Escritura epistolar, edición digital y anotación de corpus·

Resumen: En este trabajo presentamos el proyecto de investigación *Post Scriptum*, que tiene por objeto la búsqueda, edición y estudio histórico-lingüístico de cartas privadas escritas en España y Portugal durante la Edad Moderna. *Post Scriptum* reúne una amplia colección epistolar y la ofrece en dos formatos preparados para la búsqueda: el de la edición crítica digital y el del corpus lingüísticamente anotado. En esta presentación se exponen los aspectos fundamentales sobre el diseño y elaboración de *Post Scriptum*, desde la búsqueda de los manuscritos en archivos históricos hasta la edición digital y anotación semiautomática de los textos y su publicación en línea; también se recogen resultados definitivos y se aportan algunos ejemplos de explotación del corpus en diferentes niveles de análisis.

Palabras clave: lingüística de corpus, corpus histórico, edición digital, XML-TEI, cartas privadas.

Abstract: In this paper we present the project Post Scriptum, which aims to build up a campaign of collection, edition and historical-linguistic study of private letters written in Portugal and Spain along the Early Modern Ages. Not only does the Post Scriptum Project present a wide collection of private letters, but it also makes it available as a scholarly digital edition and as an annotated corpus. Key aspects about the development of Post Scriptum are explained in this paper, from the finding of manuscripts in historical archives to the digital edition and semi-automatic annotation of texts and their publication online. Final results and some examples that illustrate how the exploitation of this corpus works at different linguistic levels are also included.

Keywords: corpus linguistics, historical corpus, digital edition, XML-TEI, private letters.

· El proyecto de investigación *Post Scriptum* ha sido financiado por el Consejo Europeo de Investigación (7FP/ERC Advanced Grant - GA 295562).

1. INTRODUCCIÓN

En las últimas décadas, la progresiva expansión de las nuevas tecnologías y de los avances en el mundo de la computación han supuesto un profundo cambio en el mundo de la investigación científica, y las ciencias humanas no han permanecido ajenas a esta revolución. En el ámbito de la lingüística, resultan obvias las ventajas que ofrece el uso de los ordenadores para el estudio del lenguaje, pudiendo almacenar y procesar grandes cantidades de datos lingüísticos de manera rápida y eficaz mediante la creación de corpus en formato electrónico, que pueden ser anotados de manera automática. En el ámbito de la filología, se han desarrollado metodologías y estándares de codificación para la creación de ediciones críticas digitales de documentos, que combinan la visualización del facsímile con bases de datos que contienen la transcripción del texto, las variantes textuales, listas de manuscritos, glosarios y otro tipo de información filológica.

Los beneficios que ofrecen las nuevas tecnologías al área de la humanidades son innegables, aunque se echa en falta la creación de herramientas digitales que sean útiles a diferentes disciplinas científicas. Generalmente, las ediciones diseñadas por el filólogo o el historiador no son explotadas por el lingüista del mismo modo que lo sería un corpus anotado; y viceversa, los corpus lingüísticos, pensados fundamentalmente para la obtención de estadísticas y concordancias de palabras, constituyen un recurso de escasa utilidad para la investigación histórica o la crítica textual.

En este trabajo presentamos un proyecto de investigación ya concluido que incorpora las metodologías de las humanidades digitales y de la lingüística de corpus para ofrecer un tratamiento unitario de fuentes que pueden ser de interés en varias áreas de estudio. Se trata del proyecto *P. S. Post Scriptum. Archivo Digital de Escritura Cotidiana en Portugal y España en la Edad Moderna* (en adelante Post Scriptum), desarrollado en la Universidad de Lisboa desde el año 2012 hasta el año 2017 y centrado en la búsqueda sistemática, edición y estudio histórico-lingüístico de cartas privadas escritas entre el siglo XVI y el primer tercio del siglo XIX. Formado por un equipo de lingüistas e historiadores, este proyecto adoptó una perspectiva claramente multidisciplinar con el compromiso de cumplir un triple objetivo:

- Un objetivo histórico y cultural. Las cartas publicadas son en su mayoría inéditas y constituyen un patrimonio cultural en sí mismo. En Post Scriptum desarrollamos un trabajo de recopilación (y aun de recuperación) de un

amplio conjunto de fuentes epistolares a partir de la consulta de diferentes fondos archivísticos a lo largo y ancho de la geografía peninsular, para poder reunirlas en un único recurso electrónico: un archivo digital de escritura cotidiana.

- Un objetivo filológico. Preparamos una edición crítica digital de los manuscritos y de sus contextos históricos mediante una transcripción electrónica que conserva rigor filológico. Además, ofrecemos dos niveles adicionales de acceso al texto: una edición con grafía normalizada y la propia imagen del facsímile.
- Un objetivo lingüístico. Convertimos el contenido de las cartas en dos corpus históricos con anotación lingüística, uno para el español y otro para el portugués. En términos cuantitativos, estamos hablando de dos corpus de un millón de palabras cada uno, lo que equivale aproximadamente a 2500 cartas por lengua.

En este trabajo presentamos la metodología que se siguió en Post Scriptum para alcanzar cada uno de estos objetivos y aportamos algunos resultados ya prácticamente definitivos. También explicamos brevemente el sistema de búsqueda utilizado para recuperar datos a través de este recurso electrónico y ofrecemos algunos ejemplos de explotación del corpus. Las lenguas que constituyen el foco de interés de Post Scriptum son el español y el portugués, si bien los datos que ofrecemos en este trabajo se centran preferentemente en la parte española del proyecto.

2. BÚSQUEDA DE FUENTES

2.1 Cuestiones previas

Los datos a los que tenemos acceso para elaborar corpus contemporáneos no son comparables, ni en cantidad ni en calidad, con los que podemos obtener para construir corpus históricos. La compilación de estos últimos está condicionada por ciertas limitaciones bien conocidas, que suelen ser más acusadas a medida que retrocedemos en el tiempo: conservación fragmentaria de textos, dificultad de datación, falta de contextualización, distribución errática de géneros, etc. (Kohnen, 2007; Claridge, 2008; Kytö, 2011).

A estos problemas inherentes a la preservación de fuentes históricas hay que sumar además otra particularidad que, por obvia, no resulta menos importante: la carencia de

fuentes directas de lengua hablada. La lingüística histórica en general y especialmente la pragmática y la sociolingüística históricas necesitan acceder a muestras de uso real del lenguaje al tiempo que deben asumir la palabra escrita como fuente legítima de datos. Esta discrepancia entre lo deseable y lo disponible fue formulada por Labov en una cita recurrente que ya ha pasado a convertirse en definitoria de la lingüística histórica: "Historical linguistics can then be thought of as the art of making the best use of bad data" (Labov, 1994: 11).

