Modelado de datos en TEI en proyectos de correspondencia digital en español

Marta López Izquierdo

Modelado de datos en TEI en proyectos de correspondencia digital en español
TEI data modelling in Spanish-language digital correspondence projects

Resumen:

Se analizan en este trabajo diferentes proyectos de edición digital de cartas en español que utilizan el marcado TEI, mostrando sus particularidades con respecto a otros proyectos epistolares del ámbito internacional. Se describen distintas prácticas tanto en el modelado de metadatos como de anotaciones en el texto y se valora su interés a la hora de extraer el máximo rendimiento de las posibilidades que ofrecen las humanidades digitales hoy para el conocimiento de los textos epistolares en sus múltiples dimensiones y en particular, para la explotación de su estructura reticular.

Abstract:

This paper analyses different projects for the digital edition of letters in Spanish that use the TEI markup, showing their particularities with respect to other epistolary projects in the international sphere. Different practices are described, both in the modelling of metadata and annotations in the text, and their interest in extracting the maximum benefit from the possibilities offered by the Digital Humanities today for the knowledge of epistolary texts in their multiple dimensions and, in particular, for the exploitation of their reticular structure is assessed.

Palabras clave:

cartas en español - redes epistolares - metadatos - anotación lingüística - TEITOK

letters in Spanish - epistolary networks - metadata - linguistic annotation - TEITOK

Marta López Izquierdo es catedrática de Lingüística española y Humanidades digitales en la Universidad París 8. En sus últimos trabajos se interesa por la variación (socio)lingüística y el contacto de lenguas en corpus escritos y orales de migrantes españoles en Francia. Es coordinadora del proyecto CAREXIL-FR. marta.li@univ-paris8.fr

Marta López Izquierdo is Professor of Spanish Linguistics and Digital Humanities at the University of Paris 8. In her recent work, she is interested in (socio)linguistic variation and language contact in written and oral corpora of Spanish migrants in France. She is coordinator of the CAREXIL-FR project. marta.li@univ-paris8.fr

La carta como género está presente desde la Antigüedad, aunque en la actualidad conoce importantes mutaciones, ligadas al desarrollo de la comunicación escrita por la vía digital. Por su amplia difusión, el texto epistolar conoce múltiples formatos y estilos y puede reunir tanto cartas de alta calidad literaria como escritas en una variedad alejada de los estándares lingüísticos de una época. Su interés trasciende el marco interpersonal que une al emisor y al destinatario, pues posee valor representativo de una época, un autor, un acontecimiento histórico o un territorio. De ahí, el interés que la edición y el estudio de cartas y epistolarios ha despertado entre los diversos estudiosos desde la Edad Media (cf. Petrucci 2008).

En la actualidad, la edición digital de cartas es un terreno muy fértil en el campo de las humanidades digitales que irradia en muchas direcciones y disciplinas, con intereses comunes, pero también con sus propias especificidades: la filología, la literatura, la lingüística, la historia, la sociología... Son numerosos los proyectos internacionales que han surgido en los últimos veinte años explotando las potencialidades de la edición digital

aplicada a las cartas y en particular, utilizando el lenguaje XML-TEI. De hecho, la edición digital epistolar ha servido como terreno de reflexión para el desarrollo de la TEI, a través de la creación de un grupo específico de trabajo dentro del consorcio para el marcado de los rasgos epistolares: TEI-Correspondence-SIG (cf. Stadler, Illetschko y Seifert, 2016). Entre los proyectos pioneros, cabe citar *Model Editions Partnership: Historical Editions in the Digital Age* (MEP)[1], desarrollado en la década de los 1990 y que se interesó por las cartas en tanto que documentos históricos, o el proyecto *Digital Archive of Letters in Flanders* (DALF)[2], que recoge desde 2002 cartas de escritores flamencos fechadas entre los siglos XIX y XX, y cuyo modelo editorial ha servido de base para el epistolario digital *Vincent Van Gogh: The Letters* (2009)[3], de amplia repercusión en la esfera de la edición digital del género que nos ocupa.

Stadler, Illetschko y Seifert (2016) documentan ampliamente los epistolarios digitales de los quince primeros años del siglo XXI, con representación de proyectos europeos y americanos, aunque no se encuentran mencionados entre ellos proyectos realizados por equipos del ámbito hispánico ni sobre cartas en español[4].

Sin embargo, en esta época ya estaba en marcha el proyecto Post-Scriptum (cf. Vaamonde et al. 2014), que contó con una importante financiación europea entre 2012-2017[5], y que representa a día de hoy el proyecto de edición digital epistolar más notable en lenguas ibéricas (portugués y español). Anteriormente, la Universidad de Lisboa había llevado a cabo otros dos proyectos de edición digital de cartas en portugués, precedentes directos del proyecto Post-Scriptum: *CARDS: Unknown Letters* (Marquilhas 2006-2009) y *FLY: Forgotten Letters* (Marquilhas 2010-2013).

Es necesario también recordar otros proyectos de edición digital de cartas que se han llevado a cabo en el ámbito hispánico con otros sistemas de codificación: es el caso de corpus lingüísticos digitales que incluyen entre sus documentos cartas y epistolarios varios, como los corpus de la Real Academia Española CORDE y CREA o el más reciente CORPES XXI, aunque solo este último adopta el lenguaje TEI para la codificación[6].

Han de citarse igualmente los distintos corpus lingüísticos históricos de equipos vinculados a la Red CHARTA, que incluyen entre la documentación editada en línea numerosas cartas de diverso tipo y que llevan varios años realizando un importante trabajo de aclimatación del lenguaje TEI a la documentación histórica en español (Isasi et al. 2014, 2020).

---

[1] Puede consultarse una descripción en https://tei-c.org/activities/projects/model-editions-partnership-historical-editions-in-the-digital-age/. El proyecto, albergado en la Universidad de South-Carolina, no está ya accesible en línea.

