Scriptum digital
Vol. 14 (2025), pp. 7-36

Automatización de procesos en el desarrollo de corpus históricos: una propuesta desde las redes CHARTA Y EGPA[1]

VÍCTOR CABALLERO GÓMEZ (*Universidad de Salamanca / IEMYRhd*)
victorcaballero@usal.es
ORCID-iD: https://orcid.org/0000-0002-3522-1730

MIREIA PERIS VICENT (*Universidad Complutense de Madrid*)
mperis@ucm.es
ORCID-iD: https://orcid.org/0000-0002-1006-4687

RICARDO PICHEL (*Universidad Nacional de Educación a Distancia / Instituto da Lingua Galega, Universidade de Santiago de Compostela*)
ricardo.pichel@flog.uned.es
ORCID-iD: https://orcid.org/0000-0001-9933-0293

RESUMEN
En la creación y el desarrollo de corpus, algunas tareas pueden ejecutarse parcialmente de manera automática. La primera de ellas, en el caso de los corpus de textos manuscritos, es la propia transcripción de los testimonios. Del mismo modo, pueden automatizarse procesos como la conversión de los textos a XML, su normalización o su anotación. En este trabajo presentamos los desarrollos acometidos para automatizar algunas de estas tareas en los corpus que emplean los criterios de la Red CHARTA (Corpus hispánico y americano en la red: textos antiguos), en particular el Escritorio Galego-Portugués Antigo (EGPA), sometiéndolos a análisis y evaluando su mayor o menor grado de efectividad.

PALABRAS CLAVE: edición digital, anotación de corpus, lingüística de corpus, interoperabilidad, datos abiertos

Process automation for the development of historic corpus: a proposal from CHARTA and EGPA

ABSTRACT
Some tasks in corpus creation and development can be performed automatically. The first such task, in the case of corpora of handwritten texts, is the actual transcription of the testimony. In the same way, processes such as the conversion of texts into XML, their normalisation or their annotation can be automated. The aim of this paper is, firstly, to present the developments that have been made to automate these tasks in the corpora that follow the

[1] Este trabajo es resultado de los proyectos de investigación «CHARTA 4.0: adaptación y transferencia de recursos digitales para el desarrollo de corpus históricos» (RED2024-154111-T), «Alfonso de Cartagena. Obras completas III (ACOC III)» (PID2021-126557NB-I00) y «HERES: patrimonio textual panibérico. Recuperación y memoria» (CM/2018-T1/HUM-10230 y CM/2022-5A/HUM-24226). Asimismo, es parte de la ayuda PRE2022-105550, financiada por MCIN/AEI/10.13039/501100011033 y por el FSE+. Debemos agradecer a Belén Almeida Cabrejas, Elena Diez del Corral Areta, Miguel García-Fernández, Maarten Janssen y Gael Vaamonde la ayuda prestada en varias fases del trabajo y en múltiples dudas y problemas que incluso hoy nos siguen surgiendo.

Fecha de recepción: 22/09/2024
Fecha de aceptación: 22/09/2025
ISSN: 2014-640X
DOI: https://doi.org/ 10.5565/rev/scriptum.142

CHARTA (Corpus hispánico y americano en la red: textos antiguos) network criteria —in particular, the Escritorio Galego-Portugués Antigo (EGPA)—, and, secondly, to analyse these processes and evaluate their effectiveness to a greater or lesser extent.

KEYWORDS: Digital edition, Corpus annotation, Corpus linguistic, Interoperability, Open data

1. INTRODUCCIÓN

La creación y el desarrollo de un corpus implica la ejecución de una serie de tareas, definidas de un modo muy claro en el imprescindible trabajo que Torruella (2017: 63-65) dedica a la lingüística de corpus. Entre ellas, queremos someter a análisis las que pueden realizarse con un mayor o menor grado de automatización en el desarrollo de los macrocorpus CHARTA (Corpus hispánico y americano en la red: textos antiguos) y EGPA (Escritorio Galego-Portugués Antigo), así como en los distintos corpus que los integran. Para ello, tomaremos en consideración los desarrollos informáticos que permiten la ejecución automática o semiautomática de tareas como la conversión de los textos a XML, su normalización, lematización o etiquetado gramatical[2], así como las posibilidades de aplicación y el índice de acierto alcanzado con estas nuevas herramientas. Con ello, además, pretendemos evidenciar el flujo de trabajo seguido por CHARTA y EGPA.

Buena parte de los grupos que componen estas redes comparten el «creciente interés por los textos documentales como una fuente primordial a tener en cuenta tanto en la investigación filológica como en la histórica» (Diez del Corral Areta y Martín Aizpuru 2014: 296) y se han convertido en un auténtico foro «para el debate sobre la edición y estudio de textos archivísticos» (Sánchez-Prieto Borja 2012a: 42) en los ámbitos ibérico y americano. Por su parte, algunos de los proyectos integrados en la plataforma EGPA centran su interés también en el estudio y edición de textos de carácter literario (Pichel, García-Fernández y Caballero Gómez 2025). El indiscutible denominador común de los diversos grupos es el uso de los criterios de edición de documentos que acordó la Red CHARTA (2013), que tuvieron como precedente inmediato los trabajos de Sánchez-Prieto Borja (1998; 2009a; 2011) y cuya conveniencia ha quedado de sobra probada con su adopción en innumerables corpus y trabajos de investigación. Encontramos, entre los rasgos más característicos de estos criterios, la triple presentación de los textos, es decir, la posibilidad de acceder a tres instancias distintas del proceso de edición de los testimonios. De este modo, y de menor a mayor grado de abstracción, disponemos de la reproducción digital, de la transcripción paleográfica (TP) y de la presentación crítica (PC). Hasta la renovación en la que actualmente se encuentran inmersos los macrocorpus CHARTA y EGPA, los distintos estadios editoriales podían compararse únicamente de manera paralela. Ahora, además, se encuentran alineados por palabras[3], lo que facilita la comparación y explotación de las versiones paleográfica y crítica (Torruella 2017: 51-52)[4].

[2] Dichos recursos, que definiremos más adelante, pueden consultarse y descargarse en su versión más actualizada desde la página web del Grupo de Investigación de Textos para la Historia del Español (GITHE): https://corpora.uah.es/index.php?action=recursos [Consulta: 17/09/2025].

[3] Esta alineación se llevó a cabo en el seno de los proyectos de investigación «CHARTA 3.0: de la edición digital a la web semántica» (CM/JIN/2019-008) y el ya citado «HERES: patrimonio textual panibérico.

Esta alineación, que permite lematizar y etiquetar gramaticalmente los textos en una fase posterior, se ha visto favorecida por el uso de la edición digital en el estándar XML/TEI[5] y del entorno TEITOK (Janssen 2014; 2016; Janssen y Vaamonde 2020). Definido como *corpus management system* (Čermáková *et al*. 2021: 100), el entorno multifuncional TEITOK nos brinda la posibilidad de crear, desarrollar y explotar corpus en línea. Contamos, así, con una plataforma común empleable por todos los grupos de investigación que lo deseen, desde la que, además, pueden ejecutarse las herramientas digitales dirigidas a automatizar los procesos a los que aquí nos referimos. TEITOK se encuentra en constante evolución desde 2014 y ha servido a la creación, al desarrollo y a la explotación de corpus históricos como P.S. Post Scriptum (*Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*) u ODE (*Oralia Diacrónica del Español)*, entre muchos otros[6]. En la actualidad, sendos proyectos cuentan, por una parte, con 1.980.664 y 1.198.928 tokens; y, por otra, con 3968 y 999 documentos, respectivamente[7].

La variación cronológica y geográfica de los documentos que se incluyen en CHARTA y EGPA, expresada en términos escripto-lingüísticos, nos obliga a ser cautos respecto de la eficiencia de las diferentes tecnologías orientadas al reconocimiento automático de texto manuscrito para la edición de esta clase de fuentes. Efectivamente, una de las primeras tareas propias del desarrollo de un corpus que puede someterse a automatización es la transcripción de los textos. Hasta hace poco tiempo, el éxito de las estrategias de transcripción automática de texto manuscrito residía en la interrelación de datos estadísticos procedentes de niveles ópticos y lingüísticos (Romero *et al.* 2012: 15-18; Villegas *et al*. 2015: 831). Ello implicaba que cuanto mayor fuera la similitud, en todos esos aspectos, de los textos a transcribir respecto de los testimonios que habían servido para entrenar el modelo de transcripción automática, tanto mayor sería su índice de acierto. Era difícil concebir, por ese motivo, un modelo que sirviera para la transcripción de documentos que, a pesar de compartir cronología y usos gráficos, difirieran en su tipología o contenido. Idéntico problema encontramos en el caso opuesto. No obstante, y a pesar de que en la actualidad la mayoría de los sistemas de transcripción automática emplean únicamente el plano óptico, las pruebas iniciales realizadas sobre fuentes

Recuperación y memoria». Dicho proceso fue avalado, para el corpus CHARTA, en la Asamblea de la Red CHARTA celebrada en Granada el 9 de junio de 2022, en el marco del VII Congreso Internacional de la Red.

[4] De esta manera, las dos versiones editoriales se almacenan en un único documento fuente y, además, cada palabra y cada signo de puntuación, con su respectiva y posible variante entre versiones, se anida en un único elemento de la estructura del XML/TEI. Este sistema garantiza «la comparación entre las distintas versiones y la equiparación lingüística de sus elementos» (Torruella 2017: 52) al nivel del token.

[5] Un primer acercamiento teórico fue la propuesta de marcación XML/TEI basada en los criterios de la Red CHARTA y publicada en 2020. En ella se muestra como opción conceptual un etiquetado en fuente única, esto es, en un documento que alineara las versiones paleográfica y crítica, aunque de un modo diferente a la solución seguida finalmente en CHARTA. No obstante, en aquella ocasión se apostó, en detrimento de esta fuente única, por presentar «los diferentes casos de edición sobre el texto de la TP o de la PC» (Isasi Martínez *et al.* 2020: 11-12).

[6] Véase una lista en la que se recogen algunos de los proyectos que emplean TEITOK en http://www.teitok.org/index.php?action=projects [Consulta: 17/09/2025].

[7] Los datos han sido obtenidos en una consulta realizada el 12/03/2024.

documentales con modelos publicados todavía arrojan unas tasas de error en los caracteres transcritos superiores a las deseadas[8]. Este es el motivo por el que, al menos de momento, excluimos la transcripción como uno de los procesos que puedan realizarse de manera automática en el marco de los proyectos CHARTA y EGPAdoc[9].

Cuando disponemos de los textos transcritos, ya sea manual o automáticamente, la siguiente tarea que debe ocuparnos es su edición digital[10]. Emplear un lenguaje de marcación único como XML/TEI nos facilita compartir y explotar la información que hemos extraído al transcribir el documento. A pesar de ello, creemos que este uso no debe obligar al transcriptor a codificar el texto manualmente. La plataforma TEITOK permite integrar *scripts*, esto es, archivos en los que se incluyen una serie de órdenes o instrucciones. Gracias a ello, si definimos las órdenes concretas, podremos convertir un texto plano en un documento etiquetado en lenguaje XML/TEI. Este es el primero de los procesos que se han automatizado en los corpus que nos ocupan y al que prestaremos atención.

Como ya indicamos, la Red CHARTA prescribió la inclusión de un segundo estadio editorial, el de la presentación crítica, con el fin de facilitar la lectura del documento y su estudio morfológico, sintáctico, léxico e histórico. Se trata, de este modo, de una normalización de los «usos gráficos sin trascendencia fonética» que respeta «la variación que en algún momento de la historia de la escritura haya podido tener relevancia fónica» (Red CHARTA 2013: 20). Como veremos, un porcentaje no despreciable de este proceso se puede realizar de manera automática, gracias a un nuevo *script* y a un diccionario que cuenta con la equivalencia entre las formas propias de la TP y de la PC, por ejemplo, *vua*/*uva* o *uinna*/*viña*. Completado este proceso, nuevos *scripts* y diccionarios permiten, al menos en parte, automatizar la modernización de las formas críticas, así como su lematización y etiquetado gramatical. Estas dos últimas fases, las de la lematización y el etiquetado gramatical, pueden realizarse, en la plataforma TEITOK, con el etiquetador NeoTag, desarrollado por Janssen (2012) o con las herramientas, que aquí presentamos y analizamos, desarrolladas para dichos procesos por el grupo GITHE. Para facilitar la comprensión de los procesos y el flujo de trabajo, ofrecemos una pequeña tabla en la que incluimos los scripts a los que haremos referencia a lo largo de este texto.