Asumida la imposibilidad de contar con grabaciones de habla, la alternativa pasa por reunir muestras de lengua que, aun siendo producidas en un medio gráfico, se acerquen lo máximo posible a la dimensión de lo hablado. La carta de contenido privado se revela en este sentido como un caso paradigmático al cumplir, por lo general, una serie de parámetros que la sitúan en el polo de la inmediatez comunicativa: privacidad, familiaridad entre los interlocutores, fuerte implicación emocional, espontaneidad relativa (Koch y Oesterreicher, 2007: 29-30). En esta línea, se entiende que el uso de escritura epistolar como fuente de datos para la investigación en lingüística histórica haya sido puesto en valor en los últimos años (Jacobs y Jucker, 1995; Nevalainen y Tanskanen, 2007; Raumolin-Brunberg y Nevalainen, 2007; Elspass, 2012; Dossena y Del Lungo Camiccioti, 2012):

> Letters, and in particular private letters, are a rich source of data for historical pragmatics. They may contain more intimate and more colloquial language than other text types. It is an empirical question whether they are therefore closer to the spoken language than other more formal text types, but they contain many interactional features such as address terms, directives, politenes smarkers, apologies, and so on (Jacobs y Jucker, 1995: 8).

En esta empresa de búsqueda de la oralidad en lo escrito resultan particularmente interesantes los testimonios producidos por gentes poco instruidas, semialfabetizadas o en cualquier caso no profesionalizadas en la escritura, puesto que es de esperar que sus textos estén menos mediatizados por tradiciones discursivas, por expresiones formulaicas o por niveles de estandarización lingüística:

> Clearly, letters do not represent spoken utterances; but when persons who have had but limited experience in writing and exposure to the norms of written expression are forced to write nevertheless, their writing reflects many features of their speech fairly accurately: what they do is

> put their own "imagined" words onto paper, if only with difficulty. Thus, what we are most interested in are letters by semi-literate writers. (Schneider, 2013. 64)

El problema radica, una vez más, en la disponibilidad de material, pues acceder a este tipo de fuentes epistolares no es tarea fácil. La correspondencia de contenido privado tiene pocas posibilidades de sobrevivir al devenir histórico y, por tanto, parece lógico suponer que buena parte de la producción de cartas del pasado se haya perdido, destruida por el tiempo o por sus propios autores y destinatarios, que no debían encontrar motivos suficientes para su conservación. Con todo, y como nos recuerda Elspass, existen motivos para la esperanza: "Language historians will not come across such 'oral' texts frequently, but these texts do exist and many are still waiting to be unearthed from archives or private collections" (2012: 159).

2.1 Las fuentes en Post Scriptum

El punto de partida del proyecto Post Scriptum radica precisamente en haber constatado previamente una de esas oportunidades excepcionales sobre la conservación de fuentes históricas. Entre la documentación oficial generada por los tribunales del Antiguo Régimen se conservaron cartas particulares de gente muy diversa, cartas que llegaron hasta nuestros días archivadas dentro de procesos judiciales y que en su momento fueron utilizadas por los propios jueces como una prueba más de los delitos sobre los que deliberaban. Generalmente, se conservaron porque su contenido resultaba interesante a ojos de la ley para tomar decisiones sobre los crímenes de que eran acusados sus autores, sus destinatarios o terceras personas relacionadas con ellos o mencionadas por algún motivo en el texto. La casuística resulta casi tan variada como los contextos en que se produjeron las misivas que acabaron siendo archivadas. Veamos algunos ejemplos.

Muchas veces las cartas eran incautada por los propios medios de persecución de las instituciones, tanto de la Inquisicion como de tribunales civiles, eclesiásticos y militares. Es el caso del pleito contra Juan José Aranda, cura de Mazarulleque acusado del delito de proposiciones y hechos heréticos por la Inquisición de Cuenca en 1757. Tras varias sospechas, los inquisidores decretaron su puesto en prisión y el embargo de sus bienes, momento en el que se incautaron todas las cartas que tenía en su casa para ser incorporadas a la causa como prueba. El reo fue condenado a abjurar de levi, a dos años de reclusión en un convento y a una pena de destierro durante tres años.

Otras veces las cartas eran aportadas por alguna de las partes litigantes para demostrar algún hecho inculpatorio o exculpatorio. Es lo que sucede en el pleito civil de 1702 entre Antonia Pardo Osorio y José Bermúdez de Castro por el pago de una deuda. Junto a otro tipo de documentación, al proceso se adjuntaron seis cartas. Tres de ellas fueron aportadas por la demandante como prueba de que efectivamente existía una deuda, que se valoraba en 681 reales; las tres restantes fueron aportadas por José Bermúdez de Castro para demostrar que dicha deuda era de tan solo 260 reales.

También encontramos correspondencia producida a raíz del propio proceso judicial (entre abogados y clientes, entre acusados ya apresados y sus familiares o allegados, etc), que presentan igualmente una interacción entre bastidores y pueden ser encuadradas en términos situacionales. Así sucede en el caso de Tomas García, quien denunció en una carta a su abogado los malos tratos recibidos por parte del juez que lo había encarcelado por un delito de amancebamiento con María Cunga. Dicha carta fue utilizada por el destinatario para solicitar la puesta en libertad de su cliente, aunque la solicitud no tuvo éxito; Tomás García y María Cunga acabaron fugándose de la cárcel.

Valgan estos casos, tomados de cartas encausadas incluidas en Post Scriptum, como una pequeña muestra del tipo de situaciones que propiciaron la utilización y consecuente preservación de misivas en pleitos judiciales de la Edad Moderna. La verificación previa de que era factible reunir un número importante de estas cartas privadas, contextualizables y en su mayoría inéditas llevó al equipo de Post Scriptum a emprender un primer objetivo de localización de estas fuentes[1]. Esta tarea, que fue continuada pero a la que se concedió especial atención durante los dos primeros años del proyecto, se llevo a cabo mediante la consulta en archivos históricos, preferentemente de aquellos que contienen fondos judiciales o inquisitoriales de la época mencionada.

