CRISTINA SÁNCHEZ MARCO
Universitat Pompeu Fabra
JOSEP MARIA FONTANA
Universitat Pompeu Fabra
GEMMA BOLEDA
Universitat Politècnica de Catalunya

Propuesta de codificación de la información paleográfica y lingüística para textos diacrónicos del español. Uso del estándar TEI*

Introducción

El rápido desarrollo de la tecnología y la creación de corpus digitalizados han transformado radicalmente el mundo de la lingüística y la filología en las últimas décadas. Estos recursos han abierto nuevas vías de investigación hasta ahora inimaginables o inviables. Los datos obtenidos gracias a los corpus y otras herramientas computacionales permiten al lingüista localizar más fácilmente cambios concretos en la evolución de una lengua, y también descubrir tendencias generales en el cambio lingüístico que serían difíciles de explorar de otra manera. En esta línea de investigación destacan, por ejemplo, el estudio realizado por Han y Kroch (2000) sobre el origen del verbo auxiliar *do* a partir de datos del *Penn-Helsinki Parsed Corpus of Middle English*, o el de Sagi, Kaufmann y Clark (2009), que utilizan un corpus derivado del corpus Helsinki (Rissanen 1994) para estudiar el cambio semántico de las palabras *dog, deer* y *do*.

Actualmente no existe un corpus histórico del español accesible a la comunidad científica con el que se puedan realizar estudios empíricos con la profundidad y el alcance que caracteriza a los estudios que aca-

* Este trabajo ha sido realizado en parte gracias a una beca FPU (AP2006-03547) del Ministerio de Educación.

bamos de citar. El *CORDE*[1] de la Real Academia y el *Corpus del español*[2] (Davies 2002), a pesar de la cantidad y la calidad de los documentos que contienen, tienen una utilidad muy limitada cuando lo que se quiere estudiar son fenómenos de tipo sintáctico, semántico o discursivo, ya que no están anotados con información lingüística suficiente y sus interfaces no permiten búsquedas complejas que utilicen este tipo de información.

Con el propósito de cubrir esta laguna hemos iniciado un proyecto para construir un corpus diacrónico del español anotado con información lingüística que nos permita realizar estudios empíricos sobre la evolución de diferentes fenómenos de carácter sintáctico y semántico en la historia de la lengua española. En este trabajo describimos la parte de este proyecto que se refiere a la representación de las anotaciones en este corpus, es decir, el modo en que la información lingüística, paleográfica y estructural se presenta técnicamente y cómo estos diferentes tipos de anotación se relacionan con la fuente de datos primarios o textos originales. Concretamente, presentamos una propuesta de representación de este tipo de anotaciones en formato XML, siguiendo el estándar TEI.

Aunque el estándar TEI ya se ha utilizado para representar corpus sincrónicos en otras lenguas –como por ejemplo es el caso del *British National Corpus*[3] y el *National Corpus of Polish* (Banski y Przepiórkowski 2009)–, no existe una propuesta de uso de las especificaciones TEI para representación de corpus históricos en español.

1. El corpus: textos y anotaciones

Los textos con los que estamos trabajando son las ediciones electrónicas del *Hispanic Seminary of Medieval Studies* (en adelante HSMS), formadas por textos peninsulares de diferentes géneros de los siglos XII al XVI, y que en conjunto están formados por más de 20 millones de pala-

1 <http://www.rae.es>.

2 <http://www.corpusdelespanol.org>.

3 <http://www.natcorp.ox.ac.uk/>.

bras[4]. Estos textos, y en general los corpus diacrónicos formados por ediciones críticas, presentan algunas dificultades de representación adicional a las que tienen los corpus sincrónicos. Estos textos contienen símbolos semipaleográficos, es decir, anotaciones que describen la información de los manuscritos relativa a las inserciones, supresiones, abreviaturas del escriba o del editor, y también sobre otras características físicas de los manuscritos, como la disposición del texto en folios y columnas o la existencia de miniaturas o diagramas[5]. Este tipo de información es muy relevante para los investigadores de la historia de la lengua y por tanto es conveniente que se mantenga y represente de manera adecuada en las ediciones digitales. A esta información ya contenida en los textos originales hay que añadir la información lingüística, como por ejemplo de los lemas y las categorías morfológicas, fundamental si lo que se desea es realizar estudios sobre la evolución lingüística[6]. En la Figura 1 se puede ver el tipo de anotaciones que incluyen estos textos, a la cual habría que añadir la información lingüística deseada, como por ejemplo la de los lemas y las categorías morfológicas.

```
[fol. 2r]
{CB2.
del seso aquello que entendiesse omne
q<ue> mas su pro fuesse
```

Figura 1. Texto electrónico original. Fragmento de los *Libros de Ajedrez y Dados* de Alfonso X el Sabio, contenido en la colección del HSMS

4 Ver Corfis *et al.* (1997), Herrera y González de Fauve (1997), Kasten *et al.* (1997), Nitti *et al.* (1997), O'Neill (1999), Sánchez *et al.* (2003), Waltman (1999).