[8] Puede verse al respecto el trabajo de Fradejas Rueda (2023) como un posible punto de partida en torno a la actual eficacia de este tipo de herramientas en el contexto que aquí mencionamos.

[9] En lo relativo a la transcripción automática debemos considerar como nuevo inconveniente que, si no se emplean modelos entrenados de acuerdo con los criterios de transcripción de la Red CHARTA, es necesaria la adaptación a los mismos.

[10] Nótese, no obstante, que algunas aplicaciones dirigidas a la transcripción automática de textos manuscritos como Transkribus, permiten también exportar una versión digital, en XML/TEI, de dicho trabajo (Caballero Gómez 2019: 60).

Tarea	Scripts disponibles en la web de GITHE	Scripts desarrollados por Janssen
1. Conversión a XML	charta_teitok_def2.pl	
2. Tokenización		tokenize.php (2014)
3. Normalización (PC)	nformtreat.pl / diccionario.vrt	
4. Modernización	mformtreat.pl / moderniza.vrt	
5. Etiquetado gramatical	postreat.pl / pos.vrt	NeoTag (2012)[11]
6. Lematización	lemmatreat.pl / lemma.vrt	

Tabla 1. Tareas y scripts mencionados en el texto

Como último paso, necesitamos ajustar diferentes aspectos de los corpus: el diseño estético de la interfaz, el menú de navegación de la web, las posibilidades de búsqueda o la manera más adecuada de visualizar los textos, entre otros. De nuevo, gracias a una plataforma como TEITOK, podemos reutilizar diseños y configuraciones que responden a las necesidades comunes de aquellos grupos que siguen los criterios de la Red CHARTA y que han sido desarrollados, de un modo general, pero no exclusivo, para los corpus CHARTA y EGPA.

2. REDES DE CORPUS: CHARTA Y EGPA

Contamos con varios trabajos en los que se han caracterizado, con suficiente profundidad, los objetivos de la Red CHARTA y la metodología empleada por esta (Sánchez-Prieto Borja 2012a; Sánchez González de Herrero *et al.* 2013; Diez del Corral Areta y Martín Aizpuru 2014, entre otros). Por ello, el nuestro será únicamente un intento de sintetizar y actualizar lo que ya se ha expuesto. Algo similar sucede con EGPA, que ha sido objeto de presentación en algunos congresos[12] y en dos recientes publicaciones (Pichel y García-Fernández 2024; Pichel, García-Fernández y Caballero Gómez 2025). CHARTA y EGPA constituyen redes colaborativas integradas por diferentes investigadores o grupos de investigación que comparten el objetivo de editar fuentes históricas documentales y/o literarias y que desarrollan una serie de corpus específicos con unos criterios editoriales comunes, a los que ya nos hemos referido. Esos corpus, que centran su atención en unas tipologías textuales, en una zona geográfica o en un periodo concretos, sirven a su vez para nutrir los repositorios compartidos, los que aquí denominamos macrocorpus CHARTA y EGPA.

[11] NeoTag, que sirve tanto al etiquetado gramatical como a la lematización de los textos, no se incluye por defecto en las instalaciones actuales de TEITOK.

[12] Entre otros, en el *XIII Congreso da Asociación Internacional de Estudos Galegos. Abrindo rutas, expandindo camiños. Novas perspectivas e intersecciónsnos estudos galegos* (Varsovia, 21 a 24 de septiembre de 2022), en el *VI Congreso Internacional de Corpus Diacrónicos en Lenguas Iberorrománicas* (Venecia, 5 a 7 de octubre de 2022), en el *Simposio ILG 2022*, titulado "A edición dixital de textos antigos: modelos, proxectos e ferramentas" (Santiago de Compostela, 28 de noviembre a 2 de diciembre de 2022), en la *I Xeira CLARIAH-GAL* (Santiago de Compostela, 2 de mayo de 2024) y en la *II Xeira CLARIAH-GAL* (Santiago de Compostela, 9 de mayo de 2025).

La Red CHARTA se concibió como «un espacio para el debate sobre la edición y estudio de textos archivísticos» (Sánchez-Prieto Borja 2012: 42) y en la actualidad (julio de 2024) está compuesta por un total de veintiséis grupos de investigación que desarrollan, a su vez, treinta y cuatro corpus lingüísticos, distribuidos según detallamos en la tabla 2.

Grupo	Institución (IP)	Corpus desarrollado(s)
ACOC | Alfonso de Cartagena. Obras completas	Universidad de Salamanca	- ACOCdigital
ARINTA | Archivo Informático de Textos de Andalucía	Universidad de Málaga	- CODEMA
CEIIBA, EA 7412 | Centre d'Études Ibériques et Ibéro-Américaines, EA 7412	Université Toulouse - Jean Jaurès	- ZavalDiCor
COHIECOS | Corpus Histórico del Español de Costa Rica	Universidad de Costa Rica	- COHIECOS
Cuba19 | Investigación en Humanidades Digitales y Español de América	Universidad de Sevilla	- CODHECUN
DiLEs | Diacronía de la Lengua Española	Universidad de Granada	- CORDEREGRA - ODE
DH | Diachronica Hispanica	Université de Neuchâtel	- DHISPAM
GEDHYTAS | Grupo de Estudio de Documentos Históricos y Textos Antiguos de la Universidad de Salamanca	Universidad de Salamanca	- CMIR - CODCAR
GEECOM | Grupo de Estudio del Español Colonial Mexicano	Universidad Nacional Autónoma de México	- COREECOM
GHEN | Grupo de Historia del Español Norteño	Consejo Superior de Investigaciones Científicas	- CORHEN
GIDC | Grupo de Investigación de Documentos de Canarias	Universidad de Las Palmas de Gran Canaria	- CODDEC
GITHE | Grupo de Investigación de Textos para la Historia del Español	Universidad de Alcalá	- ALDICAM - CODEA+ 2022 - CORDEX - CORDINA[13] - Letradas - NOSTOI[14]
GLD | Grup de Lexicografia i Diacronia	Universitat Autònoma de Barcelona	- ESenCAT
GLH-ULA | Grupo de Lingüística Hispánica	Universidad de Los Andes	- CDHM
GRAFILA | Grupo de Análisis Filológico de Lausana	Université de Lausanne	- CorColombia - COSUIZA - Fueros Medievales
HIPERTEXT | Perspectivas históricas sobre textos y discurso	Universitat de les Illes Balears	- Corpus Mallorca
Historia15	Universidad de Sevilla	- DOLEO - H15corpus
EGPA | Escritorio galego-portugués antigo[15]	UNED / Instituto da Lingua Galega (Universidade de Santiago de Compostela) / Instituto de Estudios Gallegos	- EGPAdoc - EGPAlit

[13] En coordinación con la UNED.

[14] En coordinación con la UNED y la Université Toulouse - Jean Jaurès.

[15] En lo que concierne a la edición de documentos (EGPAdoc), especialmente aquellos que a su vez son objeto de integración en CHARTA, participan junto con el Instituto da Lingua Galega otras instituciones como la Universidad de Alcalá, la Universidade de Santiago de Compostela, la Universidade de Vigo, la Universidade do Minho o la Universidade de Lisboa.

	Padre Sarmiento (CSIC-Xunta de Galicia)	
ICC (Col)	Instituto Caro y Cuervo	- DHLC
inTEXTA | Investigación de Textos de Archivo	Universidad de Jaén	- COHSANRE
KCL	King's College London	
PIHUGS | Proyecto de Investigaciones Hispánicas en la Universidad de Gotemburgo, Suecia	University of Gothenburg	
SAI | Seminario Alfonso Irigoien	Universidad de Deusto	- DGSXIX
SEAH | Semillero Español Histórico de Antioquia	Universidad de Antioquia	- SEAH
TesUN | Textos del español. Universidad de Navarra	Universidad de Navarra	- CORAPRINA
UTokyo	The University of Tokyo	

Tabla 2. Grupos de investigación y corpus de la Red CHARTA (julio de 2024)

Los diferentes corpus integrados en la Red CHARTA, como anticipamos, suelen estar compuestos por documentos procedentes de zonas y periodos previamente estipulados, textos producidos por perfiles sociolingüísticos determinados o testimonios que responden a unas tipologías documentales concretas[16]. Asimismo, los corpus se encuentran en muy diversos estados de desarrollo: desde los que están en fase de incorporación de textos a sus respectivas plataformas hasta los que cuentan con varios millares de documentos integrados. Del mismo modo, disponemos de repositorios que siguen incluyendo nuevos testimonios y otros en los que, al menos de momento, no se ha previsto añadir ninguno más. Toda la información disponible sobre estos desarrollos la incorporamos en la tabla 3[17].

Nombre	Descripción	Siglos	Referencias
ACOCdigital	Alfonso de Cartagena. Obras completas	XV	
ALDICAM	Atlas lingüístico diacrónico e interactivo de la Comunidad de Madrid	XIII-XIX	Sánchez-Prieto Borja 2018
CDHM	Corpus de documentos históricos de Mérida	XVI-XVIII	
CMIR	Corpus de documentación medieval castellana de Miranda de Ebro	XIII-XIV	Marcet Rodríguez y Sánchez González de Herrero 2016
CODCAR	Corpus de cancillería real castellana del siglo XIII	XIII-XIV	Martín Aizpuru y Sánchez González de Herrero 2019
CODDEC	Corpus documental de las Islas Canarias	XVI-XVIII	
CODEA+ 2022	Corpus de documentos españoles anteriores a 1900	XII-XIX	Sánchez-Prieto Borja 2009b; 2012b; Miguel Franco y Sánchez-Prieto Borja 2016
CODEMA	Corpus diacrónico de documentación malagueña	XVI-XVIII	Carrasco Campos *et al*. 2012
CODHECUN	Corpus documental y hemerográfico de la Cuba del Novecientos	XIX	Bravo García *et al*. 2021
COHIECOS	Corpus histórico del español de Costa Rica	XVIII-XIX	Cruz Volio 2021a; 2021b

[16] Este último es el caso, entre otros, de Oralia Diacrónica del Español, cuyos tipos textuales son los inventarios de bienes, las declaraciones de testigos en juicios penales y las certificaciones de barberos y cirujanos; es decir, aquellos que «suponen el traslado al papel de las declaraciones orales de testigos, tasadores y cirujanos» (Calderón Campos y Vaamonde 2020: 169).

[17] Es de gran utilidad, para completar la información que recogemos aquí, la plataforma CORHIBER, una herramienta que recopila los distintos corpus históricos en lenguas iberorrománicas.