Se visitaron numerosos centros archivísticos, dentro y fuera del territorio peninsular, e incluyendo tanto archivos de ámbito estatal como de ámbito regional, provincial o municipal, así como archivos diocesanos y arzobispales. En la Tabla 1 recogemos la lista completa de instituciones en las que se realizaron consultas para la localización de documentación epistolar en español; dicha lista aparece ordenada en función del número aproximado de cartas localizadas por archivo:

[1] Esta verificación se inició con el proyecto CARDS (Cartas Desconhecidas), que dio como resultado un archivo digital de cerca de 2000 cartas portuguesas de la Edad Moderna. Post Scriptum es una continuación del proyecto CARDS que amplía el corpus portugués y crea un corpus de para el español similar naturaleza y tamaño.

Archivo	Cartas	Archivo	Cartas
A. Histórico Nacional	1716	A. Municipal de Burgos	9
A. de la Real Chancillería de Valladolid	601	A. Histórico Provincial de Asturias	9
A. Diocesano de Cuenca	323	A. Histórico Provincial de Pontevedra	9
A. Histórico Provincial de Sevilla	200	A. General de Simancas	8
A. Nacional da Torre do Tombo	170	Biblioteca Nacional de España	7
A. General del Arzobispado de Sevilla	148	A. General de la Nación de México	6
A. de la Real Chancillería de Granada	136	A. del Reino de Valencia	5
A. General de Indias	83	A. Histórico Provincial de Orense	2
A. Histórico Provincial de Zaragoza	64	A. Histórico Provincial de Cuenca	2
A. Histórico Provincial de Burgos	56	A. Histórico de la Ciudad de Barcelona	2
A. del Reino de Galicia	56	A. Histórico Provincial de Cantabria	1
A. Real y General de Navarra	47	A. General de la Región de Murcia	1
A. Histórico Provincial de Toledo	46	A. Générales du Royaume	1
A. Histórico Municipal de Toledo	33	A. Histórico Provincial de Huesca	0
A. Histórico de la Universidad de Valencia	29	A. Histórico Provincial de Teruel	0
A. Diocesano de Barcelona	27	A. Regional de Madrid	0
A. Histórico Provincial de Murcia	25	A. Catedralicio de Palencia	0
A. Histórico Provincial de Guadalajara	24	A. Municipal de Palencia	0
The National Archives (Kew)	19	A. Diocesano de Burgos	0
A. Histórico de la Corona de Aragón	13	A. General de la Villa de Madrid	0
A. Municipal de Murcia	11	**TOTAL**	**3889**

Tabla 1. Archivos consultados y cartas españolas localizadas.

Salvo contadas excepciones en que fue posible acceder en línea a fondos digitalizados, la búsqueda en archivo históricos se realizó in situ mediante una consulta continuada de procesos judiciales. También se llevó a cabo una lectura atenta de toda unidad procesal cuyo contenido incluyese material epistolar, con el objetivo de poder contextualizarlo y obtener información biográfica sobre autores y destinatarios, como veremos más adelante.

La búsqueda de correspondencia constituye una tarea compleja cuyo resultado está sujeto en buena medida al azar. Por norma general, los archivos históricos no disponen de catálogos o bases de datos que ofrezcan información detallada sobre el contenido de sus fondos judiciales. No es usual, por ejemplo, poder filtrar resultados en función del tipo de documentación que incluye cada proceso judicial, y mucho mas complicado resulta saber de antemano si un proceso contiene o no cartas de carácter privado. Si el fondo en cuestión es de tamaño reducido resulta factible un vaciado íntegro, pero si se trata de fondos con un gran volumen de documentación se hace obligado delimitar un subconjunto de búsqueda.

En algunos casos, sobre todo en una etapa inicial de la búsqueda de fuentes, se realizaron catas aleatorias sobre fondos documentales de gran tamaño. Es el caso, por ejemplo, del Tribunal de Distrito de la Inquisición de Toledo, un fondo inquisitoral

perteneciente al Archivo Histórico Nacional, o del conjunto de pleitos criminales incluidos en el fondo de la Real Audiencia y Chancillería de Valladolid. En otros casos, se combinaron consultas aleatorias con criterios selectivos de búsqueda, con la finalidad de incrementar las probabilidades de éxito en la localización de correspondencia, objetivo principal del proyecto en esta fase de búsqueda.

Las estadísticas obtenidas a raíz de las catas plenamente aleatorias permiten hacernos una idea de la complejidad que entrañó esta tarea de localización de fuentes: el porcentaje de procesos válidos (i.e. aquellos que revelaron una o más cartas) sobre el total de procesos consultados es del 6.24% para el caso del fondo inquisitorial de Toledo y del 5.36% para el caso del fondo criminal de Valladolid, como se desprende de los datos que recogemos en la Tabla 2:

Fondo (Archivo)	**Procesos totales**	**Procesos consultados**	**Procesos válidos**	**Cartas localizadas**
Inquisición de Toledo (AHN)	4581	2115	132	471
Pleitos criminales (ARCV)	12440	2740	147	481

Tabla 2. Relación de procesos y cartas en dos fondos con consulta aleatoria.

Sin lugar a dudas, las mayores dificultades para la obtención de cartas, tanto para el español como para el portugués, la encontramos en el siglo XVI, como reflejan los datos de la Tabla 3. Basándonos en nuestra experiencia consultando fondos 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.

Siglo	Cartas en español	Cartas en portugués
XVI	527	307
XVII	1127	1016
XVIII	1584	1247
XIX	651	791
Total	3889	3361

Tabla 3. Distribución por siglos de cartas encontradas.

Durante los cinco años del proyecto se consultaron fondos albergados en al menos 57 instituciones diferentes, contando archivos y bibliotecas: 37 en España, 13 en Portugal y 7 fuera de la península ibérica. 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.

3. EDICIÓN DIGITAL

3.1 Transcripción del texto

Una vez que las cartas han sido localizadas, el siguiente paso consiste en transcribirlas con el objeto de ofrecer una edición digital del manuscrito. Para tal fin, en Post Scriptum fue necesario adoptar una serie de decisiones técnicas y metodológicas, que explicamos y ejemplificamos a continuación.