5 En el manual de transcripción de Madison (Mackenzie y Burrus 1986) están descritos todos los símbolos y el modo en el que han sido utilizados para transcribir los documentos.

6 En Sánchez Marco *et al.* (2010, 2011) se puede ver la estrategia desarrollada para para anotar con información lingüística de manera automática los textos del HSMS.

2. Estándares de anotación y representación. Text Encoding Initiative

Para que los corpus sean reutilizables durante un largo periodo de tiempo y útiles para los objetivos de diferentes investigadores es fundamental utilizar un estándar de representación y anotación. El desarrollo de un recurso sostenible asegura su disponibilidad, accesibilidad y posterior reutilización[7].

En el caso de documentos históricos, es fundamental su digitalización y transcripción crítica y paleográfica, pero no menos importante es el modo en el que se representan los textos que estos documentos contienen y los tipos de anotación incluidos. Concretamente, el uso de modelos de anotación estándar para representar los datos facilita el desarrollo de recursos sostenibles. Un estándar de anotación es un modelo de referencia que ha sido aceptado –generalmente de manera internacional–, y está formado por un conjunto de reglas y procedimientos bien elaborados y documentados, que son ampliamente utilizados e integrados en herramientas de creación y procesamiento de corpus. Como señalan Ide y Romary (2007), un estándar de anotación permite el uso y la reutilización flexible de los datos, lo que es esencial para la siguiente generación de herramientas lingüísticas. Además un estándar de anotación normalmente facilita y provee el mantenimiento y soporte necesarios para el desarrollo de los recursos.

En las últimas décadas han surgido numerosos estándares e iniciativas para la representación de los datos primarios y sus anotaciones. Entre ellos, destacan TEI[8], CES y XCES[9], para la representación de las anotaciones; y EAGLES[10] e ISLE, para el contenido de las anotaciones[11].

Concretamente, para representar este corpus de textos diacrónicos del español hemos utilizado el estándar TEI. TEI es un consorcio que desarrolla y mantiene un estándar para la representación de textos digitales. Este

7 Ver Bird y Simons (2003), Ide y Romary (2004), Lay y Demonet (2008), Lehmberg y Wörner (2008), Zinsmeister *et al.* (2008), Zinsmeister *et al.* (2009).

8 Text Encoding Initiative, <http://www.tei-c.org>.

9 Corpus Encoding Standard, <http://www.xml-ces.org>.

10 Expert Advisory Group on Language Engineering Standards, http://www.ilc.cnr.it/EAGLES96/home.html>.

11 International Standard for Language Engineering.

estándar se viene desarrollando desde 1987 y ha sido y es utilizado para codificar grandes corpus, como por ejemplo el *British National Corpus*. El esquema y vocabulario de codificación TEI se expresa en XML (eXtensible Markup Language), el metalenguaje de marcado más utilizado para todo tipo de recursos digitales desarrollado por el World Wide Web Consortium (W3C)[12].

Entre las partes que forman un documento XML se distinguen las *etiquetas*, que aparecen entre paréntesis angulares, y su *contenido* con los datos anotados. A su vez, las etiquetas de un documento XML pueden ser *elementos*, que describen los datos, o *atributos*, que especifican características o propiedades de los elementos que los contienen. Por ejemplo, la palabra *seso* del texto de la Figura 1 se puede marcar en formato XML con la etiqueta *<w lemma="seso" type="sustantivo" >seso</w>*, utilizando el elemento *w* con los atributos *@lemma* y *@type* para señalar su lema y categoría morfológica correspondiente[13].

