Leyre Martín Aizpuru & M[a] Nieves Sánchez González de Herrero

El estudio de la documentación alfonsí: un proyecto abierto

The study of documentation from the reign of Alfonso X: an open project

Resumen: La importancia que en la historia del castellano suele atribuirse a la figura del rey Alfonso X el Sabio hizo que nos interesáramos por los testimonios originales de la cancillería. La recogida, y el estudio correspondiente, de la documentación notarial alfonsí nos han ocupado varios años y no damos aún por concluido el trabajo; durante este tiempo nos hemos ido adaptando a los avances informáticos y asumiendo las posibilidades que nos ofrecen las Humanidades Digitales. Con este trabajo pretendemos mostrar cuál ha sido el recorrido con los objetivos alcanzados, en qué punto nos encontramos hoy y cuáles son las perspectivas futuras.

Palabras clave: Historia del español; documentación alfonsí.

Abstract: The importance of King Alfonso X *El Sabio* (the Wise) in the history of the Castilian language underpins our interest in original evidence from the Royal Chancery. The collection and analysis of data from Alfonso X's notarial texts has taken several years, and work is ongoing. During this time, we have adapted our activity to new IT developments and taken advantage of opportunities afforded by the Digital Humanities. In this work, we will show the progress and results of the project to date, describing the current state of our work and the future of our research here.

Keywords: History of Spanish; Alfonso X Documentation.

1 Introducción

El *Grupo de Estudio de Documentos Históricos y Textos Antiguos de la Universidad de Salamanca* (GEDHYTAS) dedica su labor investigadora al conocimiento de

Leyre Martín Aizpuru and M[a] Nieves Sánchez González de Herrero, Universidad de Salamanca – IEMYRhd

https://doi.org/10.1515/9783110585421-009

la historia de la variación lingüística peninsular y a la edición de textos. Está formado por investigadores licenciados y doctores en Filología Románica o Hispánica, además de colaboradores de otras áreas lingüísticas como la de Estudios Árabes e Islámicos.

GEDHYTAS está integrado en la Red Internacional *Corpus Hispánico y Americano en la Red: Textos Antiguos* (*Charta*),[1] proyecto global de edición de textos hispánicos, que reúne a diversos grupos de universidades nacionales e internacionales; también forma parte del Instituto de Estudios Medievales y Renacentistas de la Universidad de Salamanca (IEMYRhd).

Desde hace unos años, tres son los corpus de documentos notariales de la época medieval en que se basa la labor de edición y estudio documental del grupo: la cancillería real castellana del siglo XIII, el concejo de Miranda de Ebro (Burgos) y el de Mombeltrán (sur de Ávila).

A continuación, nos centraremos en el desarrollo del proyecto dedicado al primero de los corpus, el alfonsí, que ha pasado por tres grandes fases: Hispanic Seminary (1995–2002), red *Charta* (2008–actualidad) y, por último, una tercera en la que exploramos el aprovechamiento de las herramientas digitales para la creación de una edición digital en XML-TEI de dichos documentos y una extracción de datos con información estadística por medio del sistema *Lyneal*,[2] ideado por el profesor Hiroto Ueda (Universidad de Tokio).

2 El corpus alfonsí en el Hispanic Seminary of Medieval Studies

La idea del proyecto partió de una propuesta que a finales del ya lejano 1995 nos transmitió John Nitti, entonces profesor de la Universidad de Wisconsin (Madison) y responsable del Hispanic Seminary of Medieval Studies (HSMS), ubicado en la citada universidad en aquel momento. Su idea era que la recogida y el estudio de la documentación de cancillería complementaría la investigación que el HSMS llevaba a cabo sobre los textos emanados de la cámara regia que sirvieron de base para la elaboración del *Diccionario de la prosa castellana del rey Alfonso X*, finalmente publicado en tres volúmenes (Kasten/Nitti 2002). Se planteaba, pues, como un trabajo complementario cuyo objetivo primero era la constitución de un corpus textual de tipo histórico que serviría para elaborar un diccionario parcial

1 https://www.redcharta.es/ (marzo de 2018).
2 http://shimoda.lllf.uam.es/ueda/lyneal/ (última consulta: marzo de 2018).

que acabaría sumándose a otros que en su conjunto constituirían el *Dictionary of Old Spanish Language* (*DOSL*). Es importante subrayar que desde el inicio se concibió como un proyecto exclusivamente filológico.

El planteamiento, asumido por el grupo de la Universidad de Salamanca que colaboró en el proyecto, formado en ese momento por las profesoras M. Nieves Sánchez González de Herrero y M. Teresa Herrera, exigía adoptar las normas de transcripción, básicamente paleográficas, y servirse de los programas informáticos del HSMS. Recordemos que en aquellas fechas no disponíamos de internet ni de plataformas de consulta, de modo que los distintos equipos trabajaban de manera autónoma utilizando las mismas herramientas informáticas.

La producción de la cancillería en los años del rey Sabio fue inmensa y los testimonios se conservan en un gran número de archivos; este hecho, unido al objetivo planteado, hizo que buscáramos un corpus representativo, sin pretender la exhaustividad. Decimos representativo ya que es muy difícil crear un corpus íntegro de este tipo de documentación cancilleresca, pues los lugares en que descansan documentos reales del Medievo son casi inabarcables: además de los archivos municipales, provinciales y catedralicios de buena parte de la Península Ibérica, hay multitud de iglesias, monasterios o conventos que conservan su documentación medieval, no siempre catalogada o accesible. Finalmente quedó formado por 660 documentos originales que, con criterios filológicos que justificamos (Sánchez González de Herrero 2002: 166–176), agrupamos en función de las regiones de destino: 213 dirigidos a Andalucía, 24 a Galicia, 115 al reino de León, 72 a Murcia, 172 a Castilla la Vieja y 64 a la Nueva.