[2] https://ctb.kantl.be/project/dalf/index.htm

[3] http://vangoghletters.org/

[4] Tampoco se recogen proyectos españoles o sobre el español en https://www.zotero.org/groups/2160280/digital_correspondence_projects/library, a excepción de Post-Scriptum.

[5] https://cordis.europa.eu/project/id/295562/reporting

[6] El corpus CORDE presenta entre los textos históricos la categoría "Cartas y relaciones", que reúne textos epistolares de naturaleza muy diversa (cartas públicas, privadas, administrativas...) y textos no epistolares; el corpus CREA ofrece la categoría Cartas: personales, empresa, propaganda; el CORPES XXI detalla más específicamente "carta oficial Circular" y "carta particular", además de "carta al Director". Sin embargo, las búsquedas efectuadas seleccionando la tipología "Carta particular" remiten a un solo documento, publicado en prensa.

En época aún más reciente arrancan otros proyectos de edición digital basados en TEI que incluyen cartas (como CODIAJE[7], en construcción desde 2012) o, más específicamente, que se conciben como epistolarios digitales (CAREXIL-FR[8], en construcción desde 2019).

Nos centraremos en este artículo en el uso que se está extendiendo en el ámbito hispánico para el marcado de cartas con TEI. Comenzaremos nuestro análisis haciendo algunas observaciones sobre las particularidades de la edición digital de correspondencias en proyectos internacionales antes de pasar a presentar con mayor detalle algunas de las experiencias que incluyen cartas en español. Nos interesa en particular la manera en que los editores explotan las potencialidades del marcado TEI para anotar las características específicas de las cartas. También nos detendremos en el entorno TEITOK, que es la herramienta hoy en día más utilizada para los proyectos de corpus epistolares, entre otras tipologías de corpus, en España. Tras exponer cómo se han modelado los metadatos de las cartas en TEI en diferentes proyectos internacionales e hispánicos (sección 1), estudiaremos el marcado de los textos epistolares, en función de los intereses de investigación (textuales, pragmáticos, lingüísticos) (sección 2), y terminaremos ofreciendo algunas conclusiones de nuestro trabajo en la sección 3.

1. Metadatos para la correspondencia en TEI

Los primeros proyectos de edición digital de cartas que hemos mencionado previamente destacaron la importancia de desarrollar anotaciones específicas para diferenciar el género epistolar de otros tipos textuales. Como explican Stadler, Illetschko y Seifert (2016), los proyectos ya mencionados MEP y DALF introducen marcas para identificar nociones como el autor y el destinatario de las cartas, las fórmulas de despedida o las postdatas, que poco a poco irán configurando el módulo <corresp Desc> 'descripción de la correspondencia', adoptado por el consorcio TEI en 2015 (Figura 1). Como se puede observar, se utilizan dos bloques, el primero, <corresp Action> informa sobre la situación de emisión y recepción de la carta (quién, dónde, cuándo), mientras que el segundo, <corresp Context> enlaza la carta dentro de una correspondencia, identificando otras cartas enviadas o recibidas por las mismas personas.

Figura 1. <corresp Desc>.

```
<corresp Desc>
        <corresp Action type="sent">
          <persName>Nombre del autor de la carta</persName>
          <placeName >Lugar donde se escribe la carta</placeName />
           <date when="1999-01-01"/>
                </corresp Action>
                <corresp Action type="received">
                  <persName>Nombre del destinatario de la carta</persName>
        <placeName />Lugar donde se manda la carta</placeName>
                  <date when="1999-01-02"/>
                </corresp Action>
                <corresp Context>
                  <ref type="prev" target="ID de la carta">
                    Carta anterior de <persName>
                      autor</persName> a <persName> destinatario </persName>: <date
```

[7] Corpus diacrónico anotado del judeo-español, http://corptedig-glif.upf.edu/teitok/codiaje/
[8] CARtas de REpublicanos Españoles REfugiados y EXILiados en FRancia, http://carexil.huma-num.fr/

```
when="1999-01-01"/>
                </ref>
            </correspContext>
        </correspDesc>
```

El mismo grupo que ha creado el módulo epistolar en TEI, ha diseñado el formato CMIF (Correspondance Metadata Interchange Format) y una herramienta <corresp Search>[9] para integrar en una sola base interrogable distintos corpus epistolares. Se utilizan los metadatos más importantes para la comparabilidad de cartas de orígenes diversos: nombre del emisor con su identificador, nombre del destinatario con su identificador, fecha de escritura y de recepción, lugar de escritura y de recepción, con identificador, número de carta en la edición de referencia (si la hay) y URL de la carta editada. La base acoge cualquier colección de cartas que disponga de los índices con el formato indicado. Actualmente, enlaza 177.000 cartas con más de 12.000 personas identificadas. Se trata en su mayoría de cartas en alemán, fechadas entre el s. XVI y el XX. Hay también colecciones de cartas noruegas, inglesas y francesas. La presencia de equipos españoles o que trabajan sobre cartas en español es anecdótica[10].

El interés de este tipo de bases que mutualizan los datos de diferentes corpus epistolares se ha puesto de manifiesto con la elaboración de proyectos como *Mapping the Republic of Letters*, repositorio que permite la visualización de las redes de correspondencia entre escritores, intelectuales y artistas, desde Erasmo hasta Franklin. Con base en la Universidad de Stanford, se compone de un conjunto de asociados entre los que figuran la Universidad de Oxford, el CNRS francés y otras instituciones de los Países Bajos e Italia. Se reúnen en él diversos estudios particulares (estudios de caso) sobre diferentes personalidades del mundo de las letras. Para el ámbito hispánico, incluye el proyecto *An Intellectual Map of Science in the Spanish Empire, 1600-1810*, dirigido por Marcelo Aranda, a partir de la base *Spanish Scientist Database*, compilada a partir del *Diccionario Histórico de la ciencia moderna en España* (López Piñero et al. 1983), que abarca el período de finales de la Edad Media hasta 1970. El proyecto se interesa en los científicos activos en la época moderna y constituye una base biográfica de cerca de 360 personalidades. Sin embargo, no está centrado en la compilación de cartas, como los otros proyectos de la base, sino en interacciones de personas y organizaciones a través del imperio español.

