Universitat Oberta de Catalunya

artnodes
REVISTA DE ARTE, CIENCIA Y TECNOLOGÍA

http://artnodes.uoc.edu

ARTÍCULO

NODO «HUMANIDADES DIGITALES: SOCIEDADES, POLÍTICAS, SABERES»

La multidisciplinariedad en la creación de corpus históricos: El caso de *Post Scriptum*

Gael Vaamonde
Centro de Lingüística de la Universidad de Lisboa

Fecha de presentación: abril de 2018
Fecha de aceptación: octubre de 2018
Fecha de publicación: noviembre de 2018

Cita recomendada

Vaamonde, Gael 2018. «La multidisciplinariedad en la creación de corpus históricos: El caso de *Post Scriptum*». En: «Humanidades digitales: sociedades, políticas, saberes». *Artnodes*. N.º 22: 118-127. UOC. [Fecha de consulta: dd/mm/aa] http://dx.doi.org/10.7238/a.v0i22.3238

Resumen

Los corpus históricos convencionales suelen estar centrados en el contenido lingüístico de los documentos recopilados, que son almacenados como texto plano para facilitar su procesamiento. Esta aproximación monodisciplinar tiende a obviar aspectos como la grafía original, las características físicas y presentacionales de los manuscritos o la información sociohistórica y contextual asociada a cada texto, aspectos que son relevantes para otros campos de investigación, cuando no para el propio estudio lingüístico. Frente a esta situación, el corpus histórico epistolar *Post Scriptum* es el resultado de un proyecto multidisciplinar formado por lingüistas e historiadores, y que combina métodos de las humanidades digitales y de la lingüística de corpus. Con esto, constituye un recurso electrónico que pretende ser de utilidad a varias disciplinas científicas, como la crítica textual, la lingüística histórica, la historia moderna, la paleografía o la cultura escrita.

Palabras clave

multidisciplinariedad, lingüística de corpus, edición digital, corpus histórico, carta privada

Multi-disciplinarity in building historical corpora: The case of Post Scriptum

Abstract

Traditional historical corpora tend to focus on the linguistic content of the compiled documents, which are stored in plain text to facilitate their processing. By adopting this mono-disciplinary approach, relevant aspects to other fields of research, and even to linguistic study itself, are disregarded: original spelling, manuscript physical description or socio-historical context. Taking this into account, the historical and epistolary corpus of Post Scriptum takes a multi-disciplinary approach involving linguists and historians, and combining methods from digital humanities and corpus linguistics. As a result of this approach, Post Scriptum becomes a useful resource for several research areas, such as textual criticism, historical linguistics, modern history, paleography or literary culture.

Keywords

Multidisciplinarity, corpus linguistics, digital edition, historical corpus, private letter

1. *Post Scriptum*: proyecto multidisciplinar

P. S. Post Scriptum. Archivo digital de escritura cotidiana en Portugal y España en la Edad Moderna (en adelante, *Post Scriptum*)[1] es un proyecto de investigación ya concluido, desarrollado en el Centro de Lingüística de la Universidad de Lisboa desde el año 2012 hasta el año 2017, y centrado en la búsqueda sistemática, edición y estudio histórico-lingüístico de unas 5000 cartas privadas escritas en España y Portugal entre el siglo XVI y el primer tercio del siglo XIX.

El punto de partida de *Post Scriptum* radica en haber constatado previamente una oportunidad excepcional sobre la conservación de fuentes históricas.[2] Entre la documentación oficial generada por los tribunales del Antiguo Régimen, se conservaron cartas particulares de gente muy distinta, 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 los que eran acusados sus autores, sus destinatarios o terceras personas relacionadas con ellos o mencionadas por algún motivo en el texto. Las cartas son inéditas en su amplia mayoría, y fueron producidas por personas de distinta condición social: hombres, mujeres, esclavos, condes, ladrones, artesanos, curas, soldados, comerciantes, presos, amantes y un largo etcétera que completa un universo heterogéneo de autores y destinatarios.

Estas misivas son interesantes como fuente de datos lingüísticos, pero también como fuente de datos históricos e incluso 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. Estamos ante un tipo de documentación que puede y debe ser abordado al menos 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). Para dar respuesta a esta triple perspectiva, *Post Scriptum* reunió un equipo multidisciplinar formado por historiadores y lingüistas y combinó métodos propios de las humanidades digitales y de la lingüística de corpus.

El resultado final es la creación de un archivo digital y un corpus anotado formado por unas 5000 cartas privadas (españolas y portuguesas), que se presenta como un recurso electrónico útil para la investigación en diferentes disciplinas científicas, entre las cuales cabe destacar la historia moderna, la lingüística histórica, la escritura cotidiana o la cultura escrita.

1. El proyecto de investigación *Post Scriptum* (http://ps.clul.ul.pt/) ha sido financiado por el Consejo Europeo de Investigación (7FP/ERC Advanced Grant - GA 295562).

2. Esta constatación se materializó en el proyecto CARDS (*Cartas Desconhecidas*), predecesor de *Post Scriptum* y cuyo objetivo se limitó al estudio de 2000 cartas portuguesas.