Resumidos, los resultados del trabajo de cinco años fueron la publicación del *Diccionario Español de Documentos Alfonsíes* (*DEDA*), los *Textos y concordancias electrónicos de documentos castellanos de Alfonso X* y varios estudios filológicos en los que tratamos de describir la norma lingüística de la cancillería en el siglo XIII, insistiendo en la variación gráfico-fonética, morfosintáctica y léxica.

3 *Codcar* (Corpus de Documentos de Cancillería Real, siglo XIII y primera década del XIV) en *Charta* (Corpus Hispánico y Americano en la Red: Textos Antiguos)

A partir de 2008, nos integramos en la red *Charta* con el propósito de consolidar y ampliar la colección documental de cancillería del siglo XIII; tras los estudios lingüísticos sobre la producción alfonsí, consideramos que los testimonios

debían analizarse en un marco más amplio que tuviera en cuenta los precedentes de Fernando III y la continuación con Sancho IV y Fernando IV.

La red *Charta*, en formación en esas fechas, se planteaba, en palabras de su fundador, como

> un proyecto global para la edición de textos archivísticos hispánicos con la intención de integrar una sólida fundamentación filológica y los desarrollos informáticos necesarios para proporcionar así a los investigadores e interesados por la lengua, la historia y la cultura una edición digital fiable, comprobable y que se pueda citar directamente en estudios de diferente ámbito y nivel, especialmente en el científico (Sánchez-Prieto 2012: 32).

Actualmente, la Red está formada por veintisiete grupos de investigación nacionales e internacionales y ha celebrado cinco ediciones de congresos internacionales dedicados a la edición y análisis del documento antiguo.

Entre sus características destacamos la triple presentación de los textos, reproducción facsimilar, transcripción paleográfica y presentación crítica, así como la sólida base de unos criterios filológicos rigurosos, que han sido ampliamente discutidos en un largo proceso de elaboración y revisión tras su puesta en práctica.

Los criterios concretos de transcripción paleográfica y presentación crítica –las cuestiones relativas al desarrollo de abreviaturas, puntuación, secuenciación de palabras, acentuación y otros aspectos– son fruto, en primera instancia, del acuerdo surgido en el encuentro *Hacia un estándar en la edición de textos antiguos españoles*, celebrado en el Centro Internacional de la Lengua Española (Cilengua), en San Millán de la Cogolla (La Rioja) el año 2007. Bajo la coordinación de Pedro Sánchez-Prieto, y desde 2017 también de Belén Almeida, los diversos criterios fueron ideados por un grupo de filólogos y aceptados por los directores de los tres institutos del Cilengua. Finalmente, tras varias reuniones de los grupos de investigación que participan en *Charta*, se publicó una edición actualizada y revisada de los criterios (Sánchez-Prieto 2011), que también se puede consultar en la página web de la Red (en versión actualizada en abril de 2013).[3]

GEDHYTAS, como grupo de investigación integrado en la red *Charta*, se acoge a estos criterios de edición y todos sus miembros realizan la transcripción y edición de sus documentos dentro de este marco. El nuevo corpus cancilleresco, con ampliación de la cronología y cantidad, nos obligó a releer y transcribir de nuevo los documentos alfonsíes transcritos para la colección del HSMS y que adaptamos a los criterios de *Charta*.

3 https://www.redcharta.es/criterios-de-edicion/ (Última consulta: marzo de 2018).

La primera tarea, que iniciamos a mediados del año 2010, consistió en la realización de un vaciado de los textos repartidos por los diferentes archivos, para lo cual partimos de obras de historiadores y especialistas en las figuras de los reyes estudiados (Torres Fontes 1977; González González 1980–1986; González Jiménez 1991; Herrera *et al*. 1999; González Jiménez/Carmona Ruiz 2012). De las citadas obras tomamos la referencia de los documentos originales,[4] así como de los archivos en los que descansan en la actualidad. Una vez dibujado el mapa, acudimos a las diferentes colecciones diplomáticas y catálogos de los propios archivos a fin de localizar el mayor número de testimonios (Chacón Gómez-Monedero *et al*. 2008; Ciérbide/Ramos 2000; De Lera Maíllo 1999; García Luján 1982; Martín Fuertes 1998; por citar solo algunos).

Confeccionada la lista, contactamos con los archivos para conocer las condiciones de acceso y consulta de los fondos y acudimos personalmente a ellos o solicitamos una reproducción digital. Hay una circunstancia que va muy unida a la creciente accesibilidad de los archivos, la llamada "revolución digital" de las últimas décadas. La comodidad de acceso a los documentos que ofrecen actualmente muchos archivos viene dada, en parte, por los monumentales avances realizados en el campo audiovisual, que permiten la digitalización del patrimonio documental a una calidad y rapidez antes desconocidas.

Además, el hecho de contar con reproducciones digitales de los documentos no solo facilita la tarea de transcribir, sino que se convierte en una parte importante de nuestro sistema de edición, cuyos criterios aconsejan que se proporcione al lector el texto transcrito acompañado de una copia o reproducción facsimilar del propio original. En resumen, parece que van quedando atrás técnicas de reproducción poco precisas y laboriosas a la hora de transcribir y editar los textos, como los microfilms o las fotocopias sacadas de estos, en las que se pierde mucha calidad de lectura, para dar paso a un nuevo sistema, más fiable y cómodo a la hora de avanzar con las investigaciones filológicas.

Entre los años 2010 y 2013, los miembros de GEDHYTAS llevamos a cabo una labor de fotografiado y petición de fotografías en los siguientes organismos:

Archivos catedralicios u otras instituciones eclesiásticas:

- Archivo de la Catedral de Cuenca, León, Orense, Salamanca, Santander, Santiago de Compostela, Segovia, Toledo, Tuy y Zamora
- Archivo de la Colegiata de San Isidoro de León y de Covarrubias
- Archivo Diocesano de Salamanca e Histórico Diocesano de Logroño

4 Para la consideración de un documento como original seguimos las indicaciones que los archiveros e historiadores facilitan en las propias colecciones documentales.