Ha de mencionarse igualmente el catálogo digital *Spanish Republic of Letters* (Lazure), que recoge las redes de correspondencia que se establecieron entre los humanistas del Renacimiento español y los humanistas de otros países europeos, con el objetivo de visibilizar la importancia de los eruditos españoles en el desarrollo de este movimiento cultural. Se recogen así las noticias bibliográficas de más de 5500 de cartas escritas entre el s. XVI y el XVII, con los metadatos siguientes: tipo de documento, autor, destinatario, fecha, lugar de origen y de destino, repositorio donde se encuentra el original, edición moderna y lengua. La plataforma permite realizar búsquedas por nombres propios de persona en el texto de las cartas. No disponemos de información sobre el protocolo de marcado que se ha utilizado en este proyecto.

---

[9] Se trata de una API (Application Programming Interface) encapsulable en una página web que permite el intercambio de datos entre proyectos epistolares a través de la transmisión de ficheros XML en formato CMIF. https://correspsearch.net/en/home.html

[10] Solo hemos encontrado una carta de de Alexander von Humboldt a Carlos IV, editada por los investigadores Puig-Samper y Garrido (2016).

Los editores de cartas en español que adoptan el marcado XML-TEI han adoptado en muy pocas ocasiones el módulo <correspDesc> y ello probablemente debido a que las cartas no se editan en proyectos especializados en correspondencias, sino que incluyen otros tipos de documentos. Ello hace que se haya propuesto una cabecera genérica para todas las categorías textuales, como ocurre por ejemplo en CORPES XXI[11] o en los proyectos de la Red CHARTA. En el corpus en línea de este último consorcio, solo figuran 35 cartas privadas sobre un total de 2.076 documentos (7 pertenecientes al corpus CODEA y las demás al corpus SAI).

El manual del proyecto CHARTA (Isasi Martínez et al. 2020) incluye la etiqueta <classDecl> para la tipología de documentos en dos niveles: en un primer nivel, una taxonomía, que engloba todas las clases de documentos reconocidas por el proyecto CHARTA en 2013[12], con 10 tipos en total, y un segundo nivel, en cada documento, como se muestra en el ejemplo, aplicado a la taxonomía de cartas privadas, con la división en cartas privadas, cartas semipúblicas y cartas secretas. (Figuras 2 y 3)

Figura 2. Primer nivel de taxonomía de la Red CHARTA.

```
<classDecl>
      <taxonomy xml:id="Tip-CH">
        <bibl>Tipología CHARTA, propuesta de octubre de 2013</bibl>
        […]
<category xml:id="car-pri">
          <category xml:id="car-pri-pri">
            <catDesc>Cartas privadas</catDesc>
          </category>
          <category xml:id="car-pri-spu">
            <catDesc>Cartas semipúblicas</catDesc>
          </category>
          <category xml:id="car-pri-sec">
            <catDesc>Cartas secretas</catDesc>
          </category>
</category>
        […]
</taxonomy>
    </classDecl>
```

Figura 3. Segundo nivel de taxonomía de la Red CHARTA.

```
<profileDesc>
      <textClass>
        <catRef target="#car-pri-spu"/>
      </textClass>
    </profileDesc>
```

[11] Como se deduce del documento que describe el sistema de anotación con TEI en este corpus. Los ficheros XML de cada documento no están disponibles para la consulta.

[12] La tipología de documentos del proyecto CHARTA está accesible en la página del proyecto: https://75debdcd3c.cbaul-cdnwnd.com/a0324a9da168ff19cd0d0e5a8d92deb8/200000025-edbc7eeb54/Propuesta%20tipolog%C3%ADa%20documental%20CHARTA%20(1).pdf En ella, se proponen las siguientes categorías de cartas: “Cartas de compraventas y contratos” y “Cartas privadas”, definidas como “Documentos de carácter privado o semiprivado”, que engloban otras subcategorías: “Cartas privadas”, “Cartas semipúblicas”, “Cartas secretas”, etc.

Las taxonomías se adaptan según el tipo de proyecto: así, el corpus CODHECUN[13], vinculado a esta misma red, recoge también cartas, entre otros tipos documentales, con una taxonomía algo diferente, pues distingue entre cartas oficiales, particulares, personales y comunicaciones[14].

Como vemos, aunque la categoría "carta" está contemplada en la taxonomía de la Red CHARTA, no se prevé un tratamiento específico para los metadatos epistolares o al menos no a través del módulo <correspDesc>. Por ejemplo, el corpus COSUIZA (Castillo Lluch y Díez del Corral), igualmente asociado a esta red, utiliza la etiqueta <particDesc>, 'descripción de participantes', que dentro de TEI se utiliza para identificar los participantes de cualquier interacción lingüística[15]. En el caso de las cartas, este elemento permite identificar a emisores y destinatarios, como se puede ver en el ejemplo de la Figura 4[16].

Figura 4. <particDesc> en COSUIZA.

```
<particDesc>
      <person role="sent" sex="M">
        <persName>Fidel Andrés de Villalobos</persName>
      </person>
      <person role="received" sex="M">
        <persName>Armand François de Saint-Saphorin</persName>
      </person>
    </particDesc>
```

Fuera ya de esta red, el proyecto CoDiAJe – Ladino Corpus (Quintana 2012-), que reúne documentos en judeo-español de los siglos XIV al XXI, clasificados por género textual, incluye seis cartas en un conjunto documental de 89 documentos. El proyecto da acceso a una serie muy completa de metadatos, entre los que figura el título, el autor, el género, el lugar de origen, la fecha, el lugar de nacimiento del autor, el lugar de residencia, la lengua, la descripción del contenido, palabras clave, la dirección de la escritura (puesto que se trata de un texto judeo-español escrito en alfabeto hebreo) y el alfabeto. Otros metadatos incluyen el repositorio, el género textual, la traducción. El proyecto utiliza un marcado genérico para todas las tipologías textuales, con la posibilidad de modelar más estrechamente las cartas en versiones posteriores si alguno de los investigadores desea trabajar sobre ese género específico[17].