2.1. La dimensión filológica en *Post Scriptum*

Durante el proceso de compilación de un corpus histórico, resulta obligado tomar una decisión sobre el tipo de fuentes documentales en las que se ha de basar la selección y el almacenamiento digital de los textos. Nos referimos a la posibilidad de partir de la fuente original, que generalmente será documentación manuscrita, o de alguna edición moderna de esta fuente. Ante esta disyuntiva, el uso de ediciones modernas de textos históricos ha predominado sobre la transcripción de la fuente original como método de recopilación de datos. La razón de esta preferencia no es difícil de imaginar: la edición moderna de un texto histórico es por lo general de fácil acceso; evita la toma decisiones editoriales, pues el trabajo filológico ya está hecho de antemano; y sobre todo, agiliza enormemente el proceso de digitalización. Si la edición seleccionada de un texto no existe ya en formato electrónico, y por tanto está accesible en red, la tecnología actual en torno al reconocimiento óptico de caracteres permite escanear un documento impreso de manera rápida y eficaz. En otras palabras, partir de ediciones modernas permite obtener mayor accesibilidad, facilidad y rapidez de digitalización, librando al lingüista del tiempo y esfuerzo que supone transcribir y editar fuentes primarias. Las evidentes ventajas de esta situación de *philological outsourcing*, como la denomina Dollinger (2004), explica el uso predominante de ediciones modernas en la construcción de corpus históricos, principalmente en la compilación de los denominados macrocorpus y corpus de referencia:

> «The compiler is confronted with the task of computerization and would like to use, and in many cases due to time and labour constraints is bound to use, the work of philologists as a base. If an edition of a given text can be found, why should any time be dedicated to the transcription of texts from manuscript sources?» (Dollinger 2004, 5).

Desde un punto de vista práctico, el recurso de la edición moderna está fuera de toda duda. Desde el punto de vista lingüístico, sin embargo, no resulta la opción más adecuada y, de hecho, supone una serie de inconvenientes que se han venido señalando en los últimos años en el ámbito de la lingüística histórica (véase Lass 2004; Dollinger 2004; Grund 2006; Claridge 2008, 250-251; Honkapohja, Kaislaniemi y Marttila 2009, 456-460, *inter alia*).

Destacamos aquí al menos dos problemas relacionados con el uso de ediciones contemporáneas en lugar de manuscritos originales. En primer lugar, cabe citar la representación inadecuada del contenido textual. Generalmente, las ediciones usadas en corpus históricos no parten de una transcripción rigurosamente fiel al original. Las abreviaturas, la puntuación original del texto, los símbolos y demás elementos pictóricos, e incluso determinados caracteres, suelen ser obviados o normalizados con diferentes criterios, en función del editor. Además, el uso de diferentes ediciones implica distintos criterios de normalización ortográfica, cuya documentación no suele aparecer recogida en el corpus histórico. En segundo lugar, existe una representación superficial de la realidad del manuscrito. Generalmente, la edición moderna se centra únicamente en el texto mismo (por ejemplo, caracteres alfanuméricos), ignorando otro tipo de rasgos del manuscrito que también pueden resultar relevantes, como por ejemplo los aspectos visuales, estructurales y paratextuales del documento original (Meurman-Solin 2013).

Conscientes de esta problemática, y de las limitaciones que se derivan de una transcripción inadecuada o incompleta de la fuente original, en *Post Scriptum* nos propusimos como un objetivo fundamental la creación de ediciones diplomáticas digitales de las cartas manuscritas. Así, el archivo digital que ofrece *Post Scriptum* se ha construido a partir de transcripciones que mantienen rigor filológico y son fieles al manuscrito original, respetando tanto su contenido textual (abreviaturas, puntuación, decoración, disposición del texto, etc.) como su dimensión presentacional (adiciones, cancelaciones, daños en el soporte, cambios de mano, conjeturas, lagunas, etc.). Siguiendo las prácticas habituales en el ámbito de las humanidades digitales, las transcripciones se han llevado a cabo mediante lenguaje XML y aplicando los estándares de codificación de fuentes primarias propuestos por la *Text Encoding Initiative* (TEI) en su versión P5. Junto con la transcripción original, para cada carta del corpus se ofrece también la imagen del facsímil, para que el usuario tenga en todo momento la posibilidad de cotejar la transcripción propuesta con el manuscrito original.

Recogemos a continuación la imagen facsimilar de una carta de *Post Scriptum* y su correspondiente transcripción en XML-TEI. Por razones de claridad, la transcripción se ha simplificado y las líneas aparecen numeradas y sangradas para facilitar su localización.

Como se puede observar, la transcripción no contempla normalizaciones sobre la ortografía original del manuscrito (línea 5, *nobedad*; línea 9, *diezynuebe*), ni alteraciones en la distinción entre mayúsculas y minúsculas (línea 11, *Obliga*; línea 12, *Papel*). Además, se conserva la forma no desarrollada de las abreviaturas (línea 2, *Sn*; línea 17, *Q B S M*), así como la puntuación original y las figuras o ilustraciones que puedan aparecer a lo largo del texto (línea 1, cruz de encabezamiento). Finalmente, cualquier fragmento cancelado (línea 10, *embyio*) o adicionado (línea 14, *en*), así como otros aspectos paratextuales, también han sido debidamente marcados para su posterior recuperación en el corpus.