- Archivo del Monasterio de Carrizo, de Santa María de Gradefes y de San Clemente
- Institución Colombina de Sevilla

o Archivos de titularidad pública:

- Archivos Históricos Provinciales de Ávila, Burgos, Palencia y Soria
- Archivos Municipales de Alcalá de Henares, Ávila, Baeza, Béjar, Bergara, Burgos, Ciudad Rodrigo, Cuéllar, Córdoba, Cuenca, Écija, El Puerto de Santa María, Huelva, La Coruña, Ledesma, León, Linares, Logroño, Miranda de Ebro, Mondragón, Murcia, Oviedo, Pamplona, Salvatierra, Segovia, Sepúlveda, Sevilla, Talavera de la Reina, Toledo, Valladolid y Vitoria
- Arquivo Nacional da Torre do Tombo (Lisboa)
- Archivo General de La Rioja y de Navarra
- Archivo Histórico Nacional
- Archivo del Territorio Histórico de Álava
- En total, aunque hubo archivos importantes a los que no pudimos acceder y otros que descartamos por falta de tiempo, en 2013 el grupo GEDHYTAS había recopilado un total de 800 documentos y dio por finalizada esta fase. Las cifras pueden parecer escasas si se comparan con el catálogo de documentos alfonsíes que publicaron Manuel González Jiménez y Mª Antonia Carmona (2012), pero debemos considerar que entre los 3397 testimonios que estos autores recogen, hay tanto documentos originales como copias o traslados, que dejamos fuera por razones filológicas.

En estos momentos, *Codcar* ofrece en línea[5] 756 documentos originales de cancillería real de Fernando III a Fernando IV, con el reparto siguiente: 45 de Fernando III, 386 de Alfonso X, 224 de Sancho IV y 101 de Fernando IV. En cuanto a la cronología, contamos con la siguiente distribución: 699 documentos del siglo XIII y 57 del XIV, es decir, de 1223 a 1312. Aquí debemos hacer dos observaciones: en primer lugar, los testimonios de los primeros años son escasos, porque eran pocos todavía los que se escribían en castellano. En segundo lugar, el objetivo se centra en el XIII, pero, teniendo en cuenta que los reinados de los dos continuadores del rey Sabio fueron breves, eliminar los primeros años del XIV implicaba una escasa representación de la cancillería de Fernando IV, por lo que dimos preferencia a dicha continuidad al menos durante 12 años. A los ya disponibles, se sumarán próximamente 43: 2 de Fernando III, 17 de Alfonso X, 17 de Sancho IV y 7 de Fernando IV.

5 http://www.corpuscharta.es/consultas.html (última consulta: marzo de 2018).

4 *Codcar* y las Humanidades Digitales

En estos años de trabajo con el ampliado corpus de cancillería castellana, no solo nos hemos valido de los avances tecnológicos en el campo de la digitalización y reproducción de los testimonios, sino que también hemos sido sensibles a las llamadas *Humanidades Digitales*, que, como señala Torruella (2017: 15), "ha(n) cambiado totalmente la manera de acceder a los textos, de trabajar con ellos y de presentar y poner a disposición de la comunidad científica los resultados obtenidos". Nuestra finalidad ha sido doble: por un lado, hemos explorado las posibilidades de la edición digital, con el lenguaje de marcación XML-TEI, y, por otro, el del manejo de herramientas digitales aplicadas a la extracción de datos filológicos.

4.1 *Codcar* en TEI

Entre 2012 y 2014, varias integrantes del grupo colaboraron en la elaboración de una guía de edición de textos mediante el lenguaje estándar TEI (*Text Encoding Initiative*);[6] este lenguaje es uno de los modelos de marcación de textos más empleado en las Humanidades Digitales, marcando con XML-TEI las transcripciones de los documentos. El proyecto o *Guía para editar textos CHARTA según el estándar TEI* (Isasi *et al.* 2014) surge como una propuesta "que aspira a ser un modelo de etiquetas TEI que cubran todos los casos de transcripción y edición de los *Criterios* CHARTA" (Martín Aizpuru 2016: 145), lo que incluye desde las abreviaturas o intervenciones de los escribanos en el texto –en forma de tachados, cancelados, etc.– hasta la regularización lingüística que supone el paso de la transcripción paleográfica (en adelante TP) a la presentación crítica (en adelante PC).

En definitiva, los principales objetivos de este proyecto son "representar los criterios de edición de la Red CHARTA mediante un lenguaje de marcación conforme a la propuesta TEI" y proponer "para todos los casos considerados en los criterios de CHARTA, etiquetas TEI que permitan visualizar tanto la transcripción paleográfica como la versión crítica" (Isasi *et al.* 2014: 12) y que pueda servir de modelo para futuros editores de textos hispánicos que quieran emplear este método.[7]

6 Página web del Consorcio TEI: http://www.tei-c.org/ (última consulta: marzo de 2018).

7 Podemos citar, por ejemplo, el trabajo de investigación, aún en curso, iniciado por Ricardo Pichel en una estancia en el King's College London (2016), bajo la orientación de Paul Spence, consistente en el desarrollo de un protomarcado (para Office Word y Oxygen) en fuente única

En este trabajo, se ha seguido, básicamente, la estructura de un documento XML: se presenta, primero, la cabecera o información metatextual –datos referentes al archivo electrónico, a la marcación digital, al documento fuente u original y al historial de revisiones del archivo electrónico–, que se corresponde a la información incluida bajo la etiqueta <teiHeader>, y, en segundo lugar, todos los casos de TP y PC para cada elemento de los criterios de *Charta* –que en un XML irían bajo la marca <text>–.