En primer lugar, hubo que tomar partido acerca del nivel de transcripción sobre el que debíamos trabajar en el proceso de digitalización de los textos, entendiendo por ello la cantidad de información contenida en el documento original que consideramos necesario preservar o incluir en la transcripción resultante. Partimos de la aceptación de que cualquier transcripción implica siempre una selección de los hechos o caracterísicas observables en el documento transcrito y de que, por tanto, no existe una transcripción, por muy precisa que sea, capaz de representar la fuente original en su totalidad (Sperberg-McQueen, 2009: 31). Esto nos deja, no obstante, con un amplio rango de posibilidades en función del grado de detalle al que nos ajustemos en términos de conservación textual, rango cuyos extremos son descritos del modo siguiente por Driscoll (2006):

> At one end of the spectrum there are transcriptions which may be called strictly diplomatic, in which every feature which may reasonably be reproduced in print is retained. These features include not only spelling and punctuation, but also capitalization, word division and variant letter forms. The layout of the page is also retained, in terms of line-division, large initials, etc. Any abbreviations in the text will not be expanded, and, in the strictest diplomatic transcriptions, apparent slips of the pen will remain uncorrected. [...] At the opposite end there are fully modernized transcriptions, where the substantives are retained but everything else is brought up to date, in some cases to such an extent as to make it questionable whether they are to be regarded as transcriptions at all. In between these two extremes a number of levels may be distinguished — 'semi-diplomatic', 'semi-normalized', etc. — depending on how the accidents of the original are dealt with.

En el caso de Post Scriptum, entendemos que los documentos recopilados son interesantes como fuente de datos lingüísticos, pero también como fuente de datos históricos y aun como objetos que representan fragmentos de una práctica, producidos manualmente por cientos de personas que vivieron en algún punto de la Edad Moderna y que plasmaron en papel sus preocupaciones diarias. En definitiva, estamos ante un tipo de documentación que puede y debe ser abordado desde tres perspectivas diferentes: como artefacto, entendido como objeto físico; como texto, entendido como contenido lingüístico; y como contexto, entendido como el conjunto de circunstancias históricas asociadas al texto y al artefacto. (Honkapohja, Kaislaniemi y Marttila, 2009: 453) .

Vista desde este triple perspectiva, nuestra labor como editores debe ser una labor minuciosa que busque preservar cualquier detalle del manuscrito. Por todo ello, en Post Scriptum partimos de una transcripción bastante conservadora de estas fuentes epistolares. Aspectos como los cambios de línea, la ortografía, las abreviaturas, los tachones, los subrayados, las correcciones del autor, los accidentes del soporte o la orientación de la escritura, entre otros aspectos, se han respetado en la transcripción digital. Tan solo se ha normalizado la segmentación de palabras y el uso de las grafías 'i', 'j', 'u' y 'v', decisiones que responden en ambos casos a razones prácticas[2]. Esta transcripción semidiplomática se traduce en lo que podríamos llamar una edición crítica digital del documento, entendiendo por ello una edición en versión electrónica que mantiene rigor filológico, permitiendo reconstruir tanto el contenido textual como el propio proceso de escritura.

En consonancia con las prácticas actuales en el campo de las humanidades digitales, la transcripción de las cartas se ha realizado utilizando el lenguaje de marcado XML (*eXtensive Mark-up Language*) y adoptando los estándares de codificación propuestos por el consorcio TEI (*Text Encoding Initiative*) para la representación de textos en formato digital[3]. El modelo XML-TEI es una convención ya consolidada en la edición virtual de fuentes primarias, lo que garantiza la integración con otros recursos electrónicos de naturaleza similar.

---

[2] En no pocas ocasiones, la caligrafía que presentan estas cartas hace imposible tomar decisiones objetivas acerca de la delimitación entre palabras o la selección entre las grafías 'u' y 'v.

[3] TEI: <http://www.tei-c.org/index.xml>.

Conviene apuntar que al inicio del proyecto, en 2012, el consorcio TEI no disponía todavía de un conjunto de etiquetas XML pensado específicamente para la marcación de material epistolar. Por este motivo, en un primer momento se partió del modelo propuesto por el proyecto DALF (*Digital Archive of Letters by Flemish Authors and Composers from the 19th & 20th century*), que está basado a su vez en una versión ya desactualizada del citado consorcio (versión TEI-P4). La adopción de este primer modelo exigió, además, numerosas modificaciones como consecuencia de las demandas que se iban imponiendo en Post Scriptum, bien por los objetivos concretos del proyecto, bien por las propias características de corpus. El resultado es un modelo altamente personalizado que no responde a criterios estandarizados y que solo tiene validez como modelo de trabajo interno (cf. Vaamonde, 2016).

Actualmente, no obstante, Post Scriptum ofrece también un modelo estandarizado que toma como referencia dos fuentes: la propuesta de la Red CHARTA (*Corpus Hispánico y Americano en la Red: Textos Antiguos*)[4] y la propuesta del módulo TEI-CORRESP-SIG para material epistolar creada por Peter Stadler, Marcel illetschko y Sabine Seifert[5]. Ambas fuentes están basadas en la versión TEI-P5, la más actual en el momento de redactar estas líneas.

Para ejemplificar el proceso de edición digital llevado a cabo en Post Scriptum ofrecemos el fragmento de una carta escrita en 1789 (Imagen 1) y, a continuación, una versión simplificada de la transcripción correspondiente (Imagen 2)[6]. Obsérvese el uso de elementos XML-TEI para marcar cambios de línea (<lb/>), segmentos añadidos fuera de línea (<add>), abreviaturas (<abbr>) o tachones (<del>) en el documento original.

Imagen 1. Fragmento de una carta escrita en 1789

Imagen 2. Transcripción en XML-TEI

[4] CHARTA: <http://www.redcharta.es/>.
[5] Correspondence SIG: <http://www.tei-c.org/Activities/SIG/Correspondence/>.
[6] Por razones de claridad, se han eliminado de la transcripción todas las etiquetas XML-TEI que no son relevantes para el ejemplo en cuestión. La transcripción completa está disponible en la dirección electrónica del proyecto: <http://ps.clul.ul.pt/index.php>. Desde esta dirección también se puede consultar el documento íntegro, al que se puede acceder a través de la búsqueda por código de la carta, que en este caso es PS9026.

3.2 Descripción de metadatos

Por cada carta que pasa a forma parte del archivo digital de Post Scriptum se genera un documento XML. Este documento consta de dos partes principales: un elemento <text>, que incluye la transcripción del texto siguiendo las pautas apuntadas en el apartado anterior, y un elemento <teiHeader>, en que se organiza diversa información de carácter extratextual. Entre los metadatos que ofrecemos para cada manuscrito destacamos los siguientes:

- Datos relativos a la referencia archivística: lugar y nombre del archivo, signatura del documento, foliación.
- Datos relativos a las características físicas: descripción del soporte, disposición gráfica del texto, medidas del papel, estado de conservación.
- Datos relativos a la contextualización: fecha de la carta, lugar de origen y destino.
- Datos relativos al contenido: clasificación general (carta de amor, de amistad, familiar, particular, anónima), clasificación particular de tipo enunciativo (confesión, extorsión, súplica, petición, elogio, etc.), palabras clave de tipo histórico, breve resumen del contenido.