COHSANRE	Corpus histórico del Santo Reino	XIV-XVI	Moreno Moreno 2018
CORAPRINA	Corpus de archivos de Navarra	XVI-XIX	Iraceburu Jiménez *et al*. 2018; Tabernero Sala 2020
CORDEREGRA	Corpus diacrónico del español del reino de Granada	XVI-XVIII	Calderón Campos y García-Godoy 2009
CORDINA	Corpus diacrónico del español de Norteamérica	XVI-XX	Sánchez Sierra 2021; Diez del Corral Areta y Pichel 2022; Giménez-Eguíbar 2022; Giménez-Eguíbar y Pichel 2023
COREECOM	Corpus electrónico del español colonial mexicano	XVI-XVIII	Arias Álvarez 2009; 2012
CORHEN	Corpus histórico del español norteño	X-XIII	Moral del Hoyo 2019
Corpus Mallorca	Documentos castellanos en archivos de las Islas Baleares	XVII-XX	Enrique-Arias y Miguel Franco 2015; Enrique-Arias 2023
COSUIZA	Corpus de documentos hispánicos de Suiza	XVI-XX	Castillo Lluch y Diez del Corral Areta 2015; 2018
DGSXIX	Documentación guipuzcoana del siglo XIX	XIX	Gómez Seibane 2013
DHISPAM	Diachronica Hispanica Americana	XVI-XVIII	
DOLEO	Documentación de lamento en español desde orígenes	XV-XIX	Pons Rodríguez *et al*. 2014
EGPAdoc	Escritorio Galego-Portugués Antigo (Textos documentais)	X-XIX	Pichel y Varela Barreiro 2014; Pichel, García-Fernández y Caballero Gómez 2025
ESenCAT	Español en Cataluña	XVIII-XIX	Torruella y Clavería 2019; Torruella 2020
H15Corpus	H15Corpus	XV-XVII	Garrido Martín y Martín Aizpuru 2024
Letradas	Corpus de textos escritos por mujeres en España (1400-1900)	XV-XIX	
ODE	Oralia diacrónica del español	XV-XIX	Calderón Campos 2019; Calderón Campos y Vaamonde 2020; Calderón Campos y Díaz Bravo 2021
SEAH	Semillero del español histórico de Antioquia	XVIII-XIX	Ospina Giraldo y González-Rátiva 2021

Tabla 3. Información localizada[18] sobre los corpus integrados en la Red CHARTA (julio de 2025)

Por su parte, EGPA se concibe como una plataforma abierta y colaborativa desde la que consultar diferentes corpus y repositorios textuales de variadas características y contenidos, todos ellos centrados en la producción escrita vinculada principalmente a Galicia y Portugal desde la Edad Media hasta el siglo XIX[19]. Por ello, se presenta como un recurso pensado para investigadores e incluso docentes de diferentes disciplinas relacionadas con el estudio, consulta y análisis de las fuentes escritas (García-Fernández y Pichel 2025b). Cuenta tanto con fuentes documentales (EGPAdoc) como con fuentes literarias (EGPAlit), editadas, igualmente, según los criterios de la Red CHARTA adaptados, en algún caso, a las especificidades de la

[18] Recogemos en esta tabla los corpus para los que hemos localizado referencias bibliográficas y aquellos que, a pesar de carecer de ellas, han contribuido con sus documentos al desarrollo del corpus CHARTA.

[19] En términos técnicos, esta plataforma y sus funcionalidades son el resultado de la combinación de las herramientas facilitadas por TEITOK y otras diseñadas específicamente en el seno del proyecto.

variedad lingüística (gallego-portuguesa) mayormente representada. Por ello, a su vez, se integra en esta Red y contribuye a la expansión del eje geográfico del macrocorpus CHARTA[20]. EGPA tiene su origen en un proyecto de investigación, centrado en la transcripción paleográfica de textos gallegos medievales, iniciado en 2007 en el seno del Instituto da Lingua Galega de la Universidade de Santiago de Compostela (Pichel y Varela Barreiro 2014: 291-292). Fruto del trabajo de dicho proyecto nació en 2015 el Corpus de Textos Antigos da Galiza (COTAGAL), que posteriormente, desde 2019 y en el marco del proyecto HERES, se convirtió en punto de partida para el desarrollo de EGPA. Hasta el momento actual, en EGPA se han integrado un total de doce proyectos, de los cuales ocho trabajan específicamente con fuentes documentales (EGPAdoc), como se muestra en la tabla 4[21].

Nombre	Descripción	Siglos	Referencias
EGPAdoc			
CORESGAL	Corpus de espontáneas de Galicia	XVIII-XX	García-Fernández 2016, 2018a, 2018-19; García-Fernández y Otero Piñeyro Maseda 2022; García-Fernández y Pichel 2025a
CORTESGAL	Corpus testamentario de Galicia	XII-XIX	García-Fernández 2017, 2018b, 2019
CORREGAM	Corpus rexio e rexinal da Galicia medieval	XI-XVI	García-Fernández, Pelaz Flores y Pichel 2020; García-Fernández y Pichel 2024a
eASPA	Edición dixital do fondo medieval do Arquivo de San Paio de Antealtares (Santiago de Compostela)	IX-XVI	García-Fernández y Pichel 2023a
eCaTui	Edición dixital dos protocolos notariais de Joán Rodríguez (Catedral de Tui, 1413-1442)	XV	Costas Fragueiro 2023
eReginae	Escrita e rainhas. As chancelarias reginais como instrumentos de poder (sécs. XI-XVI)	XI-XVI	Olaia 2023
eRiaPon	Edición dixital dos mosteiros da ría de Pontevedra	XIII-XVI	Chapela Durán 2024
GalScript	GALLÆCIÆ SCRIPTORES: edición dixital dos repertorios notariais na Galiza medieval		García-Fernández y Pichel 2025c

Tabla 4. Información sobre los proyectos y corpus integrados en EGPAdoc (julio de 2024)

Además de ofrecer un buscador general desde el que explotar la totalidad de los testimonios documentales, filtrando o no según el interés particular de la búsqueda, la

[20] A esta misma expansión contribuyen las recientes incorporaciones de nuevos corpus centrados en el estudio y edición de documentos americanos, como son los corpus CORDINA (UNED / Universidad de Alcalá) o COHIECOS (Universidad de Costa Rica).

[21] Más allá de la documentación asociada a estos proyectos específicos, EGPA integra otras iniciativas de investigación más dispersas pero igualmente relacionadas con la recuperación del patrimonio documental gallego-portugués (véase, a modo de ejemplo, García-Fernández y Pichel 2020, 2023b, 2024b, 2024-25; García-Fernández 2025). En lo que concierne al EGPAlit, véase Pichel y García-Fernández (2024) y Pichel, García-Fernández y Caballero Gómez (2025) en lo que concierne a EGPAlit. La acomodación de TEITOK y de la metodología CHARTA para la creación de corpus y ediciones digitales de textos literarios, aunque adolece de ciertos desarrollos aún en curso, se ha ido potenciando paulatinamente en los últimos años, como se puede ver, entre otros, en algunos de los subcorpus integrados en EGPA (ejs. 7PDgp, ARANHIS o MELE), los corpus NOSTOI (Pascual-Argente y Pichel 2024) y Panépica Digital (Rodríguez Molina y Vaamonde 2025) o el proyecto Lozana Digital (Díaz Bravo y Vaamonde 2020).

plataforma aloja diferentes módulos de consulta a través de los cuales se permite la navegación por las distintas fuentes y tipologías textuales. Asimismo, contamos, en algunos casos, con portales específicos, integrados en el propio EGPA, que recogen y permiten aprovechar el trabajo desarrollado específicamente por cada proyecto[22].

3. INSTALACIÓN Y CONFIGURACIÓN DIGITAL DEL CORPUS

Unas líneas atrás nos referimos a la tarea de configuración digital como último de los pasos en el proceso de creación y desarrollo de un corpus. En este apartado relacionaremos instalación y configuración, por lo que parece tener algo más de sentido tratarlo como el primero de los procesos que pueden ser, en cierto modo, automatizados, aunque el orden, en definitiva, puede verse alterado. De una manera análoga a lo que sucede con la creación de páginas web, que se ha visto facilitada gracias al desarrollo de diferentes sistemas de gestión de contenidos (*content management system* o CMS) como WordPress o Drupal, entre muchos otros, el desarrollo digital de corpus textuales ha experimentado una cierta revolución en la última década debido a la creación de TEITOK. Se trata, en sentido estricto, de un sistema de gestión de corpus que facilita y aglutina las tareas digitales de creación, desarrollo y explotación de este tipo de repositorios. Como sucede en un CMS, TEITOK permite, si se configuran adecuadamente, crear e incorporar recursos que son de igual utilidad para varios corpus.

Para disponer de TEITOK en línea es necesario, en primer lugar, contar con un servidor web[23], donde, posteriormente, se instalará la aplicación. Gracias a la estructura con la que se ha diseñado, TEITOK nos permite disponer de una primera página web que sirva, si se desea, como portal del grupo de investigación, desde el que se enlacen y aniden nuevas páginas web en las que se desplieguen los corpus. Así sucede en la instalación realizada en el servidor de la Universidad de Alcalá. La página raíz (https://corpora.uah.es/) contiene información relativa al Grupo de Investigación de Textos para la Historia del Español (GITHE). A partir de ahí, los distintos corpus ocupan carpetas diferenciadas y se accede a ellos incorporando a la URL su denominación. Sucede, de este modo y entre otros, con los macrocorpus CHARTA y EGPA, que se alojan en este servidor común y cuentan con las direcciones específicas https://corpora.uah.es/charta/ y https://corpora.uah.es/egpa/, respectivamente.

Cuando contamos con la instalación de TEITOK y creamos un corpus, es decir, habilitamos un espacio virtual en el que incluiremos contenido, debemos decidir cómo será su interfaz. El diseño de esta se realiza con una plantilla inteligente y única (*main.tpl*) para todo el corpus, en la que se incorporan elementos fijos, como la disposición de los componentes (logo, menú de navegación, cambio de lengua, etc.) y elementos de tipo variable, esto es, vínculos al contenido dinámico que debe mostrar

[22] Es el caso, entre otros, de los proyectos ARANHIS y 7PDgp, cuyos desarrollos particulares pueden verse, además de en EGPAlit, en su repositorio específico: https://corpora.uah.es/aranhis/ y https://corpora.uah.es/7pdgp/ [Consulta: 17/09/2025].

[23] Los requisitos técnicos con los que debe contar este servidor se detallan en la página web de TEITOK, espacio en el que también encontramos el instalador de la aplicación y las instrucciones para su uso: http://www.teitok.org/index.php?action=help&id=install [Consulta: 17/09/2025].

la aplicación en cada página (buscador, mapa de documentos, visualizador de textos, etc.). Con el objetivo de facilitar esta tarea, GITHE ha desarrollado y pone a disposición de cualquier interesado una sencilla interfaz, en la que, además, pueden realizarse todos los cambios que se deseen[24]. A pesar de que dicha plantilla no abandone el diseño con el menú vertical, alineado a la izquierda y habitual en TEITOK, para facilitar el uso de las páginas web en todos los dispositivos, algunos corpus como CHARTA, EGPA o Letradas ya cuentan con un diseño responsivo y en el que el menú de navegación se dispone horizontalmente[25].