En cada uno de ellos, se reproduce la misma estructura de inclusión de información. En primer lugar, en el epígrafe "Descripción", se hace un breve resumen del caso, junto con una referencia a los criterios de *Charta*. A continuación, se presenta el texto según versión de la TP o PC –o ambas si se considera pertinente–, junto con un fragmento de la imagen del testimonio. En "Marcación TEI" se ofrece una descripción de las opciones de marcación, del valor de las etiquetas seleccionadas, así como las restricciones de uso de las mismas. Además, se añaden uno o varios ejemplos etiquetados. Por último, hay dos epígrafes opcionales: "Sugerencias de visualización", donde se describen visualizaciones que van más allá de una simple reproducción de la TP y PC de *Charta* y "Comentarios", que recoge algún tipo de indicación adicional o complementaria, si así se requiere.

En el proceso de elaboración de dicha *Guía* se ensayaron las etiquetas TEI con documentos de *Codcar* y, por ello, numerosos fragmentos de textos de este corpus ejemplifican los casos de edición. A continuación, mostramos algunas de las aplicaciones filológicas, así como las principales reflexiones surgidas a partir de este proyecto de edición digital.

En cuanto a las posibilidades de hacer explícita la estructura principal del documento, los testimonios de *Codcar* requieren de indicaciones como *salto de columna*, *inclusión de signos especiales*, *anotaciones al margen*, etc., que permiten estructurar el texto transcrito. Veamos, a continuación, unos ejemplos transcritos y editados, según los criterios de *Charta*, y marcados según sus equivalentes en el lenguaje de marcación TEI:

para *Charta* preparado para su conversión automática en TEI. Por otra parte, en el marco del "Congreso Internacional de la Red de Estudios Medievales Interdisciplinares Humanidades Digitales: miradas hacia la Edad Media", celebrado del 9 al 11 de octubre de 2017 en la Facultad de Filología de la Universidad de Santiago de Compostela, Idalete Dias, Sílvia Araújo y Pedro Dono presentaron una comunicación titulada "Edición dixital de textos medievais: unha proposta editorial á volta dos documentos do notario Johán Ares (1272–1300)" en la que partían del sistema de marcación de la guía para etiquetar los textos de su corpus.

Salto de columna

Figura 8.1: Salto de columna. *Codcar*-0540 (Isasi *et al.* 2014: 44–45).

PC según los criterios de *Charta*: {a} {33} Don Joán Alfonso, obispo de Palencia {34} chanceller del rey confirma. […] {b} {33} Don Joán, fi del ifante don Manuel, confirma. {34} Don Lope confirma. […]

Etiquetado con TEI: <cb n="a"/><lb n="33"/>Don Joán Alfonso, obispo de Palencia <lb n="34"/> chanceller del rey confirma. […] <cb n="b"/><lb n="33"/ >Don Joán, fi del ifante don Manuel, confirma. <lb n="34"/>Don Lope confirma. […]

Rueda

Figura 8.2: Rueda. *Codcar*-0115 (Isasi *et al.* 2014: 74–78).

TP: [*rueda*: SIGNO DEL REY DON ALFONSO/EL ALFEREZIA DEL REY UAGA/DON IUAN GARCIA MAYORDOMO DE LA CORTE DEL REY LA CONFIRMA]
PC: [*rueda*: Signo del rey don Alfonso. El alferezía del rey vaga. Don Juan García, mayordomo de la corte del rey, la confirma.]

```
<seg type="rueda">
   <seg type="an-int">SIGNO DEL REY DON ALFONSO</seg>
   <seg type="an-ext"/>
   <seg type="an-ext-reloj">EL ALFEREZIA DEL REY UAGA</seg>
   <seg type="an-ext-contrarreloj">DON IUAN GARCIA MAYORDOMO DE LA
   CORTE DEL REY LA CONFIRMA</seg>
</seg>
```

Por otra parte, además de las características diplomáticas y formales del documento, de las que hemos visto dos ejemplos (Figura 8.1 y Figura 8.2), la labor editorial contempla criterios para representar los usos paleográficos (abreviaturas, deterioros y restitución de elementos en el texto, intervenciones en el texto –tales como el tachado, cancelado, raspado, margen, cambio de mano, etc.–) y para presentarlo de forma crítica (regularización de diferentes grafías vocálicas y consonánticas, unión y separación de palabras, reparto de mayúsculas y minúsculas, acentuación, puntuación, etc.). A continuación, incluimos algunos ejemplos de cuáles han sido las decisiones tomadas en la *Guía Charta-TEI*.

Desarrollo de abreviaturas

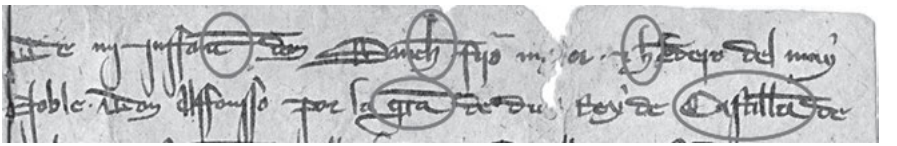

Figura 8.3: Abreviaturas. *Codcar*-0443 (Isasi *et al.* 2014: 48–50).

TP según los criterios de *Charta*: De mj jnffant<e> don Sanch<o> ffijo
Etiquetado con TEI: De mj <expan>jnffant<ex>e</ex></expan> don <expan>Sanch<ex>o</ex></expan> ffijo

Como se ve en la Figura 8.3, <expan> recoge la forma afectada por el desarrollo de una o varias abreviaturas y dentro se inserta <ex> para incluir las letras abreviadas. A partir de este etiquetado, se podría generar un índice de formas desarrolladas y otro de formas abreviadas.

Normalización para la presentación crítica

En este apartado se representa el paso de la TP a la PC, con la homogeneización de diferentes aspectos gráficos del documento que faciliten en última instancia la comprensión del texto por parte del lector. Todos los casos de normalización se marcan en TEI mediante el mismo sistema de etiquetado (<choice>, <orig> y <reg>), tal como se ve en la Tabla 8.1. Este tipo de marcación, con las dos etiquetas <orig> y <reg> combinadas sobre una misma forma y englobadas por la marca <choice>, es una de las claves de la fuente única, aspecto que trataremos a continuación.