El módulo <correspDesc> sí aparece en algunos proyectos especializados en edición de cartas, como es el caso de Post-Scriptum (CLUL 2014-2017). Se trata de una base de datos que reúne cartas privadas escritas en España y Portugal en la Edad moderna con objetivos de investigación pluridisciplinar. Desde el punto de vista de la edición, este proyecto siguió en sus primeros pasos el modelo del proyecto DALF, en una versión personalizada para responder a las necesidades de los investigadores de la Universidad

[13] (Corpus Documental y Hemerográfico de la Cuba del Novecientos): se trata de un corpus dedicado al estudio de relaciones entre Cuba y Andalucia en el siglo XIX y, para las cartas, del proyecto "Lengua, identidad y memoria a través de las cartas y la prensa de Andalucía y Cuba (s. XIX)", dirigido por Eva Bravo-García (Universidad de Sevilla). http://cuba19.us.es/cuba19/index.php?action=home

[14] Agradecemos a Leyre Martínez que nos haya comunicado información sobre el modelado de los datos del proyecto CODHECUN antes de su difusión pública.

[15] https://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/es/html/ref-particDesc.html

[16] Carta COSUIZA-0015.xml, acceso al fichero TEI-XML por gentileza de Mónica Castillo Lluch.

[17] Aldina Quintana, comunicación personal. Para una descripción más completa de este corpus y del desafío de trabajar con una lengua sin estándar y por consiguiente sujeta a gran variación interna, cf. Quintana 2020.

de Lisboa. En una segunda fase, se adoptó el módulo TEI-CORRESP-SIG, acorde por consiguiente a los estándares del consorcio (Vaamonde 2018: 149) (Figura 5).

Figura 5. <correspDesc> en Post-Scriptum[18].

```
<correspAction type="sent">
      <persName key="CDD.xml#FN8" xml:id="FN8" cert="low">
         <name>Fernando Nunes</name>
      </persName>
      <placeName confidence="0" type="origin">
         España, Madrid
         <location>
            <geo>40.4167754 -3.7037902</geo>
         </location>
      </placeName>
      <date confidence="0" when="1630-11-04" from="1630" to="1630" when-
custom="1630.11.04"/>
   </correspAction>
   <correspAction type="received">
      <persName key="CDD.xml#FPS2" cert="low">
         <name>Francisco Pacheco da Silva</name>
      </persName>
      <placeName confidence="0" type="destination">
         España, Madrid
         <location>
            <geo/>
         </location>
      </placeName>
   </correspAction>
</correspDesc>
```

Como se puede observar, el proyecto Post-Scriptum reproduce el esquema principal del módulo <correspDesc> y personaliza algunos de los elementos por medio del atributo @key, que permite enlazar algunas informaciones con un fichero central que reúne todos los datos de cierto tipo, aquí el fichero CDD.xml donde el identificador #FN8 remite al autor. El mismo procedimiento se utiliza para el destinatario.

De manera similar, el proyecto CAREXIL-FR (Author, 2019-), dedicado a la edición de cartas de petición formal del exilio republicano español, utiliza el módulo específico para correspondencias, con diversos complementos de información a través de los atributos @key (Figura 6). Se recurre en este caso a un fichero central (ner.xml) para la identificación de varios contenidos: la persona (#pers-MSE) y su función como autor o destinatario (@role=”author/addressee”), el lugar (#place-Ceilhes), al que va asociado el tipo de centro en que se encuentra el refugiado autor de las cartas (refugio, campo de internamiento, colonia infantil, etc.), y por último, la identificación de otras cartas que pertenecen al mismo intercambio epistolar (respuesta del destinatario u otras cartas del mismo emisor).

Figura 6. <correspDesc> en CAREXIL-FR.

[18] Extracto del documento editado en la base: PSCR5285.xml

```
<correspDesc>
        <correspAction type="sent">
          <persName key="ner.xml#pers-MSE" role="author"/>
          <location cert="1" type="origin">
            <placeName key="ner.xml#place-Ceilhes"/>
            <desc cert="1" corresp="ner.xml#shelter"/>
          </location>
          <date cert="1" when="1939-11-27"/>
        </correspAction>
        <correspAction type="received">
          <persName key="ner.xml#pers-RDM" role="addressee"/>
          <location cert="1">
            <placeName key="ner.xml#place-Paris"/>
          </location>
        </correspAction>
        <correspContext>
          <ref type="post" target="2051271">
            Carta anterior de <persName>
              Margarita Suárez</persName> a <persName> Renée de Monbrison
</persName>: <date when="1939-12-22"/>
          </ref>
        </correspContext>
      </correspDesc>
```

Como podemos ver, son todavía muy escasos los proyectos de ámbito hispánico que se han adueñado de esta poderosa herramienta creada en el consorcio TEI para el modelado de metadatos. Volveremos sobre esta cuestión en las conclusiones de nuestro trabajo, para explicar este relativo desinterés. Pero primero nos detendremos en las posibilidades que ofrece la TEI para la anotación del texto mismo de la carta, en función de los variados intereses de investigación.

2. TEI para la anotación pluridisciplinar de textos epistolares

Son diversos los elementos TEI que se han adoptado para caracterizar la estructura típica de las cartas. De nuevo, fue el proyecto DALF el que ofreció los primeros ejemplos en el panorama internacional, adaptando elementos como <salute>, <opener>, <closer>, para las distintas partes reconocibles en una carta, a partir de los elementos disponibles en TEI P4[19] para marcar la estructura de cualquier tipo de texto.