Se trata, en definitiva, de una edición semipaleográfica del manuscrito epistolar. El prefijo «semi» en *semipaleográfica* responde al hecho de no haber respetado al menos dos cuestiones en las transcripciones textuales de *Post Scriptum*: la frontera de palabra y la distinción entre las grafías *u*, *v*, *i*, *y*, en los dos casos por razones prácticas, pues la caligrafía que presentan muchas de las cartas dificultaba una delimitación objetiva de tales aspectos gráficos.

El trabajo filológico y paleográfico llevado a cabo en *Post Scriptum* no se limitó solo a la transcripción de los textos, sino que comprendió

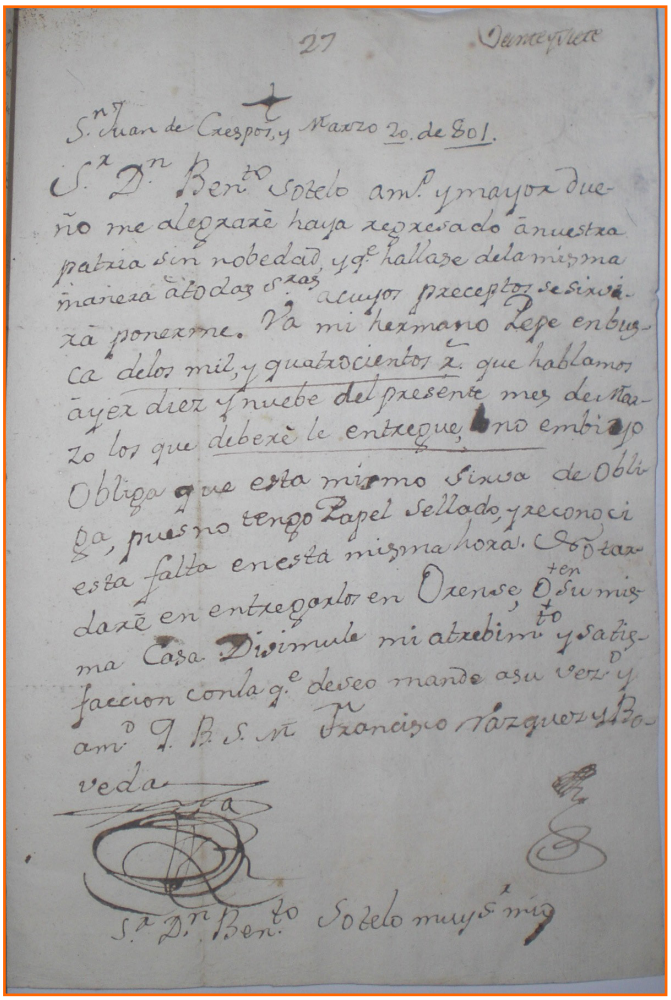

Imagen 1. Facsímil de carta escrita en 1801 por Francisco Vázquez y Bóveda[3]

```
<body><pb n="27r" facs="PS6196_2.JPG"/>
  <opener>
    <lb n="01"/> <seg type="sign">cruz</seg>
    <lb n="02"/> <placeName>Sn Juan de Crespos</placeName>,
                 <date>y Marzo <hi rend="underlined">20</hi> de <hi rend="underlined">801</hi>.</date>
  </opener>
  <p>
    <lb n="03"/> Sr Dn Bento Sotelo amo y mayor due-
    <lb n="04"/>ño me alegrare haya regresado a nuestra
    <lb n="05"/> patria sin nobedad y qe hallase de la misma
    <lb n="06"/> manera a todas sras a cuyos preceptos se servi-
    <lb n="07"/>ra ponerme. Va mi hermano Pepe en bus-
    <lb n="08"/>ca de los mil, y quatrocientos r que hablamos
    <lb n="09"/> ayer diezynuebe del presente mes de Mar-
    <lb n="10"/>zo los que debere le entregue, no embi<del hand="FVB1">y</del>o
    <lb n="11"/> Obliga que esto mismo sirva de Obli-
    <lb n="12"/>ga, pues no tengo Papel sellado, y reconoci
    <lb n="13"/> esta falta en esta misma hora. No tar-
    <lb n="14"/>dare en entregarlos en Orense, o <add hand="FVB1" place="supralinear">en</add> su mis-
    <lb n="15"/>ma casa Disimule mi atrebimto y satis-
    <lb n="16"/>faccion con la qe deseo mande a su vezo y
    <lb n="17"/> amo Q B S M <signed>Francisco Vazquez y Bo-
    <lb n="18"/>veda</signed>
  </p>
  <closer>
    <lb n="19"/> Sr Dn Bento Sotelo muy sr mio
  </closer>
</body>
```

Imagen 2. Transcripción en XML-TEI de la carta anterior

también la marcación y, por tanto, la posterior recuperación y análisis de otros aspectos relacionados con el documento en tanto que objeto físico. Concretamente, la disposición del texto, la descripción del soporte, el estado de conservación, las medidas del documento o la presencia de material adicional (sobrescrito) constituyen información que fue debidamente marcada y catalogada por cada misiva que pasó a formar parte de *Post Scriptum*.