Tabla 8.1: Normalización para la presentación crítica: distintos casos (Isasi *et al.* 2014: 78–82).

CASO	CHARTA TP	CHARTA PC	CHARTA TEI
Regularización de grafías	Seuilia	Sevilla	<choice> <orig>Seuilia</orig> <reg>Sevilla</reg> </choice>
	thenor	tenor	<choice> <orig>thenor</orig> <reg>tenor</reg> </choice>
Unión y separación de palabras	enestos	en estos	en<reg> </reg>estos
	buena miente	buenamiente	buena<orig> </orig>miente
	sobrellos	sobr'ellos	sobr<reg>'</reg>ellos*
Mayúsculas y minúsculas	dios	Dios	*de minúscula a mayúscula:* <reg type="mM">d</reg>ios
	Rey	rey	*de mayúscula a minúscula*: <reg type="Mm">R</reg>ey
Acentuación	seran	serán	<choice> <orig>seran</orig> <reg>serán</reg> </choice>
	estámos	estamos	est<choice> <orig>á</orig> <reg>a</reg> </choice>mos
Puntuación	–	–	[Tiene diferentes posibilidades de marcación que se explican más abajo]

(Continúa)

Tabla 8.1: (Continúa)

CASO	CHARTA TP	CHARTA PC	CHARTA TEI
Números	U	mil	<choice> <orig>U</orig> <reg>mil</reg> </choice>
	.xiij.	XIII	<choice> <orig>.xiij.</orig> <reg>XIII</reg> </choice>

***Nota:** En este caso no es necesario <choice> y <orig> ya que simplemente hay que añadir el apóstrofo (') o punto medio (·).

Como se ha visto en los ejemplos precedentes, la mayoría de los casos etiquetados incluidos en la descripción de la Guía son ejemplos de etiquetado sencillo (además, en cada uno solo se muestra la marcación del fenómeno del que se está hablando en cada momento). Aun así, en el proceso de marcación de los documentos del corpus los investigadores observaron y trabajaron sobre varias complejidades que, a continuación, presentamos.

Los dos problemas más comunes a la hora de enfrentarnos al etiquetado TEI son la complejidad del propio etiquetado y los casos de etiquetado solapante, es decir, con diferentes marcas anidadas unas dentro de otras, que pueden entorpecer tanto la labor del editor digital como la de sus lectores. A eso se suma la necesidad metodológica de preparar una edición digital para las dos versiones que se incluyen en la publicación *Charta* ya que hace plantear si el proceso de marcación ha de realizarse en un solo texto (*fuente única*) o en dos (*fuente doble*). Esto es:

> A un nivel práctico, el modelo TEI-XML ofrece nuevas opciones –se puede editar la versión paleográfica primero, dejando la edición crítica para después, se pueden crear las dos versiones a la vez, o emplear un método intermediario creando la versión paleográfica primero, pero con anotaciones que vienen a formar parte de la edición crítica (Spence *et al*. 2012: 480).

El primer punto, la extensión de las etiquetas y el nivel de solapamiento de las mismas, provoca que el texto editado en fuente única quede muy oculto entre las etiquetas, como puede comprobarse en el siguiente ejemplo, donde se marca (1) que en el manuscrito hay un deterioro que impide leer bien el texto (*<orig><subst>~~<gap/>~~*), (2) la reconstrucción de la letra oculta (*<add></add>*) y (3) la regularización de grafías (<reg></reg>), todo ello enmarcado dentro de <choice></choice>:

```
<choice>
<orig>
<subst>
<del>Palenç<gap unit="chars" quantity="1"/> a</del>
<add>i</add>
</subst>
</orig>
<reg>Palencia</reg>
</choice>
```

En resumen, esta sucesión de etiquetas indica que en el manuscrito –y por tanto en la TP– se lee *Palenç*a*, mientras que en la PC se reconstruye y regulariza *Palenc<i>a*, tal como podemos observar en la Figura 8.4.

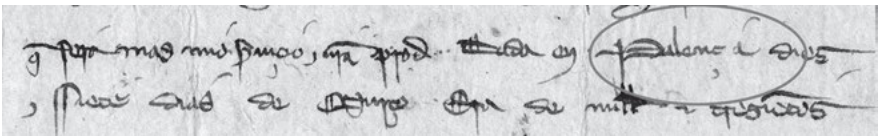

Figura 8.4: *Codcar*-0508-L12-13 (ejemplo donde se lee *Palenç*a*).

Por último, la segunda cuestión planteada interroga al investigador sobre si el proceso de marcación debe realizarse en un solo texto (*fuente* única) o en dos (*fuente doble*). En las pruebas que hemos realizado hemos optado por incluir las etiquetas en un solo texto, fuente única, porque a pesar de ser más "costoso" las ventajas de visualización, creación de índices, etc. en fuente única son mayores que si se trabajara con dos textos marcados. Este será un tema para debatir y reflexionar en el futuro ya que, a pesar de las ventajas de visualización, hay que valorar y probar cómo se desarrolla el trabajo de edición y cómo es recibido por la comunidad científica.

En las páginas previas hemos mostrado tanto desde la teoría como desde la práctica cómo TEI puede llegar a ser un lenguaje de marcación útil para los editores de documentación histórica en español. Con este lenguaje se llega a obtener resultados editoriales mucho más ricos y dinámicos tanto en su explotación científica como en su presentación visual, ya que permite confeccionar índices, realizar búsquedas de todos los elementos y partes de los documentos marcados, etc.