En los proyectos de ámbito hispánico, aquí también encontramos diferentes decisiones editoriales según si los corpus digitales contienen solo cartas o bien otros tipos de documentos además de cartas. Como ocurría con el modelado de metadatos, la anotación de las estructuras textuales en los proyectos generalistas optan por proponer una estructuración común para diferentes tipologías textuales, sin diferenciar en el marcado las particularidades epistolares.

Por el contrario, los proyectos especializados en corpus epistolares, recogen la propuesta de DALF y otros proyectos internacionales (*Vincent Van Gogh: The Letters*, *Letters and Texts: Intellectual Berlin around 1800*, entre otros), ajustándola a la última

[19] https://ctb.kantl.be/project/dalf/dalfdoc/DALFstructure.html

versión del estándar TEI P5. Es el caso de los proyectos ya mencionados supra, Post-Scriptum o CAREXIL-FR, como puede verse en la Figura 7. Las tres principales partes de todo texto epistolar (encabezamiento, cuerpo y despedida) aparecen identificadas con los elementos <opener>, <p> y <closer>, respectivamente, anidados dentro de <body>. El cuerpo se compone de tantos párrafos como aparezcan en la carta que se edita. Además, dentro de estas unidades podemos encontrar nuevas subestructuras, para saludos, <salute>, tanto dentro de <opener> como de <closer>, fecha <date>, firma <signed>, dirección <address>, etc. Asimismo, puede anotarse el nombre del destinatario, cuando aparece, asociándole un identificador único y una función con los atributos @key y @role. Por último, el elemento <postScript> permite introducir el texto de las eventuales postdatas de las cartas. Hemos de hacer notar que todos estos elementos están disponibles en el manual TEI para textos de otros tipos, en particular para textos en prosa, de manera de que no presentan soluciones de marcado ad-hoc para las cartas, pero su combinación permite reflejar adecuadamente los rasgos estructurales epistolares.

Figura 7. Plantilla para la anotación de la estructura textual de una carta en CAREXIL-FR.

```
<text>
    <body>
      <opener>
        <placeName> …</placeName>
        <date>…</date>
        <persName key="ner.xml#pers-xxx" role="addressee"> … </persName>
        <address>
          <addrLine>… </addrLine>
        </address>
        <salute>… </salute>
       </opener>
       <p>…</p>
       <closer>
           <salute>… </salute>
           <signed>…</signed>
           <address>
                <addrLine>… </addrLine>
          </address>
       </closer>
       <postscript>
        <p> </p>
      </postscript>
    </body>
  </text>
```

A estas marcas de la estructura textual de la carta se han añadido otros elementos que permiten dar cuenta de formulaciones rituales en la sección de apertura o cierre de la carta. Se trata de partes opcionales, conocidas como exordio y conclusión, cuya aparición se introduce en Post-scriptum (y siguiendo su modelo, en CAREXIL-FR) con los elementos <seg> y el atributo@type, generalmente anidados en el cuerpo (Figura 8).

Figura 8. Elementos para las divisiones estructurales opcionales de las cartas, exordio y conclusión, en Post-scriptum[20].

```
<p>
        <seg type="harangue"> …</seg>
</p>
 <p>
        <seg type="peroration">… </seg>
 </p>
```

Por el carácter pluridisciplinario del proyecto Post-Scriptum, las anotaciones que se utilizan se destinan a lingüistas e historiadores. Este mismo objetivo pluridisciplinar anima el proyecto CAREXIL-FR, con una orientación pragmática, sociolingüística e histórica. Por ello, ambos proyectos desarrollan anotaciones que permiten la interrogación del corpus epistolar con fines específicos. Así por ejemplo, en CAREXIL-FR se ha creado un sistema de marcado de actos de habla para estudiar la estructura pragmática de los actos de petición, que se descomponen en una multitud de microactos (Blum-Kulka y Olshtain 1984, Blum-Kulka, House y Kasper 1989), cuya aparición puede estudiarse en las cartas por medio de un marcado con el elemento <seg> 'segment', especificado con los atributos @type y @subtype (Figura 9):

Figura 9. Marcado de actos de habla en CAREXIL-FR[21].

```
<p>
        <seg type="act" subtype="explanation">
          <lb/>hemos pasado como hemos
         <lb/>podido por no tener de molestarle pero me encuentra ahora en
         <lb/>una situacion muy apurada
        </seg>
        <seg type="act" subtype="request">
          <lb/>si Vd. pudiera hacer algo le agra-
          <lb/>deceria mucho
        </seg>
</p>
```

Por otro lado, es posible anotar formas lingüísticas que se deseen estudiar en relación con los objetivos de los investigadores. En las recomendaciones de la TEI, se deja abierta la posibilidad de añadir anotaciones lingüísticas en función de las necesidades de los estudiosos, sin limitar las modalidades[22]. Se definen como anotaciones lingüísticas aquellas anotaciones que permiten el análisis de los rasgos lingüísticos de un texto, distintos de las propiedades estructurales del mismo o de la información descriptiva sobre su contexto de producción o uso[23]. La etiqueta puede consistir en asociar un código a cada palabra o unidad (por ejemplo, para la categoría morfosintáctica de una palabra) o a un grupo de palabras. La adscripción de estos códigos puede ser automática, manual o mixta.

[20] http://teitok.clul.ul.pt/postscriptum/files/ps_doc.html#ps_formulaic
[21] Documento CAREXIL_4140036.
[22] "The present Guidelines do not advocate any particular approach to linguistic anotation (or 'tagging'); instead a number of general analytic facilities are provided which support the representation of most forms of annotation in a standard and self-documenting manner", https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html#CCAN
[23] https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html#CCAN1

El manual TEI ofrece ejemplos de este tipo de anotación lingüística, que se reproduce aquí (Figura 10), donde el elemento <s> 'sentence' delimita la oración, <w> 'word', cada palabra, y el atributo @ana 'analysis', proporciona la interpretación del código. TEI propone también otros elementos para distintos niveles sintácticos, como sintagma <phr> 'phrase', cláusula <cl> 'clause' o morfema <m> 'morph'. Los códigos usados por cada proyecto deben ser explicitados adecuadamente e incorporados en los metadatos del documento, dentro del elemento <interpretation>[24], en la descripción de codificación (<encodingDesc>) de la cabecera (<teiHeader>).