2.2. La dimensión lingüística en *Post Scriptum*

Los corpus lingüísticos, tanto históricos como contemporáneos, se pueden dividir en dos grandes grupos: corpus no anotados y corpus anotados. Los primeros son aquellos que ofrecen únicamente el texto plano sin ningún tipo de información adicional; los segundos se caracterizan por presentar el texto enriquecido con anotación lingüística de algún tipo (morfológica, léxica, sintáctica, semántica, etc.). El uso de corpus no anotados no significa necesariamente que estos no puedan ser explotados lingüísticamente, aunque dependiendo del corpus, puede ser necesario un procesamiento previo de los datos, así como el manejo de expresiones regulares de diferente complejidad según la información que uno desee recuperar (Schulte 2009). En cualquier caso, las posibilidades de búsqueda que admite un corpus anotado serán siempre (*ceteris paribus*) mayores y más eficaces que las que puede ofrecer un corpus no anotado.[4]

Entre el tipo de anotación más habitual en la construcción de corpus lingüísticos, cabe destacar la anotación morfosintáctica (esto es, la asignación de la clase de palabra) y la lematización del corpus (esto es, la asociación de cada forma con su lema correspondiente), un proceso que se suele llevar a cabo mediante la utilización de anotadores automáticos, y cuyo resultado, dependiendo del tamaño del corpus, puede ser parcial o totalmente revisado por un anotador manual. La aplicación de este proceso a corpus históricos encuentra, sin embargo, un obstáculo adicional: el de la variación ortográfica que presentan los textos históricos. Una misma palabra puede aparecer escrita de múltiples formas, incluso dentro de un mismo texto. Por ejemplo, la forma *vergüenza* en el corpus español de *Post Scriptum* aparece atestiguada de trece formas diferentes (*verguença*, *verguenza*, *berguenza*, *berguença*, *berguenssa*, etc.). Esta variación repercute en el porcentaje de acierto del anotador automático y, por tanto, se traduce en un mayor tiempo y esfuerzo en la etapa de revisión manual.

La solución adoptada en *Post Scriptum* pasa por acometer una normalización ortográfica de los datos con posterioridad a su transcripción original, cuya información siempre se conserva (véase apartado 6), y como paso previo a la anotación lingüística. De este modo, el anotador automático se aplica únicamente sobre la forma normalizada (por ejemplo, *vergüenza*) y no directamente sobre las

3. Carta accesible en http://ps.clul.ul.pt/index.php?action=file&id=PS6196.xml.

4. En el ámbito hispánico, por ejemplo, véase la comparación entre el CORDE y el CdE (Davies 2009).

formas originales del manuscrito. Para la anotación morfosintáctica y lematización del corpus, se hizo uso del anotador 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. Además, NeoTag no impone un sistema de etiquetas propio, sino que permite utilizar un sistema personalizado. En el caso de *Post Scriptum*, se tomó como base el conjunto de etiquetas propuesto por el grupo EAGLES,[5] que se rige por un sistema de posiciones. Por ejemplo, la forma *vergüenza* es etiquetada como NCFS000, que quiere decir «Nombre Común Femenino Singular», y asociada al lema «vergüenza»; la forma *avergüenzo* es etiquetada como VMIP1S0, que quiere decir, «Verbo Principal Indicativo Presente Primera persona Singular», y asociada al lema «avergonzar». A modo de ilustración, recogemos a continuación un fragmento de una carta en versión verticalizada por columnas y con la correspondiente forma original (columna 1), forma normalizada (columna 2), etiqueta morfosintáctica (columna 3) y lema (columna 4). Todas las cartas incluidas en *Post Scriptum* se pueden descargar en un archivo TXT con este formato.

Mi	Mi	DP1CSS	mi
querida	querida	VMP00SF	querer
Amiga	amiga	NCFS000	amigo
	.	Fp	.
estoy	Estoy	VMIP1S0	estar
con	con	SPS00	con
mucho	mucho	DI0MS0	mucho
cuidado	cuidado	NCMS000	cuidado
por	por	SPS00	por
no	no	RN	no
aber	haber	VAN0000	haber
buelto	vuelto	VMP0000	volver
a	a	SPS00	a
saber	saber	VMN0000	saber
nada	nada	PI0NN00N	nada
desde	desde	SPS00	desde
nuestra	nuestra	DP1FSP	nuestro
bista	vista	NCFS000	vista
y	y	CC	y
asi	así	RG	así
espero	espero	VMIP1S0	esperar
me	me	PP1CS000	me
saque	saque	VMSP3S0	sacar
de	de	SPS00	de
este	este	DD0MS0	este
cuidado	cuidado	NCMS000	cuidado

Imagen 3. Anotación de un fragmento de carta en *Post Scriptum*[6]

Junto con la anotación morfosintáctica y la lematización, se han llevado a cabo otro tipo de anotaciones de carácter lingüístico. En primer lugar, la tarea de normalización ortográfica incluyó la aplicación de puntuación contemporánea sobre los textos, lo que permitió la división automática del corpus en oraciones ortográficas (esto es, unidades delimitadas por puntuación fuerte). Una vez hecho esto, se ha acometido también la anotación sintáctica de una pequeña parte del corpus, más reducida en el caso del español. Esta anotación toma como punto de partida el sistema originalmente creado para los *Penn Parsed Corpora of Historical English*, convenientemente adaptado a los datos del portugués y del español. Finalmente, el trabajo de enriquecimiento lingüístico del corpus contempló también una anotación de carácter discursivo. Por un lado, cada carta del corpus está asociada a un tipo temático basado en una clasificación epistolar tradicional: amor, amistad, familiar, particular o anónima. Por otro lado, se incluye en la transcripción del texto la marcación de posibles partes formulares que presente la carta, a saber: apertura, saludo, arenga, narración, peroración, cierre y posdata.