TEI recomienda un conjunto de elementos y atributos para marcar datos utilizando el formato XML. Una de las principales ventajas de las especificaciones propuestas por esta iniciativa es su estructura modular, que posibilita que el usuario elija los elementos y atributos de cada módulo más adecuados a los datos. Además de los módulos que contienen elementos y atributos globales asociados a todo tipo de documentos (específicamente los módulos *TEI Infrastructure, Common Metadata, Common Core, Default Text Structure*), también existe un módulo de codificación de manuscritos (*Manuscript Description*), que permite marcar la información paleográfica y estructural que aparece en los textos diacrónicos.

3. Propuesta de codificación de los textos utilizando el estándar TEI

Las especificaciones de TEI permiten representar las anotaciones de un corpus de manera interna o externa. En la representación interna se incluyen las anotaciones XML en los textos originales. En la representa-

12 <http://www.w3.org/XML>.

13 La @ delante de una palabra en el texto indica un atributo.

ción externa, sin embargo, se separa el texto original de los metadatos o anotaciones que lo describen[14].

3.1. Representación interna

La representación interna de un texto original del corpus, como por ejemplo el que aparece en la Figura 1, está ilustrada en la Figura 2, en el que los símbolos paleográficos del texto original han sido sustituidos por el marcado equivalente en XML y la información relativa a los lemas y las categorías morfológicas ha sido añadida, siguiendo el estándar TEI.

```
<text>
<pb n="2" type="r"/>
  <cb n="2"/>
    <lb n="1"/>
        <w lemma="del" type="SPCMS">del</w>
        <w lemma="seso" type="NCMS000">seso</w>
        <w lemma="aquel" type="PD0NS000">aquello</w>
        <w lemma="que" type="PR0CN000">que</w>
        <w lemma="entender" type="VMSI1S0">
          <choice>
            <orig>entendiesse</orig>
            <reg>entendiese</reg>
          </choice></w>
        <w lemma="hombre" type="NCMS000">omne</w>
    <lb n="2"/>
        <w lemma="que" type="PR0CN000">
          <choice>
            <abbr>q</abbr>
            <expan>que</expan>
          </choice></w>
        ....
</text>
```

Figura 2. Representación interna en XML de un fragmento de los *Libros de Ajedrez y Dados* de Alfonso X el Sabio, según el estándar TEI

14 En los anexos A y B se listan los elementos y atributos propuestos para representar la información lingüística y paleográfica utilizando las especificaciones de TEI.

Entre los elementos de TEI utilizados en la representación interna de este texto, se pueden distinguir los que marcan algunas de las características físicas de los manuscritos, como por ejemplo *pb*, *cb* y *lb*, que indican el comienzo de folio, columna y línea respectivamente; y los que marcan el comienzo y final de cada palabra (elemento *w*), y su correspondiente lema y categoría (atributos *@lemma* y *@type*).

El conjunto de etiquetas utilizado como valor del atributo *@type* está basado en las etiquetas propuestas por el grupo EAGLES (Expert Advisory Group on Language Engineering Standards) para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. Por ejemplo, el valor *NCMS000*, correspondiente a la palabra *omne*, significa 'nombre común masculino singular'.

La etiqueta *choice* permite codificar diferentes alternativas para las palabras. Por ejemplo en la Figura 2, *choice* incluye la forma original del manuscrito "entendiesse", con la doble "s" (etiqueta *orig*), y la forma regularizada o estandardizada "entendiese" (etiqueta *reg*). En el siguiente caso de la misma figura, *choice* contiene la abreviatura "q" (etiqueta *abbr*) y la expansión del editor "que" (etiqueta *expan*). La codificación de las diferentes alternativas de una misma forma, ya sea una correspondiente a una forma abreviada y expandida o paleográfica y regularizada, facilita posteriormente la visualización de la edición crítica y la original de los textos a partir de la misma representación.

A pesar de que en este tipo de representación todas las anotaciones están incluidas en un solo documento, la representación interna tiene una desventaja fundamental. El texto original es modificado durante el procesamiento, lo que dificulta la adición posterior de cualquier otro tipo de anotación. Por ejemplo, si el documento de la Figura 2 se quisiera enriquecer con información sintáctica, es decir, sobre los constituyentes de la oración, el procesamiento del texto para añadir las etiquetas XML sería más complejo, ya que se deberían tener en cuenta todas las etiquetas que ya estarían previamente incorporadas al documento original. Además, cuando hay varios niveles de anotación la representación interna puede llegar a ser muy compleja, y se pueden producir conflictos de jerarquías o anidamientos de etiquetas no permitidos por el formato XML. Por ejemplo, en los textos del HSMS puede aparecer una palabra entre dos columnas, y con este tipo de representación no se podría representar como una unidad sin perder información del manuscrito original.