Además, el hecho de que las cartas no se presenten de manera aislada sino que estén integradas en una unidad documental mayor, como es el proceso judicial, nos permite obtener dos tipos de información adicional de especial interés.

Por un lado, generalmente es posible obtener un contexto más o menos detallado de la carta, incluyendo la razón que motivó su escritura así como su relación con el proceso en que fue archivada: por qué se inició el proceso, quiénes fueron los litigantes, quién aportó la carta al pleito y con qué objetivo, cuál fue la sentencia final, etc. En definitiva, el acceso al proceso judicial constituye la vía sobre la que establecer una reconstrucción de la situación comunicativa de la carta.

Por otro lado, muchos procesos incluyen interrogatorios y declaraciones hechas a diferentes personas relacionadas con el delito juzgado; y a través de esos interrogatorios podemos obtener perfiles biográficos sobre autores y destinatarios de las cartas. En no pocas ocasiones, es posible rastrear datos como el nombre completo, la ocupación, el lugar de nacimiento y/o residencia, la religión, la edad o el estado civil, entre otros

aspectos. Por ejemplo, sabemos que el autor de la carta mostrada en la Imagen 1 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 de la carta. Toda esta información biográfica es almacenada en una base de datos independiente, creada también en lenguaje XML, y puede ser utilizada a voluntad del usuario, ya sea con un interés histórico, ya sea para ser cruzada con los datos lingüísticos del corpus. Aspectos como la edad, el sexo, la categoría social o la procedencia geográfica constituyen variables sociales de particular interés para la sociolingüísica y la dialectología históricas.

4. CORPUS LINGÜÍSTICO

4.1 Cuestiones previas

La creación de un archivo digital de escritura cotidiana, formado particularmente por cartas de contenido privado, representa el objetivo filológico de Post Scriptum y para cumplirlo nos valemos de las prácticas de marcación desarrolladas en los últimos años en el campo de la humanidades digitales. El otro gran objetivo que nos proponemos es de tipo lingüístico y consiste en la elaboración de un recurso electrónico que facilite la explotacion y el tratamiento estadístico de los datos textuales. Valiéndonos de la metodología de la lingüística de corpus, este objetivo se concretiza en la creación de dos corpus históricos, uno por cada lengua, enriquecidos con diferentes niveles de anotación lingüística.

La consideración de este doble objetivo (filológico y lingüístico) nos ha llevado a encarar un problema que ya ha sido apuntado en otras ocasiones en el ámbito de la lingüística histórica: el hecho de que los métodos de anotación desarrollados por las humanidades digitales y por la lingüística de corpus apenas presentan puntos de encuentro (Honkapohja, Kaislaniemi y Marttila 2009). Esto se debe en parte a que ambos métodos de anotación persiguen intereses diferentes. Los primeros, de acuerdo con Elena Pierazzo, están encaminados a obtener un recurso electrónico que permita inspeccionar el documento en su totalidad, esto es, una edición digital que incluya 'the source, the output and the tools to produce and display it' (Pierazzo 2011, 474). La lingüística de corpus, por otro lado, está interesada fundamentalmente en el tratamiento estadístico de expresiones lingüísticas y busca automatizar el proceso de anotación tanto como sea posible. Las humanidades digitales buscan ofrecer una exploración 'imaginativa' a través de la dimensión cultural de los documentos publicados

electrónicamente (Driscoll 2006); la lingüística de corpus pretende investigar la gramática, el léxico y el discurso del lenguaje desde una óptica más empírica, menos impresionista.

En una primera etapa del proyecto, este conflicto de intereses llevó consigo la necesidad de recurrir a diferentes herramientas de trabajo que permitiesen dar respuesta tanto a la dimensión filológica como a la dimensión lingüística de Post Scriptum. Por lo que se refiere a la edición digital, el proceso de transcripción de las cartas se realizó con el programa Oxygen, un editor de lenguaje XML. En cuanto a las tareas de corpus, para anotación morfosintáctica del español se hizo uso del anotador automático de Freeling 3.0 (Padró y Stalinovsky, 2012), mientras que para el caso del portugués se utilizó la herramienta eDictor (Piaxão de Sousa et al., 2013); la normalización ortográfica previa a la anotación lingüística se realizó también con eDictor para las dos lenguas.

Repárese en que la utilización de esta batería de herramientas conlleva una desventaja importante, pues es necesario trabajar con varios archivos de salida, que presentan diferentes formatos y cuyo contenido no siempre es posible relacionar, lo que dificulta tanto la gestión como la explotación combinada de los datos. En otras palabras, la información relativa a la marcación textual en TEI y la información relativa a la anotación lingüística acaban por ser almacenadas en archivos diferentes, lo que deriva en la necesidad de mantener dos corpus independientes que apenas sí pueden ser mutuamente aprovechados.

Para solucionar este inconveniente, desde finales de 2014 todo el tratamiento lingüístico del corpus están centralizadas en TEITOK (Janssen 2016), una plataforma interactiva que permite reunir en un único soporte XML tanto el corpus anotado como la edición crítica digital. TEITOK fue pensado y diseñado originalmente para dar respuesta a las demandas de Post Scriptum, pero actualmente son varios los proyectos de investigación que han volcado sus datos a esta plataforma, que en palabras de su creador puede ser definida del modo siguiente:

> TEITOK is a web-based system for viewing, creating, and editing corpora with both rich textual mark-up and linguistic annotation. For visitors, the system provides a graphical user interface in which te annotated document can be visualized in a number of different ways, depending on what the user is interested in. And for administrators of the corpus, TEITOK uses the same interface to easily and efficiently edit the underlying XML document. (Janssen 2014)

Por lo que se refiere a Post Scriptum, las transcripciones en TEI son generadas fuera de TEITOK mediante la herramienta Oxygen. El archivo XML generado en este programa, que contiene la transcripción del texto y los metadatos, es importado a TEITOK, en donde se procede al tratamiento lingüístico del texto así como a cualquier otro tipo de corrección posterior. Este proceso incluye fundamentalmente las tareas de tokenización, normalización ortográfica, lematización y anotación morfosintáctica, que pasamos a describir brevemente en el siguiente apartado.