Si bien estos aspectos son muy positivos, no podemos obviar las complicaciones que de su empleo se desprenden –algunas de ellas las hemos revisado en las páginas anteriores–. Estas dificultades solo pueden ser subsanadas por medio del desarrollo de la investigación, dirigida hacia el análisis de la forma en que se

ve afectado el proceso de edición con las nuevas herramientas; la manera de conseguir estas mejoras pasa por la etiquetación de corpus más extensos y de tipologías textuales más diversas. Es en esa dirección en la que se están moviendo algunos investigadores implicados en el proyecto CHARTA-TEI. En definitiva, tal como se ha desarrollado en este artículo, el campo de la edición con TEI presenta muchas oportunidades para los editores de textos antiguos y los futuros investigadores, que podrán aprovechar los corpus así marcados para agilizar sus investigaciones lingüísticas.

4.2 *Codcar* en *Lyneal* (*Letras y Números en Análisis Lingüísticos*)

Actualmente se está haciendo una tesis doctoral cuyo principal objetivo es contribuir al concepto de norma cancilleresca medieval, especialmente la llamada "norma alfonsí", con un análisis previo de si tal concepto procede. Para ello, de la totalidad del corpus editado se ha seleccionado una muestra suficientemente representativa, tanto en el cómputo general como en el reparto por reyes y escribanos, sobre la que llevar a cabo los análisis en los siguientes planos lingüísticos: grafemático, fonético y morfosintáctico. En esta investigación son fundamentales las características paleográficas y de tipología documental, ya que se distinguirá entre los ejemplos extraídos de las partes diplomáticas –protocolo, escatocolo– y entre aquellas muestras halladas en el interior del cuerpo del documento –notificación, exposición, disposición–. Por otro lado, se tendrá también en cuenta quién es el escribano y notario de la carta, así como la procedencia y destino de la misma. De esta manera se establecerá si las reglas lingüísticas eran diferentes en cada reinado o si había unas características predominantes a lo largo de todo el mencionado periodo. Dicho de otro modo, el análisis y la consideración de las variables extralingüísticas permitirán conocer el alcance de los diferentes factores sociolingüísticos en la escritura de los funcionarios reales.

Para realizar este estudio, nos valemos de *Lyneal*, un sistema de análisis de textos en español ideado por el profesor Hiroto Ueda que tiene como primer objetivo "facilitar procesamientos de datos textuales tanto de los archivos almacenados en el servidor como los propios del usuario" (Ueda 2018). El autor ha

> buscado la combinación ideal entre la parte lingüística con la estadística en la plataforma informática común, que funciona de manera lineal desde los datos lingüísticos, pasando por la búsqueda general, para llegar a las tablas de distribución y la visualización gráfica (Ueda [en prensa])

El sistema se divide en dos partes, Letras y Números, y de la combinación de los datos de ambos se extraen los resultados estadísticos. El método sigue la siguiente dirección: Datos > Letras > Números > Gráficos.

Los rasgos característicos del sistema *Lyneal* son (Ueda [en prensa]):

- facilitar el modo de buscar las formas utilizando la expresión regular simplificada
- posibilitar el reemplazo de textos a la hora de búsqueda, de modo que las etiquetas pueden ser eliminadas
- filtrar los atributos necesitados, por ejemplo, lugar, papel, sexo, edad, nivel de educación, relación entre informante y encuestador (conocido y desconocido)
- combinar los múltiples atributos para analizar de manera separada y/o unida
- seleccionar los componentes deseados
- elaborar distintos tipos de frecuencia (absoluta, relativa, normalizada, probabilística, tipificada, etc.)
- calcular distintos valores estadísticos básicos (media, varianza, desviación típica, mediana, cuartiles, rango, etc.)
- posibilitar análisis multivariantes
- ofrecer distintos tipos de gráficos

El manejo del sistema es fácil e intuitivo lo cual facilita la realización de búsquedas y extracción de datos estadísticos. Además, el autor ha preparado una Guía (Ueda 2018) en la que explica el funcionamiento, de forma pormenorizada, y la manera en que hay que introducir los patrones de búsqueda, con expresiones regulares.

Con relación a la metodología para construir *Lyneal*, Ueda ha seguido un modelo de poscategorización. Es decir, en *Lyneal* se pueden visualizar y comprobar uno a uno los ejemplos que devuelve la búsqueda para así interpretar y valorar los resultados. De esta manera, el investigador puede asegurarse de realizar, primero, búsquedas generales y, después, búsquedas más concretas, para no obviar ningún ejemplo que de primeras podría pasar desapercibido. Este es, precisamente, uno de los valores del sistema *Lyneal* pues en otras plataformas de corpus digitales el investigador no tiene acceso a la información general ya que solo puede pedir a la máquina búsquedas de ejemplos concretos. En este sentido, consideramos que el sistema ideado por Ueda respeta una de las condiciones que plantea Kabatek (2017: 13) para los trabajos de lingüística de corpus: "la 'objetividad' científica no reside solo en el tratamiento numérico adecuado, sino también en el paso previo: el rigor metodológico necesario para la transformación de textos en datos numéricos". En esta línea, en una publicación anterior, Kabatek defiende la idea de que la informática, con las nuevas técnicas de extracción de datos, no solo no ha facilitado la labor del filólogo sino todo lo contrario ya que

> [L]os problemas tradicionales de reconstrucción siguen siendo los mismos y el acceso a más datos ha causado nuevos desafíos. Las cuestiones de la frecuencia, de la estadística y de la ponderación de datos se han planteado de forma nueva y, al mismo tiempo, nuevos factores se han añadido a la lista larga de posibles condicionantes del cambio lingüístico: la teoría del cambio lingüístico ha ido identificando, en las últimas décadas, un número creciente de factores sintácticos, semánticos, fónicos y pragmáticos que pueden condicionar los cambios y, dependiendo del fenómeno estudiado, la lista puede ser larga (2016: 9–10).

En definitiva, estas nuevas posibilidades no deben desviarnos de nuestro objetivo principal, el filológico, por lo que, en la línea de lo que destacan Kabatek (2017: 11) y Torruella (2017), el "rigor científico" debe partir de un buen conocimiento filológico de los textos que analizamos.