3.2. Representación externa[15]

En el tipo de representación externa el texto original no sufre ninguna modificación, y cada tipo de anotación está contenido en un archivo diferente, almacenado de manera independiente, y unido al original por medio de los indicadores adecuados. Los archivos con las anotaciones no contienen directamente el texto, como en el anterior tipo de representación, sino una referencia a él.

El elemento recomendado por TEI para referenciar a los datos primarios es *xi:include*, que forma parte de *XInclude*, un módulo específico de procesamiento y sintaxis de inclusión definido por el W3C. Dentro de esta etiqueta, TEI sugiere utilizar los atributos *@href*, para indicar el archivo al que se refiere el documento externo, *@parse*, para indicar si el tipo de documento de datos primarios está en formato XML o en formato de texto, y *@xpointer*, para indicar los índices entre caracteres del documento referido.

La Figura 3 ilustra la arquitectura de la representación externa del corpus enriquecido con información morfológica. Todos los archivos son documentos anotados según las especificaciones de TEI, y forman una jerarquía de anotaciones, tal y como está representada en la figura, de tal manera, por ejemplo, que para añadir la información de las categorías morfológicas a las palabras es necesario que el nivel de segmentación esté correctamente marcado, ya que la anotación de este tipo de información lingüística se realiza para cada palabra.

Como se puede ver en la figura, los *datos primarios* son el texto al cual se refieren los documentos que contienen las anotaciones externas, referidas a la estructura, la segmentación, los símbolos paleográficos y la morfosintaxis. Este documento que contiene los datos primarios no es el texto editado por el HSMS sino otro obtenido del preprocesamiento de aquel y que sería similar al manuscrito original escrito por el escriba. Por ejemplo, en los datos primarios no aparece *q<ue>*, que representa la expansión de una abreviatura en el texto original del HSMS, sino *q,* que es la forma abreviada que aparece en el manuscrito.

15 También llamada anotación *stand-off.*

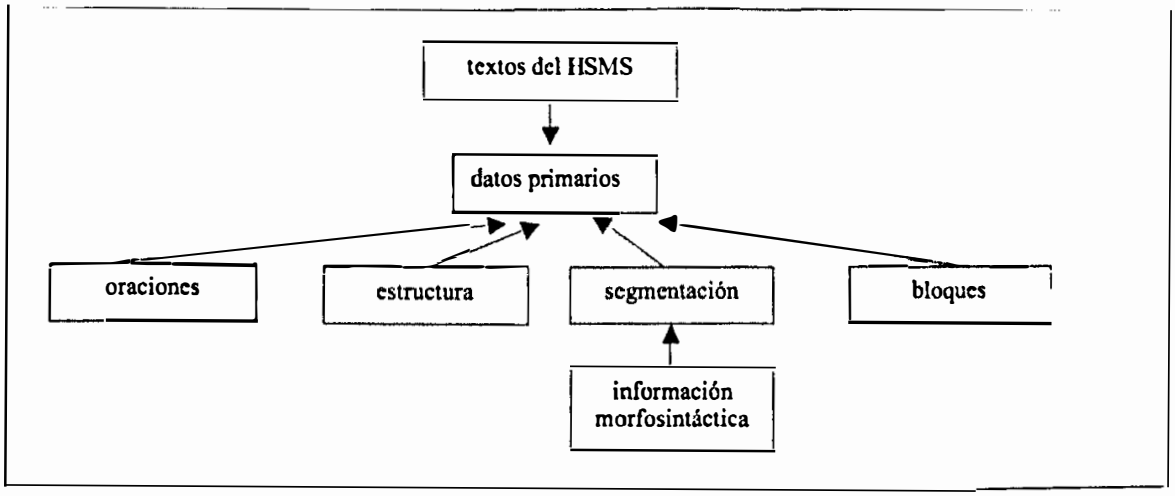

Figura 3. Arquitectura de la representación externa del corpus

En la representación externa cada texto que forma parte del corpus se conserva en un directorio separado, cada uno de los cuales contiene los documentos señalados a continuación, anotados en formato XML según las especificaciones de TEI y con los indicadores adecuados que se refieren a los datos primarios.

