos, que es el más complejo y el que demanda una mayor revisión por parte del investigador. Por eso, se ha optado por aplicar una puntuación que,

sin desatender la norma ortográfica, permita aumentar el porcentaje de acierto de los parsers (p. ej., tendencia a frases cortas)[10].

6. OBSERVACIONES FINALES

El proyecto Post Scriptum comenzó en el año 2012 y recibe financiación del Consejo Europeo de Investigación hasta el año 2017. El trabajo realizado hasta la fecha se ha centrado fundamentalmente en la búsqueda de cartas y en la transcripción XML, si bien una parte del corpus portugués cuenta ya con anotación morfológica revisada. La publicación de las cartas se está llevando a cabo en formato electrónico e incluye la siguiente información[11]:

- Visualización del facsímile del manuscrito.
- Transcripción del texto en XML (edición semipaleográfica).
- Edición normalizada del texto (en la lengua original y en inglés).
- Alineación de edición conservadora y normalizada.
- Anotación morfológica y sintáctica.
- Fichas biográficas de los participantes (autores y destinatarios).
- Palabras clave (áreas de lingüística y de historia).
- Contextualización.

Actualmente, cualquier usuario tiene acceso libre a todos los ficheros XML, a la hoja de estilo y a la DTD; además, en el futuro se podrán consultar los corpus anotados y las extracciones obtenidas a partir de esos corpus. Con todo ello, creemos que el trabajo realizado en Post Scriptum se revelará útil para un público amplio, desde usuarios no especializados pero interesados en episodios históricos de las cartas, hasta historiadores interesados en el tipo de fuentes encontradas o lingüistas interesados en realizar búsquedas sistemáticas en un corpus de textos históricos, casi espontáneos, anotados y no fragmentados. En definitiva, lo que se ofrece es un conjunto de datos que facilita la investigación en línea para múltiples campos de estudio, con diferentes herramientas electrónicas.

BIBLIOGRAFÍA

Faria, Pablo; Kepler, Fabio y de Sousa, Maria Clara, "An Integrated Tool for Annotating Historical Corpora", *Proceedings of the Fourth Linguistic Annotation Workshop,* 2010, pp. 217-221.

Galves, Charlotte y Faria, Pablo, *Tycho Brahe Parsed Corpus of Historical Portuguese,* 2010, URL: <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html> [29/09/2013].

Hendrickx, Iris y Marquilhas, Rita, "From old texts to modern spellings: an experiment in automatic normalisation", *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(2) (2011), pp. 65-76.

Kroch, Anthony; Santorini, Beatrice y Diertani, Ariel, *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*, Department of Linguistics, University of Pennsylvania, CD-ROM, first edition. URL: <http://www.ling.upenn.edu/hist-corpora/> [29/09/2013].

[10] Mediante los facsímiles de las cartas, siempre es posible consultar los signos de puntuación originales.

[11] La dirección electrónica de Post Scriptum es <http://ps.clul.ul.pt/index.php>. En el momento de redactar estas líneas, ya se han publicado en línea 757 cartas.

Nevalainen, Terttu y Tanskanen, Sanna-Kaisa (eds.), *Letter Writing*, Amsterdam, John Benjamins, 2007.

Padró, Lluís y Stanilovsky, Evgeny, "FreeLing 3.0: Towards Wider Multilinguality", *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Estambul, mayo 2012, pp. 2473-2479.

Sánchez-Marco, Cristina; Boleda, Gemma; Fontana, Josep Maria y Domingo, Judith, "Annotation and Representation of a Diachronic Corpus of Spanish", *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Malta, mayo 2010, pp. 2713-2718.

RESUMEN

El proyecto de investigación *Post Scriptum: Archivo Digital de Escritura Cotidiana (P.S.)* tiene por objeto la recuperación y publicación de cartas privadas escritas en España y Portugal durante la Época Moderna. Estas epístolas, en su mayoría inéditas, fueron producidas por autores muy diversos y de diferente condición social. Así, podemos encontrar hombres o mujeres, adultos o niños, amos o criados, ladrones, soldados, artesanos, curas, políticos, y otro tipo de categorías sociales. Sus misivas sobrevivieron excepcionalmente, al cruzarse sus vidas con los medios de persecución tanto de la Inquisición como de distintos tribunales civiles y eclesiásticos, instituciones que solían hacer uso de la correspondencia privada como prueba de los delitos que estaban siendo juzgados. Estas fuentes escritas suelen presentar una retórica (casi) oral, tematizando asuntos cotidianos que hasta ahora no se habían estudiado más que a partir de casos aislados. En este trabajo, se explicará la metodología utilizada en la edición digital de los documentos para su disponibilidad en línea; además, se dará a conocer el proceso de modernización de los textos y su posterior etiquetación morfológica y sintáctica. El objetivo final es elaborar un corpus diacrónico anotado que sirva como recurso electrónico para el estudio lingüístico e histórico del español y el portugués.

Palabras clave: Cartas, portugués, español, anotación de corpus, lingüística diacrónica.

ABSTRACT

Post Scriptum: A Digital Archive of Ordinary Writings (P.S.) is a project that aims to collect and publish Portuguese and Spanish private letters written along the Modern Ages. These documents are unpublished epistolary writings, written by authors from different social backgrounds. They could be either masters or servants, adults or children, men or women, thieves, soldiers, artisans, priests, political activists, among other kinds of social agents. Their epistolarity survived by chance, when their paths met the persecution means used by the Inquisition and the civil courts, two institutions that used private correspondence as criminal evidence. These textual resources often present an (almost) oral rhetoric, treating everyday issues of past centuries in a register that hasn't been easy to study, apart from brief examples. In the proposed paper, discussion over the methodological options that lead to the digital edition available online will be raised. Further, the modernization of texts the POS and syntactic annotation will be explained. The aim is to develop a diachronic and annotated corpus that could be used as an electronical resource for linguistic and historical research of Spanish and Portuguese.

Keywords: Letters, Portuguese, Spanish, corpus annotation, diachronic linguistics.