Figura 10. Ejemplo de anotación lingüística en TEI (Guidelines[25]).

```
<s>
   <w ana="#AT0">The </w>
   <w ana="#NN1">victim</w>
   <w ana="#POS">'s</w>
   <w ana="#NN2">friends </w>
   <w ana="#VVD">told </w>
   <w ana="#NN2">police </w>
   <w ana="#CJT">that </w>
   <w ana="#NP0">Kruger </w>
   <w ana="#VVD">drove </w>
   <w ana="#PRP">into </w>
   <w ana="#AT0">the </w>
   <w ana="#NN1">quarry </w>
   <w ana="#CJC">and </w>
   <w ana="#AV0">never </w>
   <w ana="#VVD">surfaced</w>
</s>
```

Asimismo, la clase att.linguistics proporciona una lista de atributos que pueden utilizarse para caracterizar las propiedades lingüísticas de las unidades (token), en particular las palabras (<w>): @lemma, @pos 'part-of-speech', @msd 'morphosyntactic description', etc[26].

Muchos de los proyectos de correspondencia aparecidos en los últimos años en el ámbito hispánico se basan en el entorno TEITOK creado por Maarten Janssen (2014-). Es el caso de los dos proyectos de correspondencias que hemos citado aquí, Post-Scriptum y CAREXIL-FR. Además, la Red CHARTA ha adoptado recientemente esta herramienta y se está procediendo a la migración de los corpus. Aunque TEITOK es un entorno que utiliza el marcado XML-TEI, sus especificidades lo alejan en algunos casos del estándar TEI P5, como por ejemplo en la identificación de cada token, como se ve en la figura 11:

[24] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-interpretation.html

[25] TEI Guidelines: https://tei-c.org/release/doc/tei-p5-doc/en/html/AI.html#AILA. En este ejemplo, los códigos representan la categoría morfosintáctica de las palabras anotadas: #AT0 'artículo', #NN1 'nombre común singular', #POS 'posesivo', etc.

[26] Existen otros módulos de anotación lingüística en TEI para las dependencias sintácticas (Graphs, Networks and Trees, cf. https://tei-c.org/release/doc/tei-p5-doc/en/html/GD.html) o rasgos estructurales (<fs> feature structure, cf. https://tei-c.org/release/doc/tei-p5-doc/en/html/FS.html), sobre los que no podemos extendernos aquí.

Figura 11. Anotación TEITOK/TEI P5 de la secuencia: “la tenemos enferma con una polmunia”. CAREXIL_4144042.

<tok id="w-87" lemma="la" mfs="PP3FSD00">la</tok>	<w lemma="la" xml:id="w-87">la</w>
<tok id="w-88" lemma="tener" mfs="VMIP1P0">tenemos</tok>	<w lemma="tener" xml:id="w-88">tenemos</w>
<tok id="w-89" form="enferma " mfs="AQ0FS0" lemma="enfermo">enfer-<lb n="10" id="e-13"/> ma</tok>	<w lemma="enfermo" xml:id="w-89">enfer-<lb n="10" id="e-13"/> ma</w>
<tok id="w-92" lemma="con" mfs="SPS00">con</tok>	<w lemma="con" xml:id="w-92">con</w>
<tok id="w-93" lemma="uno" mfs="DI0FS0">una</tok>	<w lemma="uno" xml:id="w-93">una</w>
<hi rend="underlined" resp="CAEERF"> <tok form="polmunia" id="w-94" nform="pulmonía" lemma="pulmonía " mfs="NCFS000">polmunia</tok></hi>	<hi rend="underlined" resp="CAEERF"> <choice> <orig> <w lemma="pulmonía " xml:id="w-94">polmunia</w> </orig> <reg> <w>pulmonía</w> </reg> </choice> </hi>
<tok id="w-95" lemma="," mfs="Fc">,</tok>	<pc lemma="," xml:id="w-95">,</pc>

La versión TEI P5 es la que ofrece el convertidor integrado en la plataforma TEITOK, gracias al cual se asegura la interoperabilidad con otros proyectos en TEI. Observemos algunas diferencias que se dan entre las dos versiones: el elemento <w> ‘word’ y <pc> ‘ponctuation character’ se marcan con un mismo elemento <tok> en la plataforma TEITOK. Además de los atributos @xml:id (en TEITOK @id) y @lemma, TEITOK ofrece para cada token un atributo de categoría morfosintáctica @mfs, al que se asocia un código lingüístico que remite al etiquetario del consorcio EAGLES[27]. En la versión TEI P5, no se ofrece ningún equivalente, dado que no hay una codificación universal para las categorías morfosintáctias, sino que cada proyecto, como se ha explicado más arriba, debe definirlas en su propio perímetro. Sí habría sido posible sin

[27] Expert Advisory Group on Language Enfineering Standards, Consejo de Europa. Cf. sus principales recomendaciones en http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html Recommendations for the Morphosyntactic Annotation of Corpora, EAG--TCWG--MAC/R, Mar, 1996.

embargo utilizar con esta misma función el atributo @pos, 'part-of-speech', propuesto en TEI.

Destaca por último el diferente tratamiento que se ofrece para las unidades normalizadas: en TEITOK, la forma original aparece identificada con el atributo @form, y la forma normalizada con el atributo @nform, mientras que en TEI P5, se utiliza el elemento <orig> 'original form' y <reg> 'regularization', anidados en <choice>, elemento que permite introducir variantes.