4.2 Tokenización

Una vez importado el archivo XML a la plataforma TEITOK, un primer paso consiste en la segmentación del texto en tokens, esto es, en ocurrencias de palabras y signos de puntuación. Durante el proceso de tokenización, que se realiza de manera automática, cada forma original del texto es marcada dentro de un elemento <tok>, al que se le asigna una identificación única también de manera automática. Esta estructura inicial permite delimitar 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 unidad <tok>. Por ejemplo, la forma *conpañia* incluida en el manuscrito de la Imagen 1 y que coincide con un cambio de línea, sería procesada en TEITOK del modo siguiente:

Imagen 3. Ejemplo de token en TEITOK

Los atributos @form, y @nform señalan la forma original y la forma normalizada de la palabra, respectivamente. Otros niveles de edición, como pueden ser la forma expandida de abreviaturas (@fform), variantes dialectales (@dform), información metalingüística (@ltags), lemas (@lemma) o etiquetas morfosintácticas (@mfs), también son añadidos de forma correlativa mediante atributos dentro de <tok>, lo que asegura siempre una vinculación entre los diferentes niveles para su posterior recuperación a través del motor de búsqueda de la interfaz.

4.3 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 interés 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 ya se explicó. Tal variedad gráfica, no obstante, constituye un problema 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, para que sirva como archivo de entrada del anotador automático y maximice su porcentaje de acierto; otra razón secundaria es la posibilidad de ofrecer al público lego una edición que facilite la lectura de las cartas en versión estandarizada.

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. Este proceso de normalización ortográfica se llevó a cabo de manera semiautomática. La plataforma TEITOK incluye una herramienta de normalización automática que realiza una primera corrección ortográfica del texto y el resultado correspondiente es revisado de forma manual antes de pasar el anotador lingüístico. Véase como ejemplo el siguiente fragmento, que representa la versión normalizada de la transcripción recogida en la Imagen 2:

> *Marcelina, quieran los cielos divinos que estas cortas letras te 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 carta a ella.*

Conviene precisar que las modificaciones realizadas sobre el texto 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.

4.4 Anotación morfosintáctica

La versión del texto con ortografía normalizada es utilizada como archivo de entrada para la anotación morfosintáctica, que se lleva a cabo nuevamente mediante un proceso de carácter semiautomático: un anotador automático asocia cada palabra contenida en el texto con un lema y una etiqueta morfosintáctica, y el resultado de esa anotación es revisado manualmente por un equipo de lingüistas.

El anotador automático integrado en TEITOK es el etiquetador Neotag (Janssen 2012). NeoTag no solo sirve para etiquetar los textos del corpus, sino que además utiliza el propio corpus ya anotado como corpus de entrenamiento, mejorando así progresivamente su porcentaje de acierto a medida que se aumenta el conjunto de datos.

El sistema de etiquetas que aplica Neotag para los textos de Post Scriptum está basado en la propuesta del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. El conjunto de etiquetas EAGLES se rige por un sistema de posiciones: cada etiqueta consta de una secuencia de letras y números, donde cada letra o número representa un rasgo morfosintáctico determinado dependiendo de su posición dentro de la secuencia. El significado de cada posición está asociado a la categoría principal, representada por la primera letra de la secuencia[7].

Por ejemplo, la forma *compañía* lleva asociada la etiqueta NCFS000, donde la N indica que se trata de un sustantivo y la C que se trata de un nombre común; la F de la tercera posición señala el género, en este caso femenino, y la S de la cuarta posición indica el número, en este caso singular. Los rasgos que no son aplicables o no son especificados para una forma particular en una lengua dada se señalan con un cero. El token correspondiente a la forma *compañía*, una vez anotado, se mostraría en TEITOK del modo siguiente:

Imagen 4. Ejemplo de token anotado en TEITOK

6. COMPOSICIÓN DEL CORPUS

Como ya hemos comentado, la búsqueda de documentación en español se llevó a cabo en 39 instituciones (archivos y bibliotecas) y arrojó un total de 3889 cartas. No obstante, el total de cartas que pasaron a formar parte del archivo digital y del corpus lingüístico fue menor, ya que sobre el total de cartas localizadas se efectuó una

[7] El conjunto de etiquetas utilizado para la anotación morfosintáctica del corpus Post Scriptum está disponible en: <http://ps.clul.ul.pt/index.php?action=tagset>.

selección en la que se tuvieron en cuenta al menos dos criterios. En primer lugar, en casos de conjuntos epistolares de gran tamaño escritos por una misma mano se seleccionaron, por regla general, un máximo de 25 cartas. En segundo lugar, se seleccionaron únicamente cartas originales, entendiendo por tales las escritas de puño y letra por su autor o los casos de escritura delegada, en las un autor mental se vale de un escriba para la redacción de la misiva. Quiere esto decir que se desecharon las cartas que constituían copias de un original, salvo en contadas ocasiones en las que su contenido resultó inusualmente interesante por razones históricas.

Teniendo esto en cuenta, el corpus que está actualmente accesible en línea presenta la composición que recogemos en la Tabla 4. Los datos ofrecidos en dicha tabla son prácticamente definitivos; faltaría añadir un conjunto de 200 cartas españolas y 80 cartas portuguesas, aproximadamente, que en el momento de redactar estas líneas se encuentran todavía en proceso de revisión:

	Español		Portugués	
Siglo	Cartas	Tokens	Cartas	Tokens
XVI	310	151439	254	136412
XVII	684	278616	578	257234
XVIII	933	383376	776	357059
XIX	512	166063	730	208018
Total	2439	979494	2338	958723

Tabla 4. Composición de los corpus español y portugués accesibles en línea.

Todos los manuscritos publicados están ya transcritos en XML-TEI, lo que nos permite hablar de una colección digital compuesta por casi 5000 cartas en línea. También está finalizado el trabajo de normalización ortográfica y de digitalización de las imágenes, posibilitando así tres vías de acceso para todo documento ya publicado: la edición semidiplomática, la edición normalizada y la edición facsimilar.

Por lo que se refiere al corpus lingüístico, ambos corpus rondan actualmente el millón de tokens. Para ambas lenguas, la anotación morfosintáctica se realizó sobre la mitad de los datos aproximadamente, si bien en el caso del español una parte del corpus anotado está pendiente de revisión manual. Post Scriptum ofrece también la anotación sintáctica de una pequeña parte del corpus, más reducida en el caso del español. Concretamente, los datos relativos a la anotación lingüística, en número de tokens, son los que recogemos en la Tabla 5:

	Español	Portugués
Total corpus	979494	958723
Total anotado (POS)	638399	605148
Total anotado (sintaxis)	63388	228105

Tabla 5. Composición del corpus con anotación lingúística.