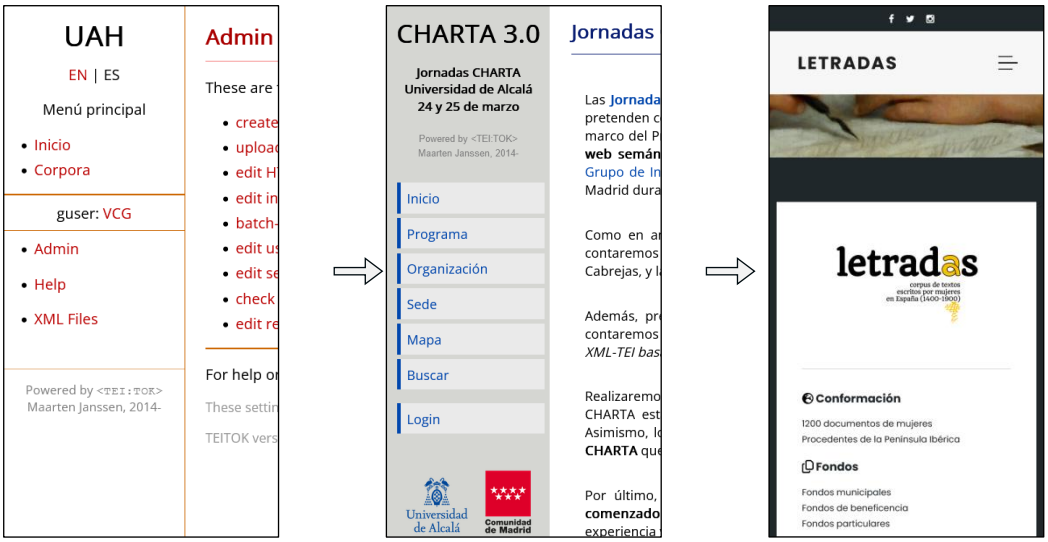

Figura 1. Diseño vertical original, diseño vertical facilitado por GITHE y diseño horizontal de Letradas

En este punto debe tenerse en cuenta si se desea o no que el corpus sea indexado y rastreable por los motores de búsqueda de Internet. En TEITOK, por defecto, los corpus impiden el rastreo. Será en esta plantilla, contenida en el archivo *main.tpl*, donde debamos modificar la metainformación relativa a este fenómeno. Si retomamos el caso de Letradas, pese a la polisemia del término y a su elevado uso en otros contextos de carácter oficial, en la actualidad, el buscador más empleado, Google, muestra la página web del corpus entre los primeros tres resultados de la búsqueda «letradas». Ello se debe a la adecuada incorporación de los metadatos y a la inclusión del sitio en el servicio que Google facilita para la indexación de dominios, Google Search Console.

```
<meta name="description" content="El Grupo de Investigación de Textos para la Historia del Español (GITHE) está
    coordinado por Pedro Sánchez-Prieto Borja y formado por investigadoras/es de la Universidad de Alcalá
    (España), y cuenta con una amplia experiencia en la edición y estudio de textos españoles antiguos.">
<meta name="keywords" content="GITHE, CHARTA, historia del español">
<meta name="author" content="GITHE - Grupo de Investigación de Textos para la Historia del Español">
<!--<meta name="robots" content="noindex,nofollow">-->
```

[24] Puede descargarse desde la misma página https://corpora.uah.es/index.php?action=recursos.
[25] No disponemos de una plantilla descargable y reutilizable con este tipo de diseño debido a que su configuración debe hacerse de manera muy específica en cada caso.

Figura 2. Metadatos relativos a la indexación presentes en el archivo *main.tpl* de https://corpora.uah.es/

La siguiente tarea que debe ocuparnos, al abordar la configuración de un corpus en el entorno TEITOK, es la elección de una serie de opciones que se recogen en un archivo de crucial importancia denominado *settings.xml*. En él encontramos la definición de: a) el teiHeader o cabecera de los documentos XML, b) los diferentes atributos que caracterizan los tokens, c) las opciones de búsqueda y explotación del corpus y d) los *scripts* ejecutables por los editores digitales. Con todo ello podemos adoptar los planteamientos teóricos y metodológicos seguidos por la Red CHARTA. Dado que lo que aquí nos ocupa son los *scripts* con los que automatizar procesos en la preparación de los textos, mostraremos, en la siguiente figura, ese apartado específico del archivo *settings.xml*. Dentro de este archivo, todos los detalles se anidan en <ttsettings>. Como puede apreciarse (fig. 3), debemos disponer de un elemento <scripts>, en el cual se deben referenciar los distintos ejecutables. Se incorporan con el elemento <item>, que a su vez cuenta con atributos que detallan el tipo de *script*, la acción que debe ejecutar o la ruta para acceder a él. Con este código seremos capaces de integrar en TEITOK los *scripts* que permiten convertir texto plano en XML/TEI, normalizar la transcripción paleográfica, modernizar las formas críticas, lematizar y etiquetar gramaticalmente los textos.

```
<scripts>
    <item key="charta_teitok_def2" action="perl Scripts/charta_teitok_def2_EGPA.pl [fn]" display="TP (CHARTA
    3.0)" recond="\{(\d|[a-z])" type="perl"/>
    <item key="nformtreat" action="perl Scripts/nformtreat.pl [fn]" display="PC (CHARTA 3.0/EGPA)"/>
    <item key="mformtreat" action="perl Scripts/mformtreat.pl [fn]" display="Modernizacion (EGPA)"/>
    <item key="lemmatreat" action="perl Scripts/lemmatreat.pl [fn]" display="Lematizacion (EGPA)"/>
    <item key="postreat" action="perl Scripts/postreat.pl [fn]" display="Etiquetado POS (EGPA)"/>
</scripts>
```

Figura 3. Detalle del archivo *settings.xml* relativo a la inclusión de *scripts*

4. ETIQUETADO AUTOMÁTICO DE LOS TEXTOS

Hasta el momento, la manera más habitual de llevar a cabo la edición digital de un texto ha implicado su codificación manual. Para ello, el editor debe conocer la sintaxis de XML, así como el lenguaje TEI, o al menos los aspectos de este que sean empleados en su trabajo editorial[26]. Además, conviene estar familiarizado con algún editor de texto que facilite la codificación en XML, aunque las alternativas que ofrecen mejores funcionalidades son *softwares* de pago. Parece deseable, por estos inconvenientes, contar al menos con la posibilidad de realizar el etiquetado de los textos de manera automática, opción ofrecida en cualquier caso como complementaria a la opción del

[26] Los inicios de la adaptación a XML-TEI de los criterios de la Red CHARTA, así como los problemas asociados a esta, pueden encontrarse en los trabajos de Spence *et al.* (2012), Spence (2014), Martín Aizpuru (2016) e Isasi Martínez *et al.* (2020).

etiquetado manual[27]. Diferentes especificaciones y un veterano lenguaje de programación denominado Perl, muy utilizado para procesar texto, han permitido su desarrollarlo e integración en la plataforma TEITOK. Tras múltiples versiones, realizadas primero para CHARTA y después también para EGPA, contamos ahora con un *script* de suficiente eficacia, rebautizado con el nombre <charta_teitok_def2_EGPA.pl>.

Para ello fue necesario, en primer lugar, definir un protomarcado sistemático que pudiera incluirse en el *script* y que se basara, en la medida de lo posible, en el lenguaje plano de los criterios de la Red CHARTA (2013). Con ese objetivo, se eliminaron las cursivas de las marcas tradicionales con las que se indican elementos diplomáticos o intervenciones en el texto del escribiente o del editor. Asimismo, se estipuló, de manera inalterable, el modo de protoetiquetar algunas intervenciones que se marcaban de formas diversas, como por ejemplo, el cambio de mano o las intervenciones correctivas coetáneas a la escritura del texto. En otro orden de modificaciones y para garantizar una adecuada interpretación de las marcas anidadas por parte del *script*, se han duplicado los caracteres de apertura y cierre, que ahora ya no son únicamente los corchetes, sino también las llaves. En la siguiente figura insertamos tres bloques de texto que intentan explicar la lógica del proceso: a) línea transcrita en texto plano, b) instrucciones del script que ejecutan transformaciones y c) resultado de la ejecución del *script* en XML.

```
{h 1r} {1} [lat.: Jn {mano 2: dej} no<m>i<n>e.] Conosçida cosa sea a q<ua>ntos esta carta viere<n> & oyere<n> como
{2} yo fray Sancho frayre de ffitero & Graniero de Naguera co<n>
```

⇩

```
#inicio de folio
$text =~ s/\{h +(\d+(r|v|))\}/<pb n="$1"\/>/g;
$text =~ s/\{h+(\d+(r|v|))\}/<pb n="$1"\/>/g;
$text =~ s/\{h. +(\d+(r|v|))\}/<pb n="$1"\/>/g;
$text =~ s/\{h.+(\d+(r|v|))\}/<pb n="$1"\/>/g;

#inicio de línea
$text =~ s/\{(\d+)\}/<lb n="$1"\/>/g;

#abreviaturas
$text =~ s/&lt;(.+?)&gt;/<ex>$1<\/ex>/g;

#cambio de lengua (latín)
$text =~ s/\[lat(\.|):\s*([^\]]+)\]/<foreign xml:lang="la">$2<\/foreign>/g;
$text =~ s/\{lat(\.|):\s*([^\]]+)\}/<foreign xml:lang="la">$2<\/foreign>/g;
$text =~ s/\[la(\.|):\s*([^\]]+)\]/<foreign xml:lang="la">$2<\/foreign>/g;
$text =~ s/\{la(\.|):\s*([^\]]+)\}/<foreign xml:lang="la">$2<\/foreign>/g;
$text =~ s/\[latín:\s*([^\]]+)\]/<foreign xml:lang="la">$2<\/foreign>/g;
$text =~ s/\{latín:\s*([^\]]+)\}/<foreign xml:lang="la">$2<\/foreign>/g;

#cambio de mano
$text =~ s/\[mano\s*(\d):\s*([^\]]+)\]/<handShift resp="\#h$1">$2<\/handShift>/g;
$text =~ s/\{mano\s*(\d):\s*([^\]]+)\}/<handShift resp="\#h$1">$2<\/handShift>/g;
$text =~ s/\[m(\d):\s*([^\]]+)\]/<handShift resp="\#h$1">$2<\/handShift>/g;
$text =~ s/\{m(\d):\s*([^\]]+)\}/<handShift resp="\#h$1">$2<\/handShift>/g;
```

⇩

[27] Única opción ofrecida, por ejemplo, en la propuesta de Isasi Martínez *et al*. (2020).

```
<pb n="1r"/> <lb n="1"/> <foreign xml:lang="la">Jn <handShift resp="#h2">dej</handShift>
no<ex>m</ex>i<ex>n</ex>e.</foreign> Conosçida cosa sea a q<ex>ua</ex>ntos esta carta viere<ex>n</ex> &amp;
oyere<ex>n</ex> como <lb n="2"/> yo fray Sancho frayre de ffitero &amp; Graniero de Naguera co<ex>n</ex>
```

Figura 4. Texto plano, líneas del *script* invocadas y resultado final

En el primero de los cuadros de texto vemos el *input* que introducimos en TEITOK, que responde a las especificaciones realizadas. Después, en el segundo, hemos recogido las instrucciones contenidas en el *script* que realizan la transformación vinculada al inicio de folio y línea, al desarrollo de abreviaturas, al cambio a lengua latina y al cambio de mano. La extensión completa del protomarcado utilizable como *input* y el conjunto de las instrucciones del *script* la incluiremos en un próximo trabajo, dedicado específicamente a estos aspectos[28]. El último cuadro de texto muestra el resultado, en lenguaje XML/TEI, de las transformaciones efectuadas por el *script* al *input* introducido en texto plano. Con este sistema y con el correcto uso del protomarcado se asegura el éxito en la conversión a XML/TEI de casi cualquier documento de la Red CHARTA, al menos de aquellos en los que se incluya hasta un nivel de marca anidada, como la contenida en el ejemplo previo, «[lat.: Jn {mano 2: dej} no<m>i<n>e.]», o cualquier otra similar como «[lat.: Jn {mano 2: dej} {mano 3: no<m>i<n>e}.]».

5. TOKENIZACIÓN DEL DOCUMENTO XML/TEI

En los próximos apartados nos referiremos a la normalización, modernización, lematización y etiquetado gramatical de los textos. El modo que estableció el estándar TEI para ejecutar estos procesos fue el uso de la etiqueta <w>, complementada por los atributos @lemma y @pos y en la que se anida un elemento <choice>, en el que se volverían a anidar los elementos <orig> y <reg>. Si tomamos en consideración la primera de las palabras del ejemplo de la figura 4, el etiquetado podría ser el siguiente:

```
<w lemma="in" pos="APPR">
    <choice>
        <orig>Jn</orig>
        <reg>In</reg>
    </choice>
</w>
```

Figura 5. Caracterización lingüística de una palabra según el estándar TEI

Ciertamente, uno de los mayores aciertos de TEITOK radica en la simplificación de este etiquetado, lo que se consigue gracias a la introducción de la etiqueta <tok>, en la que pueden incluirse ya no solo palabras como en la etiqueta <w>, sino también signos de puntuación, es decir, tokens. Las variaciones que se produzcan en los distintos estadios editoriales (tantos como deseemos) ya no deben ser anidadas dentro del

[28] Por supuesto, las órdenes de transformación que se incluyen en el *script* pueden verse en el propio archivo.

elemento <choice>, por el contrario, serán un atributo más, como lo son @lemma y @pos. La tokenización (Janssen 2016: 4038; Vaamonde 2015: 60; 2018: 154), es decir, la imposición de etiquetas <tok>, se realiza de manera automática gracias a un *script* que encontramos, por defecto, en cualquier instalación de TEITOK. Como decimos, en este proceso,

> cada forma original del texto es marcada dentro de un elemento <tok>, al que se le asigna una identificación única también de manera automática. Esta estructura inicial permite delimitar cada token para su posterior edición lingüística y permite salvaguardar además los diferentes niveles de edición, que se van almacenando en forma de atributos dentro de cada unidad <tok> (Vaamonde 2018: 154).