Fuera de las etiquetas de lema, categoría morfosintáctica o normalizaciones, TEITOK ofrece la posibilidad de introducir atributos con información lingüística específica. Dichos atributos toman la forma @ling y sus valores responden a códigos definidos libremente por cada proyecto. En el caso de CAREXIL-FR, se asocia un código para algunas de las construcciones sintácticas que se desean estudiar, como las completivas asindéticas, que se asocian al código "sint-asind" 'sintaxis asindética'. Esta misma anotación podría aparecer en lenguaje TEI P5 como se muestra en la segunda columna de la figura 12, en contraste con la anotación en TEITOK (columna 1). Nótese que la anotación lingüística aparece en CAREXIL-FR asociada al nivel del token, en este caso, el verbo de la subordinada[28].

Figura 12. Anotación lingüística en TEITOK y en TEI P5[29].

1. TEITOK	2. TEI P5
<tok>Rogando</tok> <tok>me</tok> <tok ling="sint-asind"> perdonen </tok> <tok>lo</tok> <tok>estensa</tok> <tok>que</tok> <tok>soy</tok>	<s> <w>Rogando</w> <cl ana="#asind"> <w>me</w> <w>perdonen</w> <w>lo</w> <w>estensa</w> <w>que</w> <w>soy</w> </cl> </s>

3. Conclusiones

Los párrafos anteriores han permitido ofrecer algunas calas en los diferentes usos que se están haciendo del marcado TEI en proyectos digitales de edición de cartas en español, poniéndolos en perspectiva con otros proyectos internacionales. Como se ha podido comprobar, los proyectos del ámbito ibérico (España y Portugal) se caracterizan por emanar de una fuerte tradición filológica, muy vinculada a la lingüística histórica. Así, vemos aparecer proyectos de edición digital preocupados por la fidelidad a las fuentes textuales y, a la vez, por la aplicación de metodologías propias de la lingüística de corpus. Con razón, se ha reconocido el carácter pionero de Post-Scriptum a la hora de asociar una plataforma de edición digital y un corpus lingüístico interrogable (Henny-Krahmer 2019). Esta doble necesidad se ha visto satisfecha gracias al entorno TEITOK

[28] TEITOK ofrece la posibilidad de anotar también oraciones y cláusulas, pero no se ha activado todavía esta posibilidad en CAREXIL-FR.
[29] Documento CAREXIL_7240010.

de Janssen (2014), que fue creado precisamente en el marco de este proyecto. No es de extrañar, por consiguiente, que sea precisamente esta herramienta la que está extendiéndose de manera muy amplia para otros proyectos hispánicos de edición digital asociados a búsquedas en el corpus. Aunque el uso de TEI en TEITOK no responde exactamente al estándar TEI P5, la puesta a disposición de un convertidor permite asegurar la interoperabilidad con otras plataformas.

Son escasos sin embargo los proyectos que en esta área se especializan en la edición y explotación de correspondencias, con pocas excepciones (entre ellas, las dos mencionadas aquí: Post-Scriptum y CAREXIL-FR, dos proyectos pluridisciplinares, con un fuerte componente a la vez histórico y lingüístico), pues en la mayoría de los casos los corpus editados incluyen diversos tipos de documentación para la investigación en lingüística histórica.

Este objeto de estudio más amplio no favorece la utilización de herramientas que posibilitarían sacar el máximo rendimiento de los corpus epistolares, en particular aquellos aspectos basados en la estructura reticular de este tipo de textos que han servido en proyectos internacionales para la exploración y visualización de diferentes tipos de redes (epistolares, sociales, geográficas…). (cf. sobre este punto en particular las recomendanciones del grupo Correspondance CAHIER en Walter 2018).

Por otro lado, estos proyectos internacionales, concebidos muchas veces como repositorios de cartas, se han desarrollado gracias a la anotación de metadatos específicos para la correspondencia, en particular el módulo TEI <correspDesc>. Gracias a él, se han podido comenzar a enlazar fondos epistolares diferentes y a relacionar diversas redes de corresponsales.

Cabe pensar que los proyectos hispánicos podrían beneficiarse de estos desarrollos si se incluyera en ellos, como han empezado haciendo Post-Scriptum y CAREXIL-FR, el módulo <correspDesc>. Su interés se extiende, a nuestro modo de ver, tanto a aquellos proyectos que únicamente editan cartas como a los que las incluyen entre otras tipologías de documentos, ya que, como hemos visto, permite dar cuenta de la estructura reticular característica del género epistolar: la carta no existe por sí sola, sino en el interior de redes de cartas, y en relación con redes de corresponsales así como de redes de puntos geográficos (Walter 2018). Su modelado como metadatos específicos facilitaría su posterior visualización gracias a herramientas ad-hoc como cartografías y grafos. Por otro lado, este módulo ofrece una sólida estructura para proyectar bases futuras en que se mutualicen los corpus epistolares disponibles de manera interoperable, como la herramienta <correspSearch> ya mencionada.

A su vez, la tradición filólogica y el perfil lingüístico de los corpus de ámbito hispánico pueden aportar un desarrollo más completo a los proyectos epistolares internacionales que ofrecen pocas posibilidades de búsquedas en los textos, pues, aparte de la indexación, la anotación cubre generalmente, en el mejor de los casos, algunas palabras clave y no llega al detalle que se ha mostrado en los corpus españoles o sobre el español.

Una de las grandes ventajas del marcado TEI es que puede completarse progresivamente, por medio de capas sucesivas de anotación, a medida que se van vislumbrando posibilidades de explotación y estudio de los textos. Apostemos por que los textos epistolares en español puedan beneficiarse en un futuro próximo de todas las posibilidades de su tratamiento digital por medio de bases de datos, ediciones filológicas en línea, motores de búsqueda, cartografía y visualización.