- *Estructura*. Contiene la información estructural física de los manuscritos, es decir, las referencias a las líneas (etiqueta *lb*), columnas (etiqueta *cb*), folios (etiqueta *pb*).
- *Oraciones*. Este archivo contiene las anotaciones referidas a las oraciones del texto, indicadas por medio de la etiqueta *s*.
- *Bloques*. Este documento contiene las referencias a los datos primarios que representan un elemento básico del documento, superior a una palabra y diferente a una oración. Suelen ser bloques de texto similares a los párrafos, o a otros elementos textuales o gráficos contenidos en los textos del HSMS, como las glosas (*gloss*), los comentarios del editor (*add*), las miniaturas y los diagramas (*figure* y *graphic*), los *marginalia* (*additions*), las cabeceras (*head*), y los símbolos (*g*).
- *Segmentación*. En este documento se representa la segmentación básica de palabras (con el elemento *seg*), necesaria para poder realizar el etiquetado morfológico. Por medio del atributo *@type* se marca cualquier tipo de edición realizada por el escriba o editor (atributo *@hand*), como por ejemplo una expansión, eliminación, reconstrucción o inserción. Aquí también se representan caracte-

rísticas tipográficas del texto como las rúbricas, las iniciales, o fragmentos del texto en otra lengua, por medio del atributo *@subtype*.

- *Morfosintaxis*. Este documento contiene la información morfosintáctica de lema y categoría asociada a cada palabra. El elemento de TEI utilizado para anotar las palabras y su lema y etiqueta morfológica correspondiente es *w*, con los atributos *@lemma* y *@type* respectivamente.

En las siguientes figuras se ilustran, con el mismo fragmento de texto del HSMS de la Figura 1, los índices entre caracteres de los datos primarios utilizados para incluir las referencias en los documentos que forman la anotación externa (Figura 4), y el documento que contiene las anotaciones externas referidas a la segmentación (Figura 5).

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
|d|c|l| |s|c|s|o| |a|q|u|c|l|l|o| |q|u|c| |c|n|t|c|n|d|i|c|s|s|c| |o|m|n|c|

38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
|q| |m|a|s| |s|u| |p|r|o| |f|u|c|s|s|c|.|
```

Figura 4. Representación de los datos primarios con los índices entre caracteres