Por último, la base de datos biográficos asociada al corpus contiene 5063 perfiles sumando autores y destinatarios, de los cuales se cuentan 2990 portugueses y 2073 españoles; se reparten en 4192 hombres, 848 mujeres y 23 participantes de sexo desconocido.

7. ALGUNOS EJEMPLOS DE EXPLOTACIÓN DEL CORPUS

Dedicamos este último apartado a explicar brevemente el sistema de búsqueda en línea del corpus y a ofrecer algunos ejemplos sobre el tipo de información podemos obtener. Los datos de Post Scriptum integrados en la plataforma TEITOK son accesibles a través de una interfaz de búsqueda que está dividida en tres bloques principales, a saber:

- Búsqueda del documento, que permite obtener información relacionada con el extratexto: lengua (español o portugués), año (incluyendo un intervalo de años), lugar de origen y destino, datos biográficos del autor (nombre, categoría social, sexo), entre otros aspectos. También en este bloque se puede filtrar la búsqueda mediante un conjunto amplio y cerrado de palabras clave asociadas a cuestiones temáticas e históricas (cartas sobre carlismo o sobre adulterio o sobre judeoconversos, por poner solo tres ejemplos).
- Búsqueda del discurso, que permite delimitar dos criterios adicionales relacionados con la dimensión discursiva del texto: la parte de la carta en que se aplicará la búsqueda, que puede ser el contenido narrativo del texto o un segmento formular concreto (por ejemplo, el cierre de la carta); y el tipo de carta según una clasificación general basada en cinco opciones: amor, amistad, anónima, familiar, particular.
- Búsqueda del texto, que posibilita las búsquedas propiamente lingüísticas a partir de los diferentes niveles de edición del corpus: forma original, forma normalizada, clase de palabra, lema, etc.

El usuario no interesado en el lenguaje puede destinar su búsqueda a recuperar manuscritos epistolares que cumpla determinadas condiciones recogidas en los dos primeros bloques. Por ejemplo, se pueden consultar cartas de amor escritas en el siglo XVII por autores pertenecientes al estamento eclesiástico, o cartas en español escritas desde Portugal y clasificadas con la palabra clave 'Conspiración', o cartas familiares escritas por mujeres desde América a España. Por su parte, el usuario interesado en cuestiones de lingüística histórica tiene a su disposición la búsqueda del texto que, a su vez, permite realizar dos tipos de consultas: una consulta sencilla limitada a una única palabra o una consulta más avanzada, basada en lenguaje CQP, que permite obtener resultados que combinan dos o más palabras.

Por defecto, el resultado de cualquier búsqueda realizada en el texto es siempre una lista de concordancias (*key word in context*), como muestra el ejemplo que recogemos en la Imagen 5 a partir de la búsqueda de la forma normalizada salud:

Imagen 5. Concordancias de la forma *salud*

No obstante, es posible ordenar el resultado por frecuencias de aparición de la forma original (i.e. la forma de la palabra tal como aparece escrita en el manuscrito), algo que puede resultar de interés para investigaciones de tipo ortográfico o incluso fonético. La Imagen 6 ilustra este tipo de resultado de nuevo con la forma *salud*:

Imagen 6. Variedades ortográficas asociadas a la forma *salud*.

Utilizando la consulta avanzada, el usuario puede ampliar la búsqueda a combinaciones de dos o más palabras. Por ejemplo, se pueden buscar todos los tokens cuyo lema sea *haber* seguidos de un token anotado morfosintácticamente como participio para recuperar de esta forma todas las ocurrencias de tiempos compuestos[8]. O, por ejemplo, se pueden recuperar todos los tokens anotadas como verbo seguidos de la forma normalizada *de*, que, a su vez, preceda a la forma normalizada *que*[9]. Esta última consulta permitiría iniciar un estudio sobre los casos de dequeísmo atestiguados en el corpus:

[8] En lenguaje CQP, esta consulta se podría hacer del modo siguiente: [lemma="haber"] [pos="VMP.+"].
[9] En lengauej CQP: [pos="V.+"] [nform="de"] [nform="que"].

Imagen 7. Ocurrencias de forma verbal + *de que*.

También es posible ordenar las búsquedas por frecuencia del lema, lo que puede resultar interesante para estudios de carácter léxico. Por ejemplo, se pueden obtener todos los tokens anotados como nombre común y ordenar el resultado por lema y etiqueta morfosintáctica. Obtendremos así una lista como la que se recoge en la Imagen 8, que devuelve el conjunto total de sustantivos del corpus ordenados por frecuencia de aparición:

Imagen 8. Sustantivos más frecuentes en Post Scriptum..

Sirvan estos ejemplos como muestra del tipo de datos que son fácil y rápidamente recuperables a través de la interfaz de búsqueda de Post Scriptum. Además, la posibilidad de cruzar los datos lingüísticos del corpus con variables extralingüiísticas abre todavía más las opciones de explotación del corpus. Por ejemplo, resulta factible analizar un fenómeno lingüístico determinado y asociarlo con la procedencia geográfica de los autores para observar la dimensión dialectal de dicho fenómeno. Vinculando cada lugar de procedencia con sus correpondientes coordenadas geográficas e importando los datos a un sistema de información geográfica se pueden trazar mapas dialectales de tipo histórico como el que ofrecemos en la Imagen 9, que presenta la distribución de autores laístas en Post Scriptum (i.e. autores para los que se ha atestiguado uno o más casos de uso del pronombre átono *la* o *las* en función de objeto indirecto)[10]:

Imagen 9. Mapa de autores laístas en el corpus de Post Scriptum

Finalmente, téngase en cuenta que cualquier usuario puede descargar los archivos XML en versión TEI-P5, que incluyen la transcripción y los metadatos, así como el corpus completo en formato TXT, tanto en versión original, como en versión normalizada o en versión anotada. En definitiva, junto a las posibilidades que ofrece nuestro sistema de búsqueda el usuario es libre de descargar los datos en formatos adecuados para trabajar sobre ellos con herramientas propias.

[10] Para un estudio de la variación pronominal en español con los datos de Post Scriptum, véase Vaamonde (2015).