Bibliografía:

Aranda, Marcelo (dir.). *An Intellectual Map of Science in the Spanish Empire, 1600-1810*. http://republicofletters.stanford.edu/casestudies/spanishempire.html

Baillot, Anne (ed). 2012. *Letters and Texts: Intellectual Berlin around 1800*. Berlin: Humboldt-Universität zu Berlin. <http://www.berliner-intellektuelle.eu/>

EAGLES. 1996. Recommendations for the Morphosyntactic Annotation of Corpora, EAG--TCWG--MAC/R. Consejo de Europa. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html

Bravo-García, Eva, Ana Mancera Rueda y Leyre Martín Aizpuru (dirs.). *Corpus Documental y Hemerográfico de la Cuba del Novecientos (CODHECUN)*, disponible en línea: <http://cuba19.us.es/> [fecha de consulta: 15/10/2022]

Castillo Lluch, Mónica y Elena Díez del Corral (dirs.). *COSUIZA*: *Corpus de documentos hispánicos de Suiza*. https://cosuiza.unil.ch/teitok/cosuiza/index.php?action=investigadores

Centrum voor Teksteditie en Bronnenstudie, Koninklijke Academie voor Nederlandse Taal- en Letterkunde (Centre for Scholarly Editing and Document Studies, Royal Academy of Dutch Language and Literature). 2002-. *DALF: Digital Archive of Letters in Flanders*. http://ctb.kantl.be/project/dalf/index.htm.

CLUL (ed.). 2014. *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. [15/10/2022]. URL: http://ps.clul.ul.pt.

Henny-Krahmer, 'Bridging Edition and Corpus: A Review of P. S. Post Scriptum: A Digital Archive of Ordinary Writing (Early Modern Portugal and Spain) – RIDE' <https://ride.i-d-e.de/issues/issue-10/post-scriptum-digital-archive/?hilite=post-scriptum>.

Isasi Martínez, Carmen, Leyre Martín Aizpuru, Santiago Pérez Isasi, Elena Pierazzo, y Paul Spence. 2020. *Edición digital de documentos antiguos: marcación XML_TEI basada en los criterios CHARTA*. Sevilla: Editorial Universidad de Sevilla. <https://dialnet.unirioja.es/servlet/libro?codigo=780326> [accessed 23 October 2022]

Isasi Martínez, Carmen, y Paul Spence, eds. 2014. *Guía Para Editar Textos CHARTA Según El Estándar TEI: Una Propuesta*. [En línea] http://www.charta.es/investigacion/charta-tei/.

Jansen, Leo, Hans Luijten, and Nienke Bakker (eds.). 2009. *Vincent van Gogh. The Letters.* http://vangoghletters.org/ .

Janssen, Maarten (2014-). *TEITOK. A tokenized TEI environement*. http://www.teitok.org

Lazure, Guy (dir.). *Spanish Republic of Letters.* University of Windsor (Canada). http://cdigs.uwindsor.ca/srl/

Author (coord.). 2019-. *CAREXIL-FR: Cartas de Republicanos Españoles Refugiados y Exiliados en Francia*. http://carexil.huma-num.fr

López Piñero, José M., Thomas Glick, Navarro Brotón, Víctor y Portela Marcos, Eugenio (eds.). 1983. *Diccionario histórico de la ciencia moderna en España.* 1985. Madrid. Ediciones Península. 2 vols.

Marquilhas, Rita (dir.). 2006-2009. *CARDS: Unknown Letters*. Universidade de Lisboa.

Marquilhas, Rita (coord.). 2010-2013. *FLY: Forgotten Letters. 1900-1974.* Universidade de Lisboa. http://fly.clul.ul.pt/

Stadler, Peter, Marcel Illetschko y Sabine Seifert. 2016. « Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc> », *Journal of the Text Encoding Initiative* [Online], Issue 9 | September 2016 - December 2017, Online since 24 September 2016. http://journals.openedition.org/jtei/1433 ; DOI : 10.4000/jtei.1433

Petrucci, Armando. 2008. *Scrivere Lettere. Una Storia Plurimillenaria*. Roma-Bari: Laterza.

Quintana, Aldina. 2012-. *CoDiAJe - The Annotated Diachronic Corpus of Judeo-Spanish*. http://corptedig-glif.upf.edu/teitok/codiaje

Quintana, Aldina. 2020. "*CoDiAJe - The annotated Diachronic Corpus of Judeo-Spanish.* Description of a Multi-alphabetic corpus and its textual and linguistic annotations*". Scriptum Digital,* 9, p. 209-236.

REAL ACADEMIA ESPAÑOLA. Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.* <http://www.rae.es> [15/10/2022]

REAL ACADEMIA ESPAÑOLA. Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.* <http://www.rae.es> [15/10/2022]

REAL ACADEMIA ESPAÑOLA. Banco de datos (CORPES XXI) [en línea]. *Corpus del español del Siglo XXI.* <http://www.rae.es> [15/10/2022]

REAL ACADEMIA ESPAÑOLA. 2013 (revisado 2020). *Corpus del español del siglo XXI. Descripción del sistema de codificación. Libros y prensa.* [Recurso en línea]. https://www.rae.es/banco-de-datos/corpes-xxi

Sánchez-Prieto Borja, Pedro (dir.). Red Charta. https://www.redcharta.es/

TEI Consortium. 2015. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.5.0. Last updated on October 22. https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

University of Stanford. *Mapping the Republic of Letters.* http://republicofletters.stanford.edu/

Vaamonde, Gael. 2018. "Escritura Epistolar, Edición Digital y Anotación de Corpus". *Quadernos Del Instituto Historia de La Lengua*, 11, 139–64

Vaamonde, Gael, Ana Luísa Costa, Rita Marquilhas, Clara Pinto y Fernanda Pratas, «Post Scriptum: Archivo Digital de Escritura Cotidiana», en *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, Sagrario López Poza y Nieves Pena Sueiro (editoras), Janus [en línea], Anexo 1 (2014), 473-482, publicado el 11/04/2014, consultado el 02/11/2022. URL: https://www.janusdigital.es/anexos/contribucion.htm?id=41

Walter, Richard. 2018. *L'édition Numérique de Correspondances. Guide Méthodologique*. <https://cahier.hypotheses.org/guides/guide-correspondance>