2.3. La dimensión histórica en *Post Scriptum*

La aproximación tradicional en la construcción de corpus históricos tiende a focalizar la atención en el contenido lingüístico en sí, limitando, cuando no obviando, no solo los aspectos paleográficos presentes en el documento original (disposición del texto, tipografía, etc.), sino también diferentes aspectos contextuales asociados a la producción del texto. Cierto es que sobre estos últimos, el compilador de corpus se ve muchas veces limitado porque no se conserva información suficiente que permita enriquecer un texto dado con factores de carácter histórico, social o cultura para multiplicar así las opciones de búsqueda del corpus. Por norma general, los corpus históricos se suelen ceñir únicamente a unos pocos aspectos extratextuales (marco cronológico, procedencia geográfica y género textual), que suelen además ser tratados en un nivel muy superficial (Meurman-Solin 2001).

Asumida esta situación, el caso de *Post Scriptum* no deja de representar una oportunidad excepcional. Las misivas que han podido ser localizadas y recopiladas no han llegado hasta nosotros de manera aislada, sino como parte de una pieza documental mayor, que es el proceso judicial. Desde el punto de vista histórico, este hecho resulta crucial, pues la lectura atenta de aquellos procesos que contienen cartas permitió encarar una caracterización sociohistórica de los textos, materializada en dos tipos de información: contexto situacional de la carta e información biográfica de los participantes.

5. http://www.ilc.cnr.it/EAGLES96/home.html.
6. Carta accesible en http://ps.clul.ul.pt/index.php?action=file&id=PSCR6925.xml.

Por un lado, fue posible contextualizar la carta, es decir, comprender la razón que motivó la escritura de esa carta y su relación con el proceso. En otras palabras, fue posible abordar una reconstrucción de la situación comunicativa de la carta. Por otro lado, a partir de los interrogatorios incluidos en muchos de los procesos, así como del resto de la documentación relacionada con la carta, generalmente fue posible recuperar información biográfica de los autores y destinatarios de las misivas: profesión, edad, religión, procedencia geográfica, etc.

Respecto al primer aspecto, y tomando el ejemplo de la carta mostrada en la figura 1, junto a la transcripción filológica del manuscrito y la anotación lingüística del texto, es posible acceder a la siguiente información histórica:

> «Pleito de 1804 de Benito Vázquez y Bóveda con Benito Sotelo Pérez, por la herencia y pago de una deuda. Benito Sotelo Pérez debía 1400 reales a Francisco Vázquez y Bóveda, difunto abad de San Juan de Crespos. Benito Vázquez y Bóveda, hermano de Francisco Vázquez, reclamaba el cobro de esos 1400 reales y aportó la carta aquí transcrita como prueba de la existencia de esa deuda. No obstante, durante el proceso se demostró que Benito Vázquez y Bóveda había rechazado la herencia de su hermano, por lo que no tenía derecho a recibir ningún dinero».

Siendo heterogénea la motivación que puede lleva a usar una carta como prueba instrumental de un proceso, los contextos situacionales reconstruidos en *Post Scriptum* resultan de lo más variado. Algunas veces, las cartas eran incautadas por los propios medios de persecución de las instituciones, tanto de la Inquisición como de tribunales civiles, eclesiásticos y militares. Otras veces, como en el ejemplo anterior, las cartas eran aportadas por alguna de las partes litigantes para demostrar algún hecho inculpatorio o exculpatorio. 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. Los delitos a partir de los que se generó, aportó o confiscó material epistolar presentan también una casuística variada, que va desde deudas económicas juzgadas por tribunales civiles hasta delitos inquisitoriales como los de bigamia, solicitación, alumbrados, etc. En todos los casos, suponen una excepción en el corpus las cartas para las que no fue posible establecer un contexto situacional mínimo que permita conocer las razones por las que se escribió una carta y por las que se usó como prueba en un proceso judicial.

Respecto al segundo aspecto, toda la información biográfica de autores y destinatarios, una vez recuperada a partir de la información del proceso, fue catalogada, marcada y organizada en lenguaje XML-TEI. Ofrecemos como ejemplo la entrada correspondiente a Juan Antonio Sierra, autor y destinatario de cartas entre 1745 y 1754:

Junto a la información sociohistórica recuperable a partir del proceso judicial, también se llevó a cabo una clasificación de cada

```
<person id="JAS5" role="author/addressee" sex="m" age="49" lang="ES">
  <persName>
    <name>Juan Antonio Serra</name>
  </persName>
  <affiliation>hijo de Jorge Serra y de Engracia Bernal; hermano de Rosa Serra</affiliation>
  <birth>1696</birth>
  <death/>
  <education n="learnedLanguage"/>
  <event>
    <desc>acusado por el obispo de Cuenca en 1747 de mantener un trato demasiado cercano y familiar con María García
      Almagro, su hija de confesión; acusado por la Inquisición en 1754 de cometer irregularidades en la dirección
      espiritual de María García Almagro, así como de los delitos de proposiciones erróneas e ilusa; preso en la cárcel
      del Santo Oficio de Cuenca en 1754</desc>
  </event>
  <faith>cristiano</faith>
  <floruit>1745-1754</floruit>
  <langKnowledge>
    <langKnown n="native"/>
  </langKnowledge>
  <nationality n="birthplace" key="39.424815, -2.293908">Cuenca, Vara de Rey</nationality>
  <residence n="primary" key="39.272544, -2.320589">Albacete, Minaya</residence>
  <residence n="secondary" key="40.071562, -2.137725">Cuenca</residence>
  <socecStatus key="ecclesiastical">cura, confesor y predicador</socecStatus>
  <state type="marriage">
    <p/>
  </state>
  <trait>
    <desc>dos varas de alto, cara enjuta y aguileña, piel de color cetrino, pelo y barba canos, manos y nariz
      largas</desc>
  </trait>
</person>
```

Imagen 4. Ejemplo de ficha biográfica en *Post Scriptum*

carta en términos históricos y culturales mediante la asignación de palabras clave tomadas de un conjunto amplio y cerrado de opciones. Esta clasificación no redunda sobre la normalización o la lematización del corpus, que permite obtener cualquier forma atestiguada en los textos, sino que funciona como un nivel complementario de recuperación al introducir y asignar a cada carta términos que no necesariamente aparecen en el cuerpo del texto y que permiten catalogar su contenido según parámetros históricos. Por ejemplo, la carta que venimos usando como ejemplo (figura 1) está asociada a tres términos de búsqueda: «Herencia», «Deudas» y «Petición». El número total de palabras clave con el que se trabajó asciende a unos trescientos términos aproximadamente.

3. Explotación de los datos

La idea central en *Post Scriptum* siempre fue la de ofrecer simultáneamente una edición digital de los manuscritos y un corpus histórico anotado. 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; Kytö 2011):

> «The searchability of a corpus is crucially dependent on how the corpus has been annotated. Again, there is a lack of consensus on this point, and compilers of historical corpora have been slow or even reluctant to apply standards such as the Text Encoding Initiative (TEI) Guidelines (P5). Many of the better known corpora are annotated for the main textual features but not all, and not as exhaustively as could have been the case».

A esto se suma, además, la necesidad de incorporar la información sociohistórica apuntada en el apartado anterior y, especialmente, la

caracterización biográfica de autores y destinatarios de las cartas, de modo que se puedan hacer búsquedas cruzadas entre variables sociales y datos lingüísticos. En definitiva, *Post Scriptum* demandó la existencia de una plataforma que permitiese sacar partido de todo el trabajo previo llevado a cabo en las diferentes dimensiones del proyecto: filológica, paleográfica, lingüística e histórica.

La solución técnica al problema anterior vino de la mano de TEI-TOK (Janssen 2016), una plataforma interactiva que permite reunir en un único soporte XML tanto el corpus anotado como la edición crítica digital, incluyendo asimismo la información metatextual. TEITOK fue pensado y diseñado originalmente para dar respuesta a las demandas de *Post Scriptum*, aunque actualmente son varios los proyectos de investigación que han volcado sus datos a esta plataforma:

> «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 su parte, la base de datos biográfica de *Post Scriptum* está vinculada con las transcripciones XML de las cartas mediante un identificador único para cada participante. Esta estrategia permite incorporar los datos sociales de los participantes al corpus lingüístico, abriendo así la posibilidad de hacer búsquedas útiles para estudios sobre dialectología o sociolingüística histórica. En realidad, una vez que los datos son importados a la plataforma TEITOK y que el sistema es configurado adecuadamente, pueden ser recuperados de múltiples formas y a partir de cualquier combinación que el usuario considere oportuna. Veamos algunos ejemplos ejecutados directamente desde la interfaz de consulta de TEITOK, tal y como está configurada para el corpus de *Post Scriptum*:

La interfaz consta de dos bloques principales. El bloque «Búsqueda del texto» permite hacer búsquedas sobre los datos textuales, tanto en su forma original como en su forma normalizada, así como búsquedas por etiqueta morfosintáctica y lema. El bloque «Búsqueda del documento» permite recuperar información relacionada con los datos extratextuales: lengua de la carta (español o portugués), año, lugar de origen, aspectos sobre el autor o el destinatario, etc. Finalmente, existe también un tercer bloque relacionado con aspectos discursivos, que permite filtrar la búsqueda en función del tipo de carta y/o de las partes formulares incluidas en ella.

Imagen 5. Interfaz de consulta de *Post Scriptum*

Un usuario interesado únicamente en cuestiones extratextuales puede recuperar, por ejemplo, todas las cartas de amor escritas por autores pertenecientes al estamento eclesiástico en el corpus español. Para ello, basta con que seleccione «español» en el campo «Lengua», «clero» en el campo «Estatus social del autor», y «amor» en el campo «Tipo de carta». Actualmente, obtendrá 16 resultados, es decir, 16 cartas que cumplen estos parámetros de búsqueda, con sus respectivos enlaces para poder consultar cada texto.