8. CONCLUSIONES

Post Scriptum es un recurso de acceso libre en línea que aúna metodologías y técnicas propias de las humanidades digitales y de la lingüística de corpus. Está especialmente diseñado para ofrecer a un tiempo ediciones críticas digitales y anotaciones lingüísticas del corpus, facilitando así tanto estudios de carácter histórico (incluyendo historia de la lengua) como investigaciones centradas en el cambio lingüístico y la lingüística diacrónica. Actualmente, desde la dirección electrónica del proyecto es posible consultar, entre otros, los aspectos siguientes:

- Digitalización del facsímile.
- Edición crítica digital.
- Edición con grafía normalizada.
- Diversa información extratextual: fecha, lugar de origen y destino, resumen del contenido, contexto situacional, descripción del soporte, medidas, grafismo, estado de conservación, etc.
- Lematización y anotación morfosintáctica (parcial).
- Anotación sintáctica (parcial).
- Fichas biográficas de autores y destinatarios.
- Mapas con geolocalización de autores.

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, cualquier usuario puede descargar el corpus completo o cartas individuales en formato TXT y XML, tanto en su transcripción original como en versión normalizada o anotada. En resumen, Post Scriptum constituye un recurso electrónico que responda a los intereses de varias disciplinas científicas, entre las que cabe destacar la crítica textual, la lingüística histórica (incluyendo sociolingüística, pragmática y dialectología históricas), los estudios culturales o la historia de la cultura escrita.

REFERENCIAS BIBLIOGRÁFICAS

CLARIDGE, Claudia (2008): «Historical Corpora», en A. Lüdeling y M. Kytö (eds.), Corpus Linguistics: An International Handbook (Vol.1), Walter de Gruyter, Berlin/New York, 242-259.

DOSSENA, Marina y Gabriella DEL LUNGO CAMICIOTTI (2012): *Letter Writing in Late Modern Europe*, John Benjamins, Amsterdam/Philadelphia.

DRISCOLL, Mathew James (2006): «Levels of Transcription», en John Unsworth, Katherine O'Brien O'Keeffe y Lou Burnard (eds.), *Electronic Textual Editing*. [28, 06, 2017] <http://www.tei-c.org/About/Archive_new/ETE/Preview/driscoll.xml>.

ELSPASS, Stephan (2012): «The Use of Private Letters and Diaries in Sociolinguistic Investigation», en Juan Manuel Hernández-Campoy y Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, Wiley-Blackwell, Malden, 156-169.

HONKAPOHJA, Alpo, Samuli KAISLANIEMI y Ville MARTTILA (2009): «Digital Editions for Corpus Linguistics: Representing Manuscript Reality in Electronic Corpora», en Andreas H. Jucker, Daniel Schreier y Marianne Hundt (eds.), *Corpora: Pragmatics and Discourse*, Rodopi, Amsterdam/New York, 451–475.

JACOBS, Andreas y Andreas H. JUCKER (1995): «The historical perspective in pragmatics», en Andreas H. Jucker (ed.), *Historical pragmatics: pragmatics developments in the history of English*, John Benjamins, Amsterdam/Philadelphia, 3-33.

JANSSEN, Maarten (2012): «NeoTag: a POS tagger for grammatical neologism detection», *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, Estambul, Turquía, mayo de 2012, 2118-2124.

JANSSEN, Maarten (2014): *TEITOK. A Tokenized TEI environment*. [28, 06, 2017] <http://teitok.corpuswiki.org/site/index.php>.

JANSSEN, Maarten (2016): «TEITOK: Text-Faithful Annotated Corpora», *Proceedings of the Language Resources and Evaluation Conference (LREC 2016) ELRA*. Portoroz, Eslovenia, mayo de 2016, 4037-4043.

KOCH, Peter y Wulf OESTERREICHER (2007 [1990]): *Lengua hablada en la Romania: español, francés, italiano*, Gredos, Madrid. [Versión española de Araceli López Serena].

KOHNEN, Thomas (2007): «From Helsinki through the centuries: the design and development of English diachronic corpora», en Päivi Pahta, Irma Taavitsainen, Terttu Navelainen y Jukka Tyrkkö (eds.), *Studies in Variation, Contacts and Change in English. Volume 2: Towards Multimedia in Corpus Studies*. [28, 06, 2017] <http://www.helsinki.fi/varieng/series/volumes/02/kohnen/>.

KYTÖ, Merja (2011): «Corpora and historical linguistics», *Revista Brasileira de Linguística Aplicada*, *Belo Horizonte*, 11/2, 417-457.

LABOV, William (1994): *Principles of Linguistic Change*. *Internal Factors*, Blackwell, Oxford.

NEVALAINEN, Terttu y Sanna-Kaisa TANSKANEN (2007): *Letter Writing*, John Benjamins, Amsterdam/Philadephia.

PADRÓ, Lluís y Evgeny STANILOVSKY (2012): «FreeLing 3.0: Towards Wider Multilinguality», *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, Estambul, Turquía, mayo de 2012, 2473-2479.

PAIXÃO DE SOUSA, Maria Clara, Fabio KEPLER y Pablo Picasso Feliciano DE FARIA (2013): *E-DICTOR*, Version 1.0 beta 10, 2013. [28, 06, 2017] <http://edictor.net/download>.

PIERAZZO, Elena (2011): «A Rationale of Digital Documentary Editions», *Literary and Linguistic Computing*, 26 (4) (December 1), 463–477.

RAUMOLIN-BRUNBERG, Helena y Terttu NEVALAINEN (2007): «Historical sociolinguistics. The Corpus of Early English Correspondence», en Joan C. Beal, Karen P. Corrigan y Hermann L. Moisl (eds.), *Creating and Digitizing Language Corpora: Diachronic Databases. Vol. 2*, Palgrave Macmillan, Basingstoke/New York, 148-171.

SÁNCHEZ-MARCO, Cristina, Gemma BOLEDA, Josep Maria FONTANA y Judith DOMINGO (2010): «Annotation and Representation of a Diachronic Corpus of Spanish», *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Malta, mayo de 2010, 2713-2718.

SCHNEIDER, Edgar W. (2013): «Investigating Historical Variation and Change in Written Documents: New Perspectives», en J. K. Chambers y Natalie Schilling (eds.), *The Handbook of Language Variation and Change*, Wiley-Blackwell, Malden, 57-81.

SPERBERG-MCQUEEN, C. M. (2009). «How to teach your edition how to swim», *Literary and Linguistic Computing*, 24: 27–52.

VAAMONDE, Gael (2015): «Distribución de leísmo, laísmo y loísmo en un corpus diacrónico epistolar», *Res Diachronicae*, 13, 58-79.

VAAMONDE, Gael (2016), *Guía para la Edición Digital de Textos en P.S. Post Scriptim*, Centro de Linguística da Universidade de Lisboa, Lisboa.