En los últimos años, *Lyneal* se ha empleado en varios trabajos de investigación, con corpus de documentos históricos de diversa tipología, cronología y geografía (Ueda 2015; Ueda/Moreno 2015; Torrens/Ueda 2016; Martín Aizpuru/Ueda [en prensa]) y son ya quince los corpus, de diversa naturaleza (tipología, cronología y geografía) que se pueden analizar por medio de este sistema.

En el caso de *Codcar*, hemos integrado los documentos en *Lyneal* –en formato .txt y con separadores de tabulación– ya que por medio de este sistema podemos realizar búsquedas avanzadas que tienen en cuenta parámetros multivariantes tanto de índole intralingüística (entorno textual, posición dentro de palabra, coocurrencias, etc.) como extralingüística (espacio, tiempo, estilo, registro, etc.).

Estos parámetros multivariantes son, precisamente, los atributos que el usuario puede establecer de manera ilimitada. En nuestro caso, hemos dispuesto once: referencia numérica del documento, archivo (signatura), rey o infante, fecha, año, lugar de emisión (origen), lugar o persona de recepción (destinatario), suscripción, redactor e *iussor*. De esta manera, podemos valorar la implicación de cada uno de esos atributos en relación con los datos lingüísticos. Esta metodología de trabajo la hemos aplicado recientemente en un análisis del valor discursivo de la puntuación y los resultados han sido bastante completos (Martín Aizpuru/Ueda [en prensa]).

En cuanto al tratamiento estadístico de frecuencias facilitado por *Lyneal* partiremos de la Frecuencia Absoluta, la Relativa (porcentaje) y la Normalizada por mil palabras, a los que sumaremos la Frecuencia Probabilística, recién introducida en el sistema. Esta última tiene la función de evaluar los valores estadísticamente comparables basándose en la probabilidad expectativa y la probabilidad de seguridad. Demostraremos la validez de la última frecuencia comparada con las tres anteriores tradicionales. A partir de la matriz construida de Frecuencias Probabilísticas realizaremos el Análisis de Conglomerado (Cluster),

el de Componentes Principales y finalmente el de la Correspondencia Unilateral. El último método, que presenta la distribución patronizada unilateral, de casos o de parámetros separados, es un nuevo método nuestro desarrollado a partir de la teoría del Análisis de Correspondencia de larga tradición, que ha venido ofreciendo la distribución patronizada bilateral, de casos y parámetros al mismo tiempo.

Con este sistema, en definitiva, esperamos contribuir al objetivo filológico que nos ha inquietado desde el comienzo del proyecto cancilleresco: conocer el sistema lingüístico de las cartas y comprobar cuáles son los factores extralingüísticos que afectan directamente a su extensión y evolución. Pretendemos ofrecer resultados lo más completos posibles tanto desde el punto de vista cuantitativo como cualitativo.

5 Conclusiones

El recorrido, descriptivo y expositivo, de este trabajo no pretende llegar a una serie de conclusiones, simplemente busca poner de manifiesto cómo los estudios filológicos pueden y deben apoyarse en medios técnicos para obtener resultados sólidos. Los proyectos de carácter amplio, como el que nos ocupa, pueden desarrollarse gracias al avance de los medios digitales que permiten análisis más exhaustivos y al mismo tiempo más refinados.

Es inevitable que los investigadores que se acercan al estudio de la historia de la lengua o a la labor de edición de los textos antiguos recurran a los medios digitales. Coincidimos con algunas aportaciones recientes que defienden que, pasado un periodo inicial de euforia, hoy en día la Lingüística de corpus en el ámbito iberorromance atraviesa una etapa más reflexiva y crítica que aspira a continuar mejorando no solo las herramientas de trabajo, sino también la metodología de uso por parte de los filólogos (Kabatek 2016: 1).

Con nuestra aportación en el mundo de la edición digital, marcación de textos con XML-TEI, el trabajo ha consistido, por ahora, en una reflexión teórica sobre la metodología, así como en la elaboración de una propuesta de guía para etiquetar textos históricos escritos en castellano. Esperamos que tenga continuidad con el objeto de resolver las cuestiones problemáticas que en esta revisión hemos destacado.

En cuanto a la extracción de datos lingüísticos y al tratamiento de datos estadísticos, debemos tener presente que el aprovechamiento de herramientas digitales no implica que el trabajo del filólogo sea más sencillo, sino todo lo contrario.

La posibilidad de cuantificar y visualizar cada vez más datos es la que nos está permitiendo repensar algunas afirmaciones sobre la evolución de la lengua.

Con la revisión del trabajo que ha realizado GEDHYTAS con relación al "Corpus de documentos de Cancillería Real, siglo XIII y primera década del XIV" en las últimas décadas, hemos querido mostrar la evolución que ha supuesto la incorporación de herramientas digitales destinadas a la elaboración de ediciones digitales y a la extracción de datos lingüísticos, sin olvidar el fin último de nuestras investigaciones que es el de contribuir a la reconstrucción de la historia textual y de la lengua.

Referencias bibliográficas

Chacón Gómez-Monedero, Francisco Antonio/Canorea Huete, Julián/Salamanca López, Manuel (2008): *Catálogo de la sección institucional del archivo de la catedral de Cuenca. I. Siglos XII-XIV*. Madrid: UAM Ediciones Cuenca, Ediciones de la Universidad de Castilla-La Mancha.

CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*, Red Internacional. [En línea, <http://www.corpuscharta.es/>, 10/03/2018].

CHARTA (2013): *Criterios de edición de documentos hispánicos (Orígenes-siglo XIX) de la Red Internacional CHARTA*. [En línea, https://www.redcharta.es/criterios-de-edicion, 10/03/2018].

Ciérbide, Ricardo/Ramos, Emiliana (2000): *Documentación medieval del Archivo Municipal de Pamplona (1357–1512)*. Donostia: Eusko Ikaskuntza.