Un usuario interesado en cuestiones puramente lingüísticas puede obtener, por ejemplo, todas las variantes ortográficas atestiguadas en el corpus español para la forma «salud». Para ello, debe seleccionar el corpus español («español» en el campo «Lengua») y teclear «salud» en el campo «Normalización». En primer lugar, obtendrá la lista de concordancias de la forma normalizada «salud» en el corpus.

No obstante, en la parte inferior de la ventana de resultados, el usuario puede agrupar los datos por diferentes criterios. Ordenando por el campo «Transcripción provisional», que es el que se corresponde con la forma original de la palabra tal y como aparece en el manuscrito, obtendrá una tabla similar a la que recogemos a continuación. Esta tabla ofrece la lista, por orden de frecuencia, de formas originales asociadas a la forma normalizada «salud» en el corpus español:

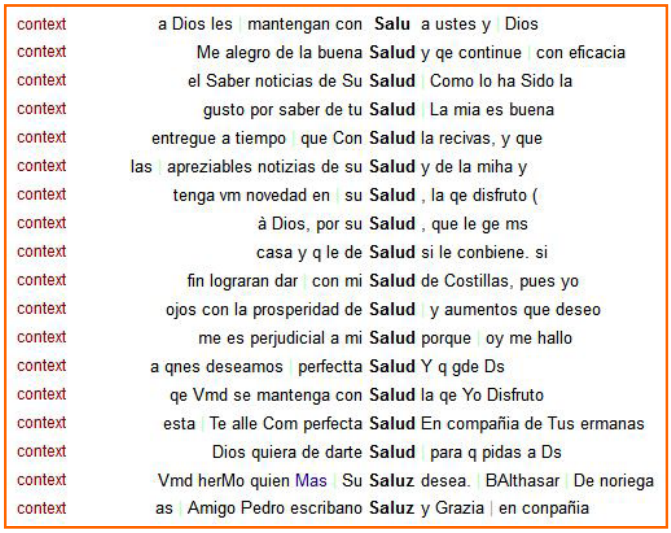

Imagen 6. Lista de concordancias de la forma «salud»

Graph: Table | Count: Count | Download: [seleccionar]

Transcripción provisional	Count	WPM
salud	873	445.585
saluz	32	16.333
salu	20	10.208
Salud	15	7.656
ssalud	5	2.552
Saluz	2	1.021
sallud	2	1.021
salus	2	1.021
Salu	1	0.51
salad	1	0.51
salut	1	0.51

Imagen 7. Formas originales asociadas a la forma normalizada «salud»

También es posible hacer búsquedas sobre dos o más palabras consecutivas. Por ejemplo, un usuario puede estar interesado en recuperar formas compuestas de «haber + participio» en el corpus español. Para ello, basta con teclear «haber» en el campo «Lema» y «VMP0000» en el campo «POS» o, alternativamente, ejecutar la siguiente orden en lenguaje CQP:

```
[lemma="haber"] [pos="VMP0000"]
```

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 extralingüísticas abre todavía más las opciones de explotación del corpus. Por ejemplo, la búsqueda anterior podría limitarse en función de si el autor es hombre o mujer; u obtener solo los casos atestiguados en un intervalo temporal concreto; o filtrar únicamente por una zona geográfica determinada o por el archivo histórico en el que se localizó la documentación epistolar.

4. 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, y que presta atención no solo a la dimensión lingüística de los textos, sino también a su tratamiento filológico y a su contextualización histórica. Actualmente, desde la dirección electrónica del proyecto es posible consultar, entre otros, los aspectos siguientes:

- Digitalización del facsímile.
- Edición semipaleográfica.
- Edición con grafía normalizada.
- Anotación morfosintáctica y lematización.
- Diferente información extratextual: fecha, lugar de origen y destino, resumen del contenido, contexto situacional, descripción del soporte, medidas, grafismo, estado de conservación, etc.
- Anotación sintáctica de una parte del corpus.
- Fichas biográficas de autores y destinatarios.
- Mapas con geolocalización de autores.

Toda esta información se integra en una interfaz de búsqueda que facilita no solo la consulta de cualquiera de los aspectos mencionados, sino también la búsqueda cruzada de los datos. *Post Scriptum* constituye, así, un recurso electrónico que responde a los intereses de varias disciplinas científicas, entre las cuales cabe destacar la crítica textual, la lingüística histórica (incluyendo sociolingüística, pragmática y dialectología históricas), la historia moderna o la cultura escrita. Creemos que este tipo de corpus especializados, que permiten un análisis meticuloso y multidisciplinar de los datos, son necesarios y aun complementarios de los grandes corpus existentes para avanzar en la investigación de la historia de la lengua. Finalmente, esperamos que en un futuro próximo surjan otros corpus similares que primen la calidad de los datos ofrecidos, tan demandada en lingüística histórica, sobre la cantidad de textos recopilados.

> «Bigger may not always be better, and size may not win all. Indeed, there are signs that the first decade of this new century will well turn out to be the decade of the small specialized corpus» (Swales 2006).

Referencias bibliográficas

Claridge, C. 2008. «Historical Corpora». En: A. Lüdeling; M. Kytö (eds.). *Corpus Linguistics: An International Handbook (Vol. 1)*, Berlín / Nueva York: Walter de Gruyter, 242-259.