De ese modo, la forma transcrita será introducida en su correspondiente etiqueta <tok>, que a su vez podrá contar, cuando se añada esa información, con los atributos @lemma para su lema y @pos para su categorización gramatical, y, si hubiera variación, con los atributos @nform para su normalización gráfica y @mform para su forma moderna. Asimismo, si la palabra estuviera abreviada, se incluirían los atributos @form para la forma sin desarrollar y @fform para la forma desarrollada. Estos dos últimos atributos, @form y @fform, se cumplimentan de manera automática en el mismo proceso de tokenización. Tomamos como ejemplo la palabra «q<ua>ntos»:

```
<tok form="qntos" fform="quantos" nform="cuantos" lemma="cuanto" pos="PR0MP000">
    q<ex>ua</ex>ntos
</tok>
```

Figura 6. Caracterización lingüística y paleográfica de una palabra en TEITOK

A continuación caracterizaremos los elementos relativos a la forma normalizada, forma modernizada, lema y etiqueta gramatical o POS (*part-of-speech*), así como las posibilidades de automatizar su cumplimentación.

6. NORMALIZACIÓN Y MODERNIZACIÓN (SEMI)AUTOMÁTICA

En el I Congreso Internacional de la Red CHARTA, celebrado en Madrid en 2009, así como en el volumen monográfico que derivó de aquel encuentro, Horcajada Diezma (2012) presentó una aplicación informática, denominada *BConcord* 2010, que permitía, entre muchas otras opciones, automatizar lo que el autor denominaba la estandarización gráfica de los textos, esto es, lo que nosotros, siguiendo otras tradiciones, denominamos normalización, presentación crítica o edición normalizada. Por desgracia, nada sabemos del estado actual de dicho proyecto, aparentemente descontinuado, ni de los recursos que se podrían haber generado en su entorno, a pesar de lo enormemente beneficiosos que habrían resultado para los desarrollos que analizamos en las siguientes líneas. Nos ocuparemos, en ellas, de las funcionalidades que nos permiten automatizar algunos aspectos de la normalización y de la modernización de los textos que previamente han sido transcritos, etiquetados en XML/TEI y tokenizados.

Como anticipábamos, los criterios de edición de documentos acordados por la Red CHARTA facilitan un triple acceso al texto: a) la reproducción digital, b) la transcripción paleográfica y c) la presentación crítica. Las formas recogidas en la presentación crítica son, en definitiva, la normalización gráfica de las formas paleográficas y, por ello, podemos establecer una equivalencia unívoca entre ambas (paleográfica y crítica) en una significativa cantidad de ocasiones. Sucede, por ejemplo y según indicamos antes, con los pares de valores *vua*/*uva* o *uinna*/*viña*. No podemos, sin embargo, establecer la misma correlación en casos de ambigüedad, como, entre muchos otros, en la forma paleográfica *donde*, que podría normalizarse *donde* pero también *dónde*. Dado que el sistema que aquí presentamos funciona únicamente con un diccionario y no tiene en cuenta el contexto de aparición, dichos casos deben ser desambiguados manualmente[29]. El mismo procedimiento manual ha de seguirse para regularizar la puntuación del texto[30]. Pese a que los criterios CHARTA no incluyen un estadio editorial en el que se modernice el texto, contamos igualmente con un *script* y un diccionario que permiten realizar, del mismo modo que en el supuesto anterior, dicha modernización. El único objetivo que persigue este paso es el de facilitar los posteriores, es decir, la lematización y el etiquetado gramatical. En ningún caso este último estadio, el de la modernización, sería consultable por los usuarios del corpus, puesto que no está sometido a revisión, como sí el resto de las etiquetas lingüísticas. En la figura 7 contamos con una pequeña muestra de los diccionarios y los pares de valores con los que se realizan las transformaciones de la forma paleográfica a la forma crítica y de la forma crítica a la forma modernizada.

Diccionario normalización	
Forma paleográfica	**Forma crítica**
çibdad	cibdad
çibdadano	cibdadano
çibdades	cibdades
çibdat	cibdat

Diccionario modernización	
Forma crítica	**Forma moderna**
cibdá	ciudad
cibdad	ciudad
cibdades	ciudades
cibdat	ciudad
ciubdades	ciudades
ciudá	ciudad
ciudat	ciudad
civdad	ciudad
siudad	ciudad
ziudad	ciudad

Figura 7. Formas del lema «ciudad» incluidas en los diccionarios normalizador y modernizador

En la elaboración de estos diccionarios se han empleado los inventarios léxicos del corpus CODEA, en concreto, los que recogen las distintas formas de los documentos procedentes de Castilla la Vieja (Sánchez-Prieto Borja y Ueda 2018) y Castilla la Nueva (Agujetas Ortiz, Sánchez-Prieto Borja y Ueda 2022). Se encuentran en constante revisión, realizada de manera manual, para introducir o excluir aquellos pares de

[29] Uno de los retos a los que nos debemos enfrentar en el futuro es introducir en este sistema un módulo capaz de realizar, también de manera automática, la desambiguación morfosintáctica, al estilo de lo presentado por Aguilar *et al*. (2004-2005).

[30] Para facilitar la reasignación de tokens tras esta revisión manual, contamos con un nuevo script realizado por Maarten Janssen.

valores que conviene añadir o eliminar en función de las necesidades detectadas con su uso. En el momento de redacción de este texto, el primero de los diccionarios, que se encarga de la normalización, contiene un total de 11.616 formas en las que existe variación entre la transcripción paleográfica y la presentación crítica. El segundo de ellos, el que ejecuta la modernización de los textos, contiene más de 50.000 pares de valores.

Para evitar posibles sesgos derivados de la prueba de estas herramientas con textos del corpus CODEA, utilizado, en último término, para generarlas, las someteremos a revisión con dos fragmentos de documentos procedentes de otros corpus que también emplean los criterios editoriales propuestos por la Red CHARTA[31]. En el primer caso empleamos las dos primeras líneas de una carta de compraventa del CORHEN, datada en Revilla de la Fuente (Burgos, España) en el año 1287. Para el segundo, recurrimos a las cuatro primeras líneas de un inventario de bienes datado en Heredia (Costa Rica) en 1803, que forma parte del COHIECOS. Con el objetivo de evidenciarlas, en las siguientes tablas subrayamos en gris las transformaciones realizadas, primero, por el *script* de normalización y, después, por la revisión manual.

CORHEN-0474 | Archivo del Monasterio de Las Huelgas de Burgos, leg. 35, n. 1663		
Transcripción paleográfica	**Resultado *script* normalización**	**Presentación crítica**
{1} Sepan. quantos esta carta vieren. Como. yo don ffray yenego por la gracia de dios. Abbat del monesterio de sant xristoual de {2} Eueas. Conplazimjento & con otorgamjento de todo el conuento desse mismo monesterio. vendemos Auos domjng yuannez el clerigo del	{1} Sepan. cuantos esta carta vieren. como. yo don fray yenego por la gracia de Dios. abat del monesterio de sant Cristóval de {2} Eueas. Conplazimjento e con otorgamiento de todo el convento d'esse mismo monesterio. vendemos a vós domjng Yuáñez el clérigo del	{1} Sepan cuantos esta carta vieren cómo yo don fray Yénego, por la gracia de Dios abat del monesterio de Sant Cristóval de {2} Eveas, con plazimiento e con otorgamiento de todo el convento d'esse mismo monesterio, vendemos a vós Doming Yuáñez, el clérigo del

Tabla 5. Prueba de normalización en CORHEN-0474

COHIECOS-0001 | Archivo Nacional de Costa Rica, Mortuales Coloniales de Heredia, exp. 1991, ff. 3r-4r		
Transcripción paleográfica	**Resultado *script* normalización**	**Presentación crítica**
{1} En la Poblacion de Eredia a los veinte y sie{2}te dias del mes de mayo de mil ochocientos once; yo {3} Don Julian Rodriguez Alcalde primero de ella, {4} y sus terminos; para proceder a los Ynventarios y	{1} En la población de Eredia a los veinte y sie{2}te días del mes de mayo de mil ochocientos once; yo {3} don Julian Rodríguez alcalde primero de ella, {4} y sus términos; para proceder a los Ynventarios y	{1} En la población de Eredia a los veinte y sie{2}te días del mes de mayo de mil ochocientos once; yo, {3} don Julián Rodríguez, alcalde primero de ella {4} y sus términos; para proceder a los inventarios y

Tabla 6. Prueba de normalización en COHIECOS-0001

En el primero de los casos (CORHEN-0474), de un total de cuarenta y una formas paleográficas, requieren normalización dieciocho y el *script* realiza doce, mientras que

[31] Intencionadamente, omitimos aquí la prueba basada en la modernización de los textos, ya que su función es instrumental, no editorial, y su índice de éxito debe ponerse en relación con las siguientes fases, las de la lematización y el etiquetado gramatical. Por su parte, los fragmentos han sido elegidos por su especial distancia cronológica y geográfica.

seis deben realizarse de manera manual. En el siguiente (COHIECOS-0001) contamos con treinta y cinco formas paleográficas, de las cuales ocho deben normalizarse. En seis casos el *script* interviene y en dos debe intervenir el editor. Dado que existe el riesgo de que una forma paleográfica sea normalizada erróneamente, tomamos en consideración, para el cálculo del índice de éxito, no solo las formas que deben someterse a normalización, sino también aquellas en las que no hay variación. Obtenemos, así, una tasa de acierto de palabras normalizadas que se sitúa en el 89,47%. Si analizamos las formas cuya normalización no se ha realizado adecuadamente con el *script*, descubrimos que, en su mayor parte, se trata de nombres propios que no se encuentran en el diccionario (con casos como «Yénego» o «Eveas») y formas que requieren una desambiguación morfosintáctica (por ejemplo, «cómo» o «Sant»).

7. LEMATIZACIÓN Y ETIQUETADO GRAMATICAL

Janssen (2012) desarrolló una herramienta, denominada NeoTag, diseñada para lematizar y anotar morfosintácticamente los textos, que mejoraba el reconocimiento y etiquetado de neologismos, es decir, de palabras con una nueva acepción o categoría gramatical, diferente de la que se recogía previamente en el diccionario utilizado. La tasa de acierto, tanto de este como de otros etiquetadores de primer nivel, se aproxima al 98%, gracias a que calculan la probabilidad de que a una palabra le corresponda una etiqueta u otra en función de su posición dentro de la secuencia de palabras en la que aparece (Janssen 2012: 2118). De ese modo se consigue un elevado índice de éxito incluso en casos que requieran una desambiguación morfosintáctica. A pesar de ello, es inevitable revisar de manera manual el etiquetado. Un caso paradigmático a este respecto es, por ejemplo, el de la palabra «dicho», como acertadamente trataron Calderón Campos y García-Godoy (2023: 163-164), que puede ser anotada como participio invariable, participio concordado, determinante demostrativo y nombre común.

Con el objetivo de facilitar la desambiguación manual del etiquetado gramatical y de la lematización en los textos que siguen los criterios de la Red CHARTA, los mencionados proyectos CHARTA 3.0 y HERES desarrollaron una nueva técnica que recopila las distintas opciones de anotación para cada palabra y permite al editor elegir la correcta en cada caso, todo ello dentro de la propia plataforma TEITOK. Asimismo, se pueden gestionar aquellos supuestos en los que el etiquetador no debe ofrecer estas posibilidades, sino optar por un etiquetado de manera automática y descartar otro por su baja probabilidad de aparición. Un ejemplo de esto último lo encontramos en la palabra «mayo», que será etiquetada como nombre común, aunque también pueda ser una forma del presente de indicativo del verbo «mayar». Su funcionamiento, como en la normalización y modernización, depende de dos *scripts* escritos en lenguaje Perl y de dos diccionarios elaborados a partir de los datos que para el castellano moderno ofrece FreeLing (Lloberes, Castellón y Padró 2010)[32]. De este modo, el *script* busca el

[32] Los nombres propios, en cambio, son incorporados desde los ya mencionados inventarios léxicos del corpus CODEA.

atributo @mform de cada token en el correspondiente diccionario y le añade los atributos @pos y @lemma asociados a cada palabra.