De Lera Maíllo, José Carlos (1999): *Catálogo documental medieval de la catedral de Zamora*. Zamora: Instituto de Estudios zamoranos "Florián de Ocampo".

García Luján, Antonio(1982): *Los privilegios reales de la Catedral de Toledo* (1086–1462). Granada: Universidad de Granada.

González González, Julio (1980–1986): *Reinado y diplomas de Fernando III*. Monte de Piedad y Caja de Ahorros: Madrid.

González Jiménez, Manuel (1991): *Diplomatario Andaluz de Alfonso X*. Sevilla: El Monte, Caja de Huelva y Sevilla.

González Jiménez, Manuel/Carmona Ruiz, María Antonia (2012): *Documentación e itinerario de Alfonso X el Sabio*. Sevilla: Universidad de Sevilla.

Herrera, María Teresa/Sánchez González, María Nieves/González de Fauve, María Estela (1999): *Textos y concordancias electrónicos de documentos castellanos de Alfonso X*. Madison: Hispanic Seminary of Medieval Studies (ed. en CD).

Isasi, Carmen/Spence, Paul (coords.)/Lobo Puga, Ana/Martín Aizpuru, Leyre/Pérez Isasi, Santiago/Pierazzo, Elena (2014): *Guía para editar textos CHARTA según el estándar TEI: una propuesta*. [En línea, https://www.redcharta.es/investigacion/, 10/03/2018]

Kabatek, Johannes (2016): "Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. Introducción a este volumen". En: Kabatek, Johannes (ed.)/de Benito Moreno, Carlota (col.): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, pp. 1–17.

Kabatek, Johannes (2017): "Prólogo". En: Torruella Casañas, Joan (ed.): *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística*. Nueva York: Peter Lang edition, pp. 11–13.

Kasten, Lloyd A./Nitti, John J. (2002): *Diccionario de la prosa castellana del rey Alfonso X*. Nueva York: Hispanic Seminary of Medieval Studies, 3 volúmenes.

LYNEAL = *Letras y Números en Análisis Lingüísticos*. [En línea, http://shimoda.lllf.uam.es/ueda/lyneal/, 10/03/2018].

Martín Aizpuru, Leyre (2016): "Algunos recursos informáticos al servicio de la edición de textos: la edición en XML-TEI". En: Albertin, Chiara/del Rey Quesada, Santiago (coords.): *Hispanica Patavina. Estudios de historiografía e historia de la lengua española en homenaje a José Luis Rivarola*. Padua: CLEUP, pp. 139–154.

Martín Aizpuru, Leyre/Ueda, Hiroto (en prensa): "Parámetros multivariantes para el tratamiento de datos lingüísticos digitales (LYNEAL): puntuación manuscrita con función discursiva en documentación cancilleresca (siglo XIII)". Comunicación presentada en el *III Congreso de la Sociedad Internacional Humanidades Digitales Hispánicas (HDH), Sociedades, Políticas, Saberes*. Universidad de Málaga (18-20/10/2017).

Martín Fuertes, José Antonio (1998): *Colección documental del Archivo Municipal de León (1219–1400)*. León: Caja España de Inversiones y Archivo Histórico Diocesano de León.

Sánchez González de Herrero, Mª Nieves (2002): "Rasgos fonéticos y morfológicos de los documentos alfonsíes". En: *Revista de Filología Española*, LXXXII.1–2, pp. 139–177.

Sánchez-Prieto Borja, Pedro (2011): *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*. San Millán de la Cogolla : Cilengua.

Sánchez-Prieto Borja, Pedro (2012): "La red CHARTA: proyecto global de edición de documentos hispánicos". En: Torrens Álvarez, Mª Jesús/Sánchez-Prieto Borja, Pedro (eds.): *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Berna: Peter Lang, pp. 17–44.

Spence, Paul/Isasi, Carmen/Pierazzo, Elena/Vicente Miguel, Irene (2012): "Cruzando la brecha: la marcación digital con criterios filológicos". En: Sánchez-Prieto Borja, Pedro/Torrens Álvarez, Mª Jesús (eds.): *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Bern: Peter Lang, pp. 465–484.

TEI Consortium (2011): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [En línea, http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf, 10/03/2018].

Torrens Álvarez, Mª Jesús/Ueda, Hiroto (2016): "El nacimiento de la grafía jota como grafía consonántica. Análisis de documentos burgaleses de los siglos X-XIII". En: Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín: De Gruyter, pp. 299–321.

Torres Fontes, Juan (1977): *Colección de documentos para la historia del reino de Murcia. IV. Documentos de Sancho IV*. Murcia: Academia Alfonso X el Sabio.

Torruella Casañas, Joan (2017): *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística*. Nueva York: Peter Lang.

Ueda, Hiroto (2015): "La vocal débil en la apócope extrema medieval: Observaciones sobre el *Corpus de Documentos Españoles Anteriores a 1700*". En: Sánchez Méndez, Juan/de la Torre, Mariela/Codita, Viorica (coords.): *Temas, problemas y métodos para la edición y el estudio de documentos hispánicos antiguos*. Valencia: Tirant lo Blanch, pp. 585–607.

Ueda, Hiroto (2018): *Cómo usar Lyneal*. [En línea, <http://shimoda.lllf.uam.es/ueda/lyndat/doc/how-to-es.pdf>, 10/03/2018].

Ueda, Hiroto (en prensa): "Preguntas confirmativas en español. Análisis numérico de los datos de PRESEEA en el sistema LYNEAL". Comunicación presentada en el *XVIII Congreso Internacional de la Asociación de Lingüística y Filología de América Latina (ALFAL)*. Universidad Nacional de Colombia (24-28/07/2017).

Ueda, Hiroto/Moreno-Sandoval, Antonio (2015): "LETRAS and NÚMEROS: two integrated web-based tools for research in Linguistcs and Humanities". Comunicación presentada en el *7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived* Corpora *and* Beyond</i> (CILC 2015). Universidad de Valladolid (5-7/03/2015).