Davies, M. 2009. «Creating useful historical corpora: a comparison of CORDE, the Corpus del Español and the Corpus do Português». En: E. Arias (ed.). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana/Vervuert, 137-166.

Dollinger, S. 2004. «"Philological computing" vs. "philological outsourcing" and the compilation of historical corpora: a Late Modern English test case». En: C. Dalton-Puffer y otros (eds.). *Vienna English Working Papers (VIEWS),* n.º 13: 3-23.

Grund, P. 2006. «Manuscripts as sources for linguistic research: A methodological case study based on the Mirror of Lights». *Journal of English Linguistics,* n.º 34: 105-125.

Honkapohja, A.; Samuli, K.; Ville, M. 2009. «Digital Editions for Corpus Linguistics: Representing Manuscript Reality in Electronic Corpora». En: A. H. Jucker; D. Schreier; M. Hundt (eds.). *Corpora:*

Pragmatics and Discourse, Ámsterdam / Nueva York: Rodopi, 451–475.

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

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

Kytö, M. 2011. «Corpora and historical linguistics». *Revista Brasileira de Linguística Aplicada*, *Belo Horizonte,* n.º 11, vol. 2:417-457.

Lass, R. 2004. «Ut custodiant litteras: Editions, Corpora and Witnesshood». En: M. Dossena; R. Lass (eds.). *Methods and Data in English Historical Dialectology (Linguistic Insights16)*. Berna: Peter Lang, 21–48.

Meurman-Solin, A. 2001. «Structured text corpora in the study of language variation and change». *Literary and Linguistic Computing,* n.º 16, vol. 1: 5-27.

Meurman-Solin, A. 2013. «Principles and Practices for the Digital Editing and Annotation of Diachronic Data». En: A. Meurman-Solin; J. Tyrkkö (eds.). *Studies in Variation, Contacts and Change in English* (vol. 14). Helsinki: Varieng.

Schulte, K. 2009. «Using non-annotated diachronic corpora: benefits, methods and limitations». En: E. Arias (ed.). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana/Vervuert, 167-180.

Swales, J. M. (2006). «Corpus Linguistics and English for Academic Purposes». En: E. Arnó y otros (eds.). *Information Technology in Languauges for Specific Purposes*, Springer, 19-33. https://doi.org/10.1007/978-0-387-28624-2_2

CV

Gael Vaamonde
Centro de Lingüística de la Universidad de Lisboa
gaelvmnd@gamil.com

Faculdade de Letras da Universidade de Lisboa
Alameda da Universidade (Lisboa) 1600-214. Portugal

Gael Vaamonde es licenciado en Filología Hispánica por la Universidad de Vigo (2002) y doctor en Lingüística por esta misma Universidad (2011). Entre 2002 y 2011 desarrolló su actividad investigadora en el Departamento de Traducción y Lingüística de la Universidad de Vigo, formando parte del proyecto *ADESSE: Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*. ADESSE es una base de datos con información sintáctica, semántica y léxica para todos los verbos y cláusulas de un corpus del español. Desde el 2002 hasta el 2011, compaginó su colaboración en ADESSE y en sus sucesivas ampliaciones con su formación investigadora en el área de la lingüística de corpus. La Universidad de Vigo le concedió una ayuda predoctoral para los años 2004 y 2005, periodo en el que obtuvo el diploma de estudios avanzados en Lingüística. En el 2006, defendió su tesis de licenciatura sobre la estructura argumental en español. Desde el 2006 hasta el 2010, fue beneficiario de una beca predoctoral FPI concedida por el Ministerio de Educación y Ciencia, que le permitió realizar estancias breves de investigación en Aarhus (2007), Berkeley (2009) y Leipzig (2010), sumando un total de 8 meses de estancia en universidades fuera de España. Defendió su tesis doctoral en julio del 2011, centrada en el dativo posesivo y otras construcciones afines en español.

En el año 2012, obtuvo una beca posdoctoral vinculada a un proyecto subvencionado por el ERC: *Post Scriptum: A Digital Archive of Ordinary Writings* (7FP/ERC Advanced Grant - GA 295562). *Post*

Scriptum reúne una amplia colección de cartas privadas escritas en español y portugués durante la Edad Moderna, y las ofrece en dos formatos preparados para la búsqueda: el de la edición digital y el del corpus lingüísticamente anotado. En este periodo, desarrollado íntegramente en la Universidad de Lisboa, continuó su investigación en el campo de la lingüística de corpus, al tiempo que adquirió competencias significativas en las áreas de la lingüística computacional y las humanidades digitales. Desde el 2017 hasta la actualidad, está contratado por el Centro de Lingüística de la Universidad de Lisboa como becario posdoctoral, ofreciendo asesoramiento y apoyo técnico a proyectos relacionados con la anotación de corpus y/o la edición digital de textos.

Actualmente, su perfil investigador combina experiencia en la compilación y anotación de corpus contemporáneos e históricos, conocimientos sobre edición digital de fuentes primarias y capacidades técnicas en programación y procesamiento del lenguaje natural. Sus líneas de investigación preferentes son la gramática del español, la lingüística de corpus, el procesamiento del lenguaje natural y la edición digital en TEI/XML.