En cuanto al etiquetado gramatical, siguiendo la práctica común en la plataforma TEITOK, se recurre a las *Recommendations for the Morphosyntactic Annotation of Corpora* desarrolladas por el grupo EAGLES (1996). Gracias a ellas podemos asignar un código con un número variable de caracteres alfanuméricos, en función de los cuales se definen los diferentes accidentes gramaticales de cada palabra. Así, por ejemplo, al pronombre interrogativo «cúyo» le corresponde la etiqueta POS «PT0MS00», ya que se trata de un pronombre (P), de tipo interrogativo (T), sin persona definible (0), de género masculino (M), de número singular (S), sin caso (0) ni poseedor (0)[33]. Las recomendaciones de EAGLES permiten añadir, además, datos de tipo semántico. Para ello, reservan las posiciones quinta y sexta del código atribuido a los nombres. Queda pendiente, en este sentido, la inclusión de dicha información, que debería aprovechar las posibilidades exploradas por Agujetas Ortiz y Sánchez-Prieto Borja (2022) acerca de la clasificación del vocabulario y su recuperación en corpus históricos.

En la siguiente tabla ofrecemos un ejemplo con la primera línea del documento CORHEN-0474, gracias a la cual podemos observar el funcionamiento de los procesos automáticos de lematización y etiquetado gramatical basados en estos recursos, así como las necesidades de intervención manual derivadas.

@form	**@nform**	**@mform**	**@lemma**	**@pos**
Sepan	Sepan	sepan	saber	@ VMM03P0 | VMSP3P0
quantos	cuantos	cuantos	cuanto	@ DI0MP0 | PR0MP00
esta	esta	esta	esto	@ DD0FS0 | PD0FS00
carta	carta	carta	carta	NCFS000
vieren	vieren	vieren	ver	VMSF3P0
Como	cómo	cómo	cómo	@ PT00000 | PE00000
yo	yo	yo	yo	PP1CSN0
don	don	don	don	NCMS000
ffray	fray	fray	fray	NCMS000
yenego	Yénego	-	-	-
por	por	por	por	SPS00
la	la	la	la	@ DA0FS0 | PP3FSA0
gracia	gracia	gracia	gracia	NCFS000
de	de	de	de	SPS00
dios	Dios	Dios	Dios	NP00000
Abbat	abat	abad	abad	NCMS000
del	del	del	de+el	SPCMS
monesterio	monesterio	monasterio	monasterio	NCMS000
de	de	de	de	SPS00
sant	Sant	San	San	AQ0MS0
xristoual	Cristóval	Cristóbal	Cristóbal	NP00000
de	de	de	de	SPS00

[33] Las tablas a partir de las que se definen las categorías y los accidentes gramaticales, así como algunos ejemplos, pueden verse en: https://corpora.uah.es/charta/index.php?action=pos [Consulta: 17/09/2025]. Asimismo, es del todo útil a este respecto el cuidado etiquetado de *Oralia Diacrónica del Español*, que creemos puede servir como guía en el tratamiento de múltiples soluciones: http://corpora.ugr.es/ode/index.php?action=tagset [Consulta: 17/09/2025].

Tabla 7. Prueba de lematización y etiquetado gramatical en CORHEN-0474

Si analizamos los datos obtenidos, comprobamos que, del total de veintidós palabras, se atribuye correctamente la forma moderna y el lema a veintiuna (95,45%). La forma que falta por etiquetar es «Yénego», un nombre propio no incluido en los diccionarios. Por su parte, el etiquetado gramatical atribuye el código correcto en dieciséis ocasiones, mientras que en cinco ofrece dos posibilidades entre las que elegirá el editor y en una de ellas, la misma que en el caso anterior, no asigna etiqueta POS. El recuento de atributos @pos asignados de manera correcta e inequívoca es del 72,72%, lo que reduce considerablemente el índice de acierto que, como apuntamos anteriormente, se obtiene con NeoTag. No obstante, el método que aquí presentamos cuenta con la ventaja de ofrecer las distintas posibilidades en caso de ambigüedad, así que podemos aislar dichos casos y revisarlos de un modo, creemos, más sencillo y controlado.

8. CONCLUSIONES

Un primer balance de lo aquí expuesto nos debe hacer reparar en la enorme y merecida importancia que ha adquirido la herramienta TEITOK para la elaboración de corpus digitales. Un gran número de grupos de investigación la emplea ya en sus proyectos, pues permite, como hemos visto, la creación, el desarrollo y la explotación de corpus, entre otros, de carácter histórico. Esa relevancia es extensible, en el ámbito de la Red CHARTA, al corpus *Oralia Diacrónica del Español*, pionero en el uso de esta herramienta y guía, según apuntamos, en la ejecución de una serie de prácticas que han facilitado los desarrollos necesarios para CHARTA y EGPA.

Creemos que puede valorarse de un modo muy positivo el diseño e implementación de las variadas soluciones que hemos analizado aquí y que facilitan la automatización de procesos en el desarrollo de corpus, más aún cuando siguen las orientaciones y criterios de la Red CHARTA, que han tenido un relevante impacto en los ámbitos europeo y americano. Esto, sumado a la interoperabilidad de estas herramientas, permite que sean empleadas por los múltiples grupos que componen CHARTA y EGPA o que, simplemente, siguen sus criterios. De hecho, si hacemos un repaso de los corpus que ya las utilizan, contamos más de una veintena, incluidos los macrocorpus CHARTA y EGPA, los subcorpus que componen EGPA, ACOCdigital, CODEMA, CODHECUN, COHIECOS, CorColombia, CORDEX, CORDINA, COSUIZA, Fueros Medievales, Letradas y ZavalDiCor. No puede ignorarse que en la base de todo ello se encuentran los nunca suficientemente reconocidos esfuerzos por crear unas redes que sirvan de catalizador y promotor de iniciativas como la redacción de unos criterios editoriales para la edición de documentos antiguos o como la elaboración de las herramientas que aquí presentamos.

Como esperamos haber demostrado, estos desarrollos dirigidos a la automatización de diferentes procesos agilizan la creación y el desarrollo de corpus históricos que emplean los criterios de la Red CHARTA y, al mismo tiempo, facilitan la incorporación y ejecución de soluciones homogéneas. En efecto, una de las grandes ventajas de incorporar herramientas que automaticen los procesos puede medirse no solo en términos temporales, sino también en un sentido propiamente científico. No es

extraño ni erróneo que dos editores propongan soluciones diferentes al mismo problema, pero esto puede suponer un inconveniente a la hora de explotar un corpus cuyos criterios se presuponen homogéneos. Automatizar ese tipo de soluciones disminuye las ocasiones en las que el editor debe enfrentarse a este tipo de dificultades, lo que redunda en una menor disparidad de criterios.

Todo ello no obsta para que siga habiendo retos que afrontar, así como actuaciones muy necesarias que llevar a cabo en lo relacionado con el desarrollo y la explotación del tipo de corpus que aquí hemos presentado. Entre otros problemas a los que esperamos se pueda ofrecer una próxima solución encontramos el relativo a la posibilidad de descargar los textos editados (por ejemplo, para citarlos en una publicación) en un formato que incluya todos los detalles del etiquetado y las marcas habituales que se recogen en los criterios de la Red CHARTA. También debe arbitrarse una solución convincente para el etiquetado de los textos escritos con anterioridad al triunfo del romance, a la luz, especialmente, de lo establecido por Torrens Álvarez (2019). Asimismo, esperamos ofrecer en el futuro una información más específica y desarrollada de varios de los procesos y las herramientas aquí presentadas y su aplicación en contextos de trabajo real en los corpus.

REFERENCIAS BIBLIOGRÁFICAS

7PDgp = PICHEL, Ricardo (dir.): *7PDgp. Edição digital das Sete Partidas na Galiza e Portugal*. Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/7pdgp/ [Consulta: 17/09/2025].

ACOCdigital = VALERO MORENO, Juan Miguel (dir.): *ACOCdigital: Alfonso de Cartagena. Obras completas*. http://corpus.usal.es/acoc/ [Consulta: 17/09/2025].

AGUILAR, Lourdes, Ana Belén AVILÉS, Jordi FONTSECA, Carme DE LA MOTA, Yolanda RODRÍGUEZ SELLÉS, Paola Guadalupe CAYMES SCUTARI, Sergi BALARI (2004-2005): «Un módulo de desambiguación morfosintáctica para el castellano basado en conocimiento lingüístico», *Revista española de lingüística aplicada*, 17-18, pp. 7-17.

AGUJETAS Ortiz, María y Pedro SÁNCHEZ-PRIETO BORJA (2022): «Nuevas vías para la recuperación de información en corpus históricos: clasificación del vocabulario», *Scriptum digital*, 11, pp. 5-54. https://raco.cat/index.php/scriptumdigital/article/view/412601. [Consulta: 17/09/2025].

AGUJETAS ORTIZ, María, Pedro SÁNCHEZ-PRIETO BORJA e Hiroto UEDA (2022): *Inventario léxico de Castilla la Nueva*. https://h-ueda.sakura.ne.jp/lyneal/il/cn/ [Consulta: 17/09/2025].

ALDICAM = GRUPO DE INVESTIGACIÓN DE TEXTOS PARA LA HISTORIA DEL ESPAÑOL [GITHE]: *Atlas lingüístico diacrónico e interactivo de la Comunidad de Madrid*. http://aldicam.blogspot.com/ [Consulta: 17/09/2025].

ARANHIS = PICHEL, Ricardo, Carmen BENÍTEZ GUERRERO y Filipe Alves MOREIRA (coords.): *ARANHIS: Archivum Annalisticum Hispanum*. Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/aranhis/ [Consulta: 17/09/2025].

ARIAS ÁLVAREZ, Beatriz (2009): «Confección de un corpus para conocer el origen, la evolución y la consolidación del español en la Nueva España», en Andrés Enrique-Arias (ed.): *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid-Frankfurt am Main: Iberoamericana-Veuvert, pp. 55-76. https://doi.org/10.31819/9783865278685-005 [Consulta: 17/09/2025].

ARIAS ÁLVAREZ, Beatriz (2012): «Configuración de un corpus colonial y caracterización de subcorpus que ayuden al conocimiento del español colonial mexicano», en María Jesús

Torrens Álvarez y Pedro Sánchez-Prieto Borja (eds.): *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Berna: Peter Lang, pp. 307-322.

BRAVO GARCÍA, Eva, Ana MANCERA RUEDA y Leyre MARTÍN AIZPURU (2021): «Edición de un corpus de documentos sobre los movimientos de insurgencia en Cuba durante la segunda mitad del siglo XIX», *Scriptum digital*, 10, pp. 53-73. https://raco.cat/index.php/scriptumdigital/article/view/395976 [Consulta: 17/09/2025].

CABALLERO GÓMEZ, Víctor (2019): «La codificación XML en el ámbito de las Ciencias Historiográficas. Algunas propuestas para su uso y desarrollo», *Revista de Humanidades Digitales*, 4, pp. 57-68. https://doi.org/10.5944/rhd.vol.4.2019.25136 [Consulta: 17/09/2025].

CABALLERO GÓMEZ, VÍCTOR, Miguel GARCÍA-FERNÁNDEZ y Ricardo PICHEL (2025): «O Escritorio Galego-Portugués Antigo (EGPA). Novos avances técnicos e proxectos en andamento», en II Xeira CLARIAH-GAL (Santiago de Compostela, 9 de mayo de 2025). https://ilg.usc.gal/gl/actividades/II-xeira-clariah-gal [Consulta: 17/09/2025]

CALDERÓN CAMPOS, Miguel (2019): «La edición de corpus históricos en la plataforma TEITOK. El caso de "Oralia diacrónica del español"», *Chimera: Romance Corpora and Linguistic Studies*, 6, pp. 21-36.