```
<text xmlns:xi="http://www.w3.org/2001/XInclude">
<seg xml:id="1-seg">
<xi:include href="datos_primarios.xml" parse="xml" xpointer="string-range(element(/1), 0, 3)"/>
</seg>
<seg xml:id="2-seg">
<xi:include href="datos_primarios.xml" parse="xml" xpointer="string-range(element(/1), 4, 8)"/>
</seg>
...
<seg xml:id="8-seg" type="scribal_abbreviation">
<xi:include href="datos_primarios.xml" parse="xml" xpointer="string-range(element(/1), 38, 39)"/>
</seg>
...
</text>
```

Figura 5. Fragmento del documento externo con las anotaciones de la segmentación

Aunque los diferentes niveles de información están almacenados separadamente en la representación externa, dependen los unos de los otros. Esto significa, por ejemplo, que cualquier error en el documento que contiene las anotaciones sobre la segmentación se extiende también al nivel de la marcación morfosintáctica. En Banski y Przepiórkowski (2009) se señalan algunas de las dificultades de representación encontradas al utilizar el estándar TEI para realizar la representación externa del *National Corpus of Polish* (*NCP*), como por ejemplo cuando lo que se

quieren anotar son elementos enclíticos. En general, a pesar de estas dificultades, las ventajas sobre la representación interna son evidentes, sobre todo si se quieren añadir varios tipos de anotación a los textos en el mismo momento de crear la primera versión del corpus, o posteriormente. El texto original no es modificado, y por tanto permanece estable y seguro, lo que facilita la incorporación de nuevos niveles de anotación e implica que cualquier modificación hecha sobre un nivel de anotación tiene efectos mínimos en otros niveles. Además, el hecho de que cada tipo de anotación está almacenado de manera independiente posibilita tener diferentes niveles de anotación que con la representación interna formarían entrecruzamientos de etiquetas. Una ventaja adicional en cuanto al proceso de creación del corpus es que el equipo de trabajo puede trabajar en diferentes niveles al mismo tiempo.

4. Conclusiones

La representación estandardizada de las anotaciones y los datos en los corpus y las ediciones digitales es fundamental para el uso posterior y sostenible del recurso desarrollado. TEI, por su estructura modular, su flexibilidad y su soporte y documentación, es uno de los estándares más adecuados para representar documentos antiguos.

En este trabajo se ha presentado una propuesta concreta para la codificación de la información paleográfica y lingüística para corpus diacrónicos en formato XML utilizando las etiquetas recomendadas por este estándar. La representación externa, que almacena separadamente los textos de las anotaciones, es la más adecuada, ya que evita los conflictos de anidamientos de etiquetas, que surgen cuanto mayor es la cantidad de anotaciones, y permite incorporar otros niveles de anotación con relativa facilidad.

Aunque TEI ha sido y es utilizado para representar corpus sincrónicos y diacrónicos, esta es posiblemente la primera vez que se propone codificar no sólo la información lingüística, sino también la paleográfica en un corpus diacrónico siguiendo estas especificaciones. La representación de este tipo de información permitirá realizar estudios sobre la his-

toria de la lengua considerando factores que hasta ahora no se han tenido en cuenta en los análisis que utilizan datos obtenidos de corpus.

5. Anexo A. Elementos del estándar TEI para representar la información lingüística y paleográfica de textos diacrónicos

La siguiente tabla contiene los elementos de TEI propuestos en este trabajo para representar la información paleográfica y lingüística de textos diacrónicos, ordenados alfabéticamente. En la primera columna aparecen los nombres de los elementos, en la segunda el módulo de TEI al que pertenecen y en la tercera una breve descripción de su significado y de los atributos que pueden contener. La barra que aparece al final de algunos elementos indica que el elemento en cuestión está vacío e indica un punto concreto en el documento.

Etiquetas	Módulo	Descripción
abbr	Common Core	Indica una abreviatura.
add	Common Core	Marca letras, palabras o frases insertadas en el texto por el escriba o editor. Puede contener el atributo *@hand.*
additions	Manuscript Description	Señala una descripción de cualquier adición significativa encontrada en un manuscrito, como *marginalia* u otras anotaciones. Puede contener el atributo *@hand.*
catchwords	Manuscript Description	Describe el sistema utilizado para asegurar el orden correcto de los folios que forman un códice o un incunable, normalmente por medio de notas a pie de página.
cb/	Common Core	Marca el comienzo de una columna.
choice	Common Core	Agrupa codificaciones alternativas para una forma en el texto. Puede contener los elementos que marcan una abreviatura (*abbr*) y su expansión (*expan*), una supresión (*del*), un comentario del editor (*add*), la forma original (*orig*) o la estandardizada (*reg*).

del	Common Core	Contiene una letra, palabra o pasaje marcado como suprimido. Puede contener el atributo *@hand.*
expan	Common Core	Contiene la expansión de una abreviatura.
figure	Tables, Formulae, Figures	Marca fragmentos del texto que contienen información de un gráfico, una ilustración o una figura.
foreign	Common Core	Identifica una palabra o fragmento del texto en una lengua diferente.
G	Non-standard Characters and glyphs	Identifica un carácter o caracteres no escritos en el alfabeto estándar occidental.
gap/	Common Core	Indica una parte del documento donde se ha omitido material, ya sea por razones editoriales, ya sea porque el material es ilegible.
gloss	Common Core	Marca un texto utilizado para comentar o explicar otra parte del texto. Puede contener el atributo *@hand.*
graphic/	Common Core	Indica el lugar de un gráfico, ilustración o figura.
head	Common Core	Contiene cualquier tipo de cabecera incluida en el manuscrito original.
Hi	Common Core	Marca una palabra o una frase diferente gráficamente del texto que la rodea, como por ejemplo, las iniciales.
lb/	Common Core	Marca el comienzo de línea.
orig	Common Core	Contiene una forma del documento original.
P	Common Core	Contiene fragmentos del texto considerados párrafos.
pb/	Common Core	Marca el inicio de folio o página. Contiene el atributo @type para señalar si el folio es vuelto (valor "v") o recto (valor "r").
pc	Analysis and Interpretation	Marca un caracter o caracteres que constituyen un signo de puntuación. Contiene el atributo *@type*, para indicar el tipo de puntuación.
reg	Common Core	Contiene una forma que ha sido regularizada o normalizada. Puede contener el atributo *@hand.*
rubric	Manuscript Description	Marca el texto de una rúbrica o cabecera referida a un elemento particular del manuscrito.
S	Simple Analytic Mechanisms	Contiene una oración.
seg	Linking, Segmentation, and Alignment	Representa cualquier tipo de segmentación inferior a un sintagma.

supplied	Transcription of Primary Sources	Señala texto facilitado por el escriba o el editor, normalmente porque el original no se puede leer debido a daño físico o pérdida del original. Puede contener el atributo *@hand*.
text	Default Text Structure	Contiene cada uno de los textos que forman el corpus.
unclear	Common Core	Contiene una palabra, frase o fragmento de texto que no puede transcribirse con certeza porque es ilegible en el documento.
W	Analysis and Interpretation	Incluye una palabra. Contiene los atributos *@lemma* y *@type*, para indicar el lema y la categoría morfológica de la palabra, respectivamente.
xi:include	Linking, Segmentation, and Alignment	Incluye las referencias a los datos primarios en la anotación externa. Contiene los atributos *@href, @parse* y *@xpointer*.

6. Anexo B. Atributos del estándar TEI para representar la información lingüística y paleográfica de textos diacrónicos

La siguiente tabla contiene los atributos de TEI propuestos en este trabajo para representar la información paleográfica y lingüística de textos diacrónicos, ordenados alfabéticamente. En la primera columna aparecen los nombres de los atributos, en la segunda el módulo de TEI al que pertenecen y en la tercera una breve descripción de su significado y de los valores que pueden indicar.

@hand	Critical Apparatus	Indica la mano responsable del tipo de edición indicada por el elemento que lo contiene.
@href	Linking, Segmentation, and Alignment	Indica el nombre del archivo que contiene los datos referidos en el documento de anotación externa.
@lemma	Simple Analytic Mechanisms	Señala el lema de una forma.

@n	The TEI Infrastructure	Indica el número del elemento que lo contiene. El valor de este atributo no tiene que ser único en el documento. Utilizado en como atributo de los elementos que marcan el inicio de página, columna y línea (*pb, cb, lb*).
@parse	Linking, Segmentation, and Alignment	Especifica el tipo de texto al que se refiere el atributo *@href*. Sus valores pueden ser 'text' o 'xml', según sea el texto original.
@rend	The TEI Infrastructure	Marca el modo como el elemento que lo incluye está representado en el texto original. En esta propuesta, se utiliza como atributo de *hi* para indicar el tamaño de la inicial.
@subtype	The TEI Infrastructure	En el nivel de segmentación de palabras de la representación externa, indica el número de características tipográficas del texto como rúbricas, iniciales o fragmentos de texto en otras lenguas.
@target	The TEI Infrastructure	Especifica los identificadores de los elementos o pasajes a los que se refiere el elemento *ref*.
@type	The TEI Infrastructure	Atributo utilizado para clasificar o subclasificar elementos de una manera determinada.
@xml:id	The TEI Infrastructure	Atributo identificador en los documentos que forman la anotación externa utilizado para asignar un identificador único al elemento que lo contiene.
@xml:lang	The TEI Infrastructure	Indica la lengua del texto contenido utilizando una etiqueta de acuerdo con el BCP 47[16]. Se utiliza en esta propuesta como atributo de *foreign*. Puede tener uno de los siguientes valores: "la" (Latin), "ga" (gallego), "po" (portugués), "ca" (catalán), "ba" (vasco), "fr" (francés), "pr" (provenzal), "it" (italiano), "en" (inglés), "de" (alemán), "ar" (árabe), "he" (hebreo).