CALDERÓN CAMPOS, Miguel y M.ª Teresa GARCÍA-GODOY (2009): «El Corpus diacrónico del español del Reino de Granada (CORDEREGRA)», en Andrés Enrique-Arias (ed.): *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid-Frankfurt am Main: Iberoamericana-Veuvert, pp. 229-250. https://doi.org/10.31819/9783865278685-014 [Consulta: 17/09/2025].

CALDERÓN CAMPOS, Miguel y Gael VAAMONDE (2020): «Oralia Diacrónica del Español: un nuevo corpus de la Edad Moderna», *Scriptum digital*, 9, pp. 167-189. https://raco.cat/index.php/scriptumdigital/article/view/377292 [Consulta: 17/09/2025].

CALDERÓN CAMPOS, Miguel y Rocío DÍAZ BRAVO (2021): «An online corpus for the study of historical dialectology: Oralia diacrónica del español», *Digital Scholarship in the Humanities*, 1, pp. 1-19. https://doi.org/10.1093/llc/fqaa066 [Consulta: 17/09/2025].

CALDERÓN CAMPOS, Miguel y M.ª Teresa GARCÍA-GODOY (2023): «"Y el dicho testigo dijo su dicho". Gramaticalización y etiquetado de *dicho* en el corpus *Oralia Diacrónica del Español* (ODE)», en Patricia Giménez-Eguíbar *et al*. (eds.): *Despertar palabras, renacer historias. Estudios lingüísticos en homenaje a M.ª Nieves Sánchez González de Herrero*. Salamanca: Universidad de Salamanca, pp. 155-167.

CARRASCO CAMPOS, Inés, Livia Cristina GARCÍA AGUIAR y Pilar LÓPEZ MORA (2012): «El corpus CODEMA: una base documental para el estudio de la norma meridional», en M.ª Ángeles Peinado Herreros (coord.): *I Congreso Internacional «El patrimonio cultural y natural como motor de desarrollo: investigación e innovación»*. Sevilla: Universidad Internacional de Andalucía, pp. 2140-2152.

CASTILLO LLUCH, Mónica y Elena DIEZ DEL CORRAL ARETA (2015): «El fondo Balbueno de la Universidad de Lausana», *Scriptum digital*, 4, pp. 109-123. https://raco.cat/index.php/scriptumdigital/article/view/316361 [Consulta: 17/09/2025].

CASTILLO LLUCH, Mónica y Elena DIEZ DEL CORRAL ARETA (2018): «Fondos documentales hispánicos en Suiza: una exploración de conjunto», *Scriptum digital*, 7, pp. 95-105. https://raco.cat/index.php/scriptumdigital/article/view/343467 [Consulta: 17/09/2025].

CDHM = GRUPO DE LINGÜÍSTICA HISPÁNICA-UNIVERSIDAD DE LOS ANDES [GLH-ULA]: *Corpus de Documentos Históricos de Mérida*. https://corpora.uah.es/charta/ [Consulta: 17/09/2025].

CHARTA = RED CHARTA: *Corpus hispánico y americano en red: textos antiguos (CHARTA)*. https://corpora.uah.es/charta/ [Consulta: 17/09/2025].

CHAPELA DURÁN, Daniel (2024): *Edición e glosario da documentación en galego do século XIII do mosteiro de San Xoán de Poio*. Trabajo Final de Máster. Santiago de Compostela: Universidade de Santiago de Compostela.

CMIR = GRUPO DE ESTUDIO DE DOCUMENTOS HISTÓRICOS Y TEXTOS ANTIGUOS DE LA UNIVERSIDAD DE SALAMANCA [GEDHYTAS]: *Corpus de documentación medieval castellana de Miranda de Ebro*. https://campus.usal.es/~gedhytas/index.php/txt/doc/cmir [Consulta: 17/09/2025].

CODCAR = GRUPO DE ESTUDIO DE DOCUMENTOS HISTÓRICOS Y TEXTOS ANTIGUOS DE LA UNIVERSIDAD DE SALAMANCA [GEDHYTAS]: *Corpus de cancillería real castellana del siglo XIII*. https://campus.usal.es/~gedhytas/index.php/txt/doc/ccan [Consulta: 17/09/2025].

CODEA+ 2022 = GRUPO DE INVESTIGACIÓN DE TEXTOS PARA LA HISTORIA DEL ESPAÑOL [GITHE]: *Corpus de documentos españoles anteriores a 1900*. https://corpuscodea.es/ [Consulta: 17/09/2025].

CODEMA = ARCHIVO INFORMÁTICO DE TEXTOS DE ANDALUCÍA [ARINTA]: *Corpus Diacrónico de Documentación Malagueña*. http://teitok.uma.es/codema/ [Consulta: 17/09/2025].

CODHECUN = BRAVO GARCÍA, Eva, Ana MANCERA RUEDA y Leyre MARTÍN AIZPURU (dirs.) (2022): *Corpus Documental y Hemerográfico de la Cuba del Novecientos*. http://cuba19.us.es/ [Consulta: 17/09/2025].

COHIECOS = CRUZ VOLIO, Gabriela (dir.): *Corpus Histórico del Español de Costa Rica*. https://teitok.ucr.ac.cr/ [Consulta: 17/09/2025].

COHSANRE = GRUPO INTEXTA: *Corpus Histórico del Santo Reino*. https://corpora.uah.es/charta/ [Consulta: 17/09/2025].

CORAPRINA = TESUN: *Corpus de archivos privados de Navarra*. https://corpora.unav.edu/ [Consulta: 17/09/2025].

CorColombia = DIEZ DEL CORRAL ARETA, Elena (dir.): *Corpus del español de Colombia*. https://grafila.unil.ch/corcolombia/ [Consulta: 17/09/2025].

CORESGAL = GARCÍA-FERNÁNDEZ, Miguel (dir.): *Corpus de espontáneas de Galicia*. Integrado en Escritorio Galego-Portugués Antigo. https://corpora.uah.es/coresgal/ [Consulta: 17/09/2025].

CORDEREGRA = CALDERÓN CAMPOS, Miguel y María Teresa GARCÍA-GODOY (dirs.) (2015): *Corpus Diacrónico del Español del Reino de Granada*. http://corpora.ugr.es/ode/ [Consulta: 17/09/2025].

CORDEX = SÁNCHEZ SIERRA, Diego (dir.): *Corpus diacrónico de Extremadura*. https://corpora.uah.es/cordex/ [Consulta: 17/09/2025].

CORDINA = PICHEL, Ricardo y Diego SÁNCHEZ SIERRA (dirs.): *Corpus diacrónico del español de Norteamérica*. https://corpora.uah.es/cordina/ [Consulta: 17/09/2025].

COREECOM = GRUPO DE ESTUDIO DEL ESPAÑOL COLONIAL MEXICANO [GEECOM]: *Corpus Electrónico del Español Colonial Mexicano*. https://doi.org/10.19130/coreecom.clh.2019 [Consulta: 17/09/2025].

CORHEN = TORRENS ÁLVAREZ, María Jesús (dir. y ed.) (2016): *Corpus Histórico del Español Norteño*. http://corhen.es/ [Consulta: 17/09/2025].

CORHIBER = TORRUELLA, Joan y Johannes KABATEK (2018-): *Portal de Corpus Históricos Iberorrománicos*. http://www.corhiber.org/ [Consulta: 17/09/2025].

Corpus Mallorca = ENRIQUE-ARIAS, Andrés (dir.): *Documentos castellanos en archivos de las Islas Baleares*. https://www.corpusmallorca.es/ [Consulta: 17/09/2025].

CORREGAM = GARCÍA-FERNÁNDEZ, Miguel, Diana PELAZ FLORES y Ricardo PICHEL (coords.): *Corpus rexio e rexinal da Galicia medieval*. Integrado en Escritorio Galego-Portugués Antigo. https://corpora.uah.es/egpa/corregam [Consulta: 17/09/2025].

CORTESGAL = GARCÍA-FERNÁNDEZ, Miguel (dir.): *Corpus testamentario de Galicia*. Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/egpa/cortesgal [Consulta: 17/09/2025].

COSTAS FRAGUEIRO, Brais (2023): *Primeira aproximación aos protocolos notariais de Johán Rodríguez (Catedral de Tui, 1413-1442)*. Trabajo Final de Grado. Santiago de Compostela: Universidade de Santiago de Compostela. http://hdl.handle.net/10347/31088 [Consulta: 17/09/2025].

COSUIZA = GRUPO DE ANÁLISIS FILOLÓGICO DE LAUSANA [GRAFILA]: *Corpus de documentos hispánicos de Suiza*. https://grafila.unil.ch/cosuiza/ [Consulta: 17/09/2025].

COTAGAL = PICHEL, Ricardo y Xavier VARELA BARREIRO (dirs.): *Corpus de Textos Antigos da Galiza*. https://ilg.usc.gal/es/proxectos/corpus-de-textos-antiguos-de-galicia-cotagal [Consulta: 17/09/2025].

CRUZ VOLIO, Gabriela (2021a): «Hacia la conformación de un corpus histórico para el español colonial de Costa Rica», *Diseminaciones*, 4 (7), pp. 79-98.

CRUZ VOLIO, Gabriela (2021b): «Cuestiones sobre la selección y la edición de documentos coloniales para un corpus histórico del español de Costa Rica», en Alexánder Sánchez Mora, Gabriela Cruz Volio y José Luis Ramírez Luengo (eds.): *La palabra olvidada: la lengua y la literatura de Centroamérica entre la Colonia y la Independencia*, vol. I. San José: Encino, pp. 17-57.

ČERMÁKOVÁ, Anna, Jarmo JANTUNEN, Tommi JAUHIAINEN, John KIRK, Michal KŘEN, Marc KUPIETZ y Elaine Uí DHONNCHADHA (2021): «The International Comparable Corpus: Challenges in building multilingual spoken and written comparable corpora», *Research in Corpus Linguistics*, 9 (1), pp. 89-103. https://doi.org/10.32714/ricl.09.01.06 [Consulta: 17/09/2025].

DGSXIX = GÓMEZ SEIBANE, Sara (dir.): *Documentación guipuzcoana del siglo XIX*. https://corpora.uah.es/charta/ [Consulta: 17/09/2025].

DHISPAM = Diacrhonica Hispanica [DH]: *Diachronica Hispanica Americana*. https://www.unine.ch/espagnol/home.html [Consulta: 17/09/2025].

DHLC = RUIZ VÁSQUEZ, Néstor Fabián (dir.): *Documentos para la historia lingüística de Colombia; siglos XVI a XIX*. Integrado en el *Corpus Diacrónico y Diatópico del Español de América*. https://www.cordiam.org/ [Consulta: 17/09/2025].

DÍAZ BRAVO, Rocío y Gael VAAMONDE (2020): «Creación de ediciones digitales para lingüistas de corpus: el caso del *Retrato de la Loçana andaluza*», en José R. Belda-Medina y Ricardo Casañ-Pitarch (eds.): *Análisis del Discurso en la Era Digital: Una Recopilación de Casos de Estudio*. Granada: Comares, pp. 17-34. https://hdl.handle.net/10481/89863 [Consulta: 17/09/2025].

DIEZ DEL CORRAL ARETA, Elena y Leyre MARTÍN AIZPURU (2014): «Sin corpus no hay historia: la Red CHARTA como un proyecto de edición común», *Cuadernos de Lingüística*, 2, pp. 287-314.

DIEZ DEL CORRAL ARETA, Elena y Ricardo PICHEL (2021): «Fenómenos de contacto español-francés en un corpus epistolar franco-chileno (s. XIX)», *Cuadernos del Instituto Historia de la Lengua*, 14, pp. 187-212. https://doi.org/10.58576/cilengua.vi14.19 [Consulta: 17/09/2025].

DOLEO = HISTORIA15: *Documentación de lamento en español desde orígenes*. https://corpora.uah.es/charta/ [Consulta: 17/09/2025].

GalScript = GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (dirs.): *GALLÆCIÆ SCRIPTORES: edición dixital dos repertorios notariais na Galiza medieval*. https://corpora.uah.es/egpa/galscript [Consulta: 17/09/2025].

eASPA = PICHEL, Ricardo y Miguel GARCÍA-FERNÁNDEZ (dirs.): *e-ASPA. Edición dixital do patrimonio documental medieval de San Paio de Antealtares*. Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/egpa/easpa [Consulta: 17/09/2025].

eRiaPon = CHAPELA, Daniel (dir.), Miguel GARCÍA-FERNÁNDEZ y Ricardo PICHEL (coords.): *e-RiaPon. Edición dixital dos mosteiros da ría de Pontevedra*. Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/egpa/e-riapon [Consulta: 17/09/2025].

eCaTui = COSTAS FRAGUEIRO, Brais: *eCaTui. Edición dixital dos protocolos notariais de Joán Rodríguez (Catedral de Tui, 1413-1442)*. Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/egpa/ecatui [Consulta: 17/09/2025].

EGPA = PICHEL, Ricardo y Miguel GARCÍA-FERNÁNDEZ (dirs.): *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/egpa/ [Consulta: 17/09/2025].

ENRIQUE-ARIAS, Andrés y Ruth MIGUEL FRANCO (2015): «Una nueva herramienta para el estudio histórico del castellano en contacto con el catalán en Mallorca», en Juan Pedro Sánchez Méndez, Mariela de La Torre y Viorica Codita (eds.): *Temas, problemas y métodos para la edición y el estudio de documentos hispánicos antiguos*. Valencia: Tirant lo Blanch, pp. 407-426.

ENRIQUE-ARIAS, Andrés (2023): «El *Corpus Mallorca*». Una herramienta para el estudio histórico del castellano en contacto con el catalán en Mallorca», en Miguel Calderón Campos e Inmaculada González Sopeña (eds.): Scripta manent. *Historia del español, documentación archivística y humanidades digitales*. Lausanne: Peter Lang, pp. 535-561.

eReginae = RODRIGUES, Ana Maria S. A. y Maria Manuela Tavares dos Santos SILVA (dirs.): *eReginae. Escrita e rainhas. As chancelarias reginais como instrumentos de poder (sécs. XI-XVI)* Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/egpa-ereginae/ [Consulta: 17/09/2025].

ESenCAT = TORRUELLA, Joan: *Español en Cataluña* [en fase de desarrollo].

FRADEJAS RUEDA, José Manuel (2023): «El Coloso Español», en José Manuel FRADEJAS RUEDA (ed.), *7PartidasDigital. Edición crítica digital de las «Siete Partidas»*. Valladolid: Universidad de Valladolid. https://doi.org/10.58079/vemr [Consulta: 17/09/2025].

Fueros Medievales = CASTILLO LLUCH, Mónica: *Fueros Medievales*. https://grafila.unil.ch/fuerosmedievales/ [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel (2016): «As espontáneas de San Xoán de Río. Sexualidade extraconxugal e tentativas de control municipal das mulleres no século XIX», en *Álbum de mulleres*. Santiago de Compostela: Comisión de Igualdade do Consello da Cultura Galega. http://culturagalega.gal/album/detalle.php?id=1044 [Consulta: 28/05/2024].

GARCÍA-FERNÁNDEZ, Miguel (2017): «Las últimas voluntades como expresión de la voz femenina en la Edad Media. Dos nuevas aportaciones al "Corpus testamentario de la Galicia medieval"», en Manuel Cabrera Espinosa y Juan Antonio López Cordero (eds.): *IX Congreso virtual sobre Historia de las Mujeres (15 al 31 de octubre de 2017). Comunicaciones*. Jaén: Asociación de Amigos del Archivo Histórico Diocesano de Jaén, pp. 233-284. https://corpora.uah.es/egpa/publicaciones/ultimasvoluntades.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel (2018a): *As espontáneas. Historias de San Xoán de Río I*. [San Xoán de Río]: Asociación Cultural RioMola / Concello de San Xoán de Río. https://culturagalega.gal/album/detalleextra.php?id=5264 [Consulta: 13/11/2024].

GARCÍA-FERNÁNDEZ, Miguel (2018b): «As derradeiras vontades dos señores de Castroverde: edición de «novos» documentos para o "Corpus testamentario da Galicia medieval"», *Lucensia. Miscelánea de cultura e investigación*, XXIX, 57, pp. 197-218. https://corpora.uah.es/egpa/publicaciones/asderradeiras.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel (2018-19): «Los expedientes de las "Espontáneas" de San Xoán de Río. Nuevas fuentes para la historia de las mujeres gallegas del siglo XIX", *Boletín Avriense*, 48-49, pp. 287-314. http://consellodacultura.gal/mediateca/extras/CCG_ig_album_Espontaneas_sobre_003.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel (2019): «Testamentos femeninos para el estudio de la realidad señorial gallega a finales de la Edad Media: una aproximación comparada a las últimas voluntades de Guiomar Méndez de Ambía (1484) y doña Isabel González Noguerol (1527-1533)», en Manuel Cabrera Espinosa y Juan Antonio López Cordero (eds.): *XI*

Congreso virtual sobre Historia de las Mujeres (15 al 31 de octubre de 2019). Comunicaciones. Jaén: Asociación de Amigos del Archivo Histórico Diocesano de Jaén, pp. 279-330. https://corpora.uah.es/egpa/publicaciones/testamentosfemeninos.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel (2025): «Da cantiga aos documentos: dona Milia Íñiguez de Mendoza nos tempos do trobadorismo galego-portugués», en Ricardo Pichel (ed.): *«Contarte he maravillas...» Estudios hispánicos dedicados a Joseph T. Snow*. *Vol. 2 - El patrimonio literario de Alfonso X*. Berlín: Peter Lang (en prensa).

GARCÍA-FERNÁNDEZ, Miguel; Diana PELAZ FLORES y Ricardo PICHEL (2020): «Galicia e El-Rei ou como reinar desde a distancia: comunicación política arredor de dous novos privilexios rodados de Xoán II», *Madrygal. Revista de Estudios Gallegos*, 23, pp. 139-180. https://doi.org/10.5209/madr.73069 [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Pablo S. OTERO PIÑEYRO MASEDA (2022): «Cinco «espontáneas» lucenses: achegas á muller, sexualidade e violencia na Galicia interior (s. XVIII)», *Murguía. Revista Galega de Historia*, 45-46, pp. 57-77. https://corpora.uah.es/egpa/publicaciones/cincoespontaneas.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2020): «Novas achegas documentais para o estudo da realidade monástica da Ribeira Sacra: tres pergamiños inéditos de Santa Cristina de Ribas de Sil (ss. XV-XVI)», *Murguía. Revista Galega de Historia*, 41-42, pp. 13-41. https://corpora.uah.es/egpa/publicaciones/novasachegas.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2023a): «A documentación de San Salvador de Sobrado de Trives máis alá do ASPA: achega para unha nova edición da súa colección documental», en *Desde San Paio de Antealtares. Historia, patrimonio e vida monástica. Estudos dedicados a sor Mercedes, arquiveira de San Paio, e á Comunidade de Antealtares*. Santiago de Compostela: Consorcio de Santiago / Alvarellos Editora, pp. 89-121. https://corpora.uah.es/egpa/publicaciones/sansalvadordesobrado.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2023b): «O final do casal Limia-Seixas: a sentenza de anulación matrimonial de dona Xoana Vázquez das Seixas e Fernán Eanes de Limia (1398)», *Madrygal. Revista de Estudios Gallegos*, 26, e94791. https://doi.org/10.5209/madr.94791 [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2024a): «Tres "novos" pergamiños de Afonso VII referidos a Galiza», *Estudios Mindonienses*, 37, pp. 929-943. https://corpora.uah.es/egpa/publicaciones/tresnovospergaminos.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2024b): «Novos testemuños para a colección documental de Santa María de Montederramo», *Madrygal. Revista de Estudios Gallegos*, 27, e104943. https://doi.org/10.5209/madr.104943 [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2024-25): «Os pergamiños de Manuel Murguía (I). Documentación medieval de Santa María de Sobrado», *Murguía. Revista Galega de Historia*, 50-51, pp. 13-36. https://corpora.uah.es/egpa/publicaciones/manuelmurguia.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2025a): «O Corpus de Espontáneas de Galicia (CORESGAL). Revalorización de un tipo de documental singular a través de las Humanidades digitales», en II Xeira CLARIAH-GAL (Santiago de Compostela, 9 de mayo de 2025). https://ilg.usc.gal/gl/actividades/II-xeira-clariah-gal [Consulta: 17/09/2025]

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2025b), «Fontes documentais da Galicia medieval para o profesorado en formación: materiais para a inclusión da perspectiva de xénero na Educación Secundaria a partir do EGPA», en Uxío-Breogán Diéguez Cequiel y Miguel García-Fernández (eds.): *Ciencias Sociais. Investigación, fontes documentais, didáctica e*

recursos dixitais. Unha mirada desde a lexislación, a historia, a memoria histórica, a socioloxía, a antropoloxía e a didáctica das ciencias sociais. Santiago de Compostela: Instituto Galego de Historia, pp. 29-67. https://corpora.uah.es/egpa/publicaciones/fontesprofesorado.pdf [Consulta: 17/09/2025].

GARCÍA-FERNÁNDEZ, Miguel y Ricardo PICHEL (2025c): *GALLÆCIÆ SCRIPTORES. O Rexistro de escrituras do notario Vasco Gómez Varela (1448-1461)*. Madrid: Sílex (en preparación).

GARRIDO MARTÍN, Blanca y Leyre MARTÍN AIZPURU (2024): «Filología en Teitok: la edición digital y algunas muestras de escritura epistolar», en María García Antuña (coord.): *Filología y nuevas tecnologías*. Sevilla: Universidad de Sevilla, pp. 91-98.

GIMÉNEZ-EGUÍBAR, Patricia y Ricardo PICHEL (2022): «"Se acordará de escribir a los pobres desterrados en este valle de dullness". Prácticas del translenguar en la correspondencia privada de María Amparo Ruiz de Burton», en Belén Almeida, Ricardo Pichel y Delfina Vázquez Balonga (eds.), *Escritura en mano de mujeres en el ámbito hispánico de la Edad Media a la Modernidad*. Madrid: Sílex, pp. 405-429.

GIMÉNEZ-EGUÍBAR, Patricia y Ricardo PICHEL (2023): «Cartas desde la California recién anexionada: rasgos lingüísticos de la correspondencia privada de María Amparo Ruiz de Burton en los fondos de la Huntington Library (1852-1857)», *Revista internacional de lingüística iberoamericana*, 41 (Sección temática. Escritura femenina en el ámbito hispánico: enfoques para su estudio lingüístico y textual II), pp. 89-108. https://doi.org/10.31819/rili-2023-214107 [Consulta 15/07/2024].

GÓMEZ SEIBANE, Sara (2013): «Documentos guipuzcoanos 2. Cartas privadas y familiares», en Carmen Isasi Martínez y José Luis Ramírez Luengo (eds.): *Una muestra documental del castellano norteño en el siglo XIX*. Lugo: Axac, pp. 143-176.

HORCAJADA DIEZMA, Bautista (2012): «De la transcripción paleográfica a la presentación crítica. Automatización del proceso», en María Jesús Torrens Álvarez y Pedro Sánchez-Prieto Borja (eds.): *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Berna: Peter Lang, pp. 427-445.

IRACEBURU JIMÉNEZ, Maite, Concepción MARTÍNEZ PASAMAR y Cristina TABERNERO SALA (2018): «Presentación del proyecto de investigación del grupo TesUN (Universidad de Navarra)», *Chimera: Romance Corpora and Linguistic Studies*, 5 (2), pp. 321-327. https://doi.org/10.15366/chimera2018.5.2.011 [Consulta: 17/09