@xpointer	Linking, Segmentation, and Alignment	Describe el rango de unidades que ocupa el texto referido en *@href*.

16 <ftp://ftp.isi.edu/in-notes/bcp/bcp47.txt>.

Bibliografía

BANSKI, Piotr, y PRZEPIÓRKOWSKI, Adam (2009), «Stand-off TEI Annotation: the Case of the National Corpus of Polish», *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP, Suntec, Singapur, 64-67.

BIRD, Steven, y SIMONS, Gary (2003), «Seven Dimensions of Portability for Language Documentation and Description», *Language*, 79, 557-582.

BIBER, Douglas (1993), «Representativeness in corpus design», *Literary and Linguistic Computing*, 1993, 8(4), 243-257.

CORFIS, Ivy A., O'NEILL, John, y BEARDSLEY, Theodore S. Jr., eds. (1997), *Early Celestina Electronic Texts and Concordances*, Ltd. Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

DAVIES, Mark (2002), *Corpus del Español* (100 millones de palabras, 1200-1900), <http://www.corpusdelespanol.org>.

DIPPER, Stefanie, FAULSTICH, Lukas, LESER, Ulf, y LÜDELING, Anke (2004), «Challenges in Modelling a Richly Annotated Diachronic Corpus of German», *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, Lisboa, 21-29.

HAN, Chung-hye, y KROCH, Anthony (2000), «The rise of do-support in English: implications for clause structure», *Proceedings of the North East Linguistic Society (NELS 30)*, Washington D.C., Georgetown University.

HERRERA, María Teresa, y GONZÁLEZ DE FAUVE, María Estela, eds. (1997), *Textos y Concordancias Electrónicos del Corpus Médico Español*, Ltd. Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

IDE, Nancy, y ROMARY, Laurent (2004), «International Standard for a Linguistic Annotation Framework», *Natural Language Engineering*, 10, 3/4, 211–225.

IDE, Nancy, y SUDERMAN, Keith (2006), «Merging Layered Annotations», *Proceedings of Merging and Layering Linguistic Information, Workshop held in conjunction with LREC 2006*, Génova.

IDE, Nancy, y ROMARY, Laurent (2007), «Towards International Standards for Language Resources», L. Dybkjaer, H. Hemsen, W. Minker, eds., *Evaluation of Text and Speech Systems*, Springer, 263-84.

KASTEN, Llyod, John NITTI, y Wilhemina JONXIS-HENKEMENS, eds. (1997), *The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio*, Ltd. Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

LEHMBERG, Timm y Kai WÖRNER (2008), «Annotation Standards», Anke Lüdeling and Merja Kytö, eds., *Corpus Linguistics. An International Handbook*, Berlin, Mouton de Gruyter.

LAY, Marie-Hélène, y DEMONET, Marie-Luce (2008), «Sustainability and Sharability of the Humanist Virtual Library (BVH): Experiment Feed-back», Andreas Witt, Georg Rehm, Thomas Schmidt, Khalid Choukri y Lou Burnard, eds., *Proceedings of the LREC 2008 Workshop "Sustainability of Language Resources and Tools for Natural Language Processing"*.

MACKENZIE, David, y BURRUS, Victoria A. (1986), *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language*, 4ª ed., Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

NITTI, John, y KASTEN, Lloyd, eds. (1997), *The Electronic Texts and Concordances of Medieval Navarro-Aragonese Manuscripts*, Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

O'NEILL, John (1999), *Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings*, Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

REAL ACADEMIA ESPAÑOLA: *Banco de datos (CORDE) [en línea]. Corpus diacrónico del español,* <http://www.rae.es> (10/01/10)

RISSANEN, Matti (1994), «The Helsinki Corpus of English Texts», Merja Kytö, Matti Rissanen y Susan Wright, eds., *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*, Amsterdam, Rodopi.

SAGI, Eyal, KAUFMANN, Stefan, y CLARK, Brady (2009), «Semantic density analysis: Comparing word meaning across time and phonetic space», Roberto Basili and Marco Pennacchiotti, eds., *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, Athens.

SÁNCHEZ GONZÁLEZ DE HERRERO, María Nieves, HERRERA, María Teresa, y ZABÍA, María Purificación, eds. (2003), *Textos medievales misceláneos*, Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

SÁNCHEZ-MARCO, Cristina, BOLEDA, Gemma, FONTANA, Josep Maria, y DOMINGO, Judith (2010), «Annotation and Representation of a Diachronic Corpus of Spanish», *Proceedings of Language Resources and Evaluation Conference (LREC).*

SÁNCHEZ-MARCO, Cristina, BOLEDA, Gemma, PADRÒ, L. (2011), «Extending the tool, or how to annotate historical language varieties», *Proceedings of 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.*

TEI CONSORTIUM, eds., *TEI P5: Guidelines for Electronic Text Encoding and Interchange,* TEI Consortium, <http://www.tei-c.org/Guidelines/P5/> (12/01/10).

WALTMAN, Franklin, ed. (1999), *Textos y concordancias del Fuero general de Navarra*, Madison, Wisconsin, Hispanic Seminary of Medieval Studies.

ZINSMEISTER, Heike, HINRICHS, Erhard, KÜBLER, Sandra, y WITT, Andreas (2008), «Linguistically Annotated Corpora: Quality Assurance, Reusablity and Sustainability», Anke Lüdeling y Merja Kytö, eds., *Corpus Linguistics. An International Handbook,* Berlin, Mouton de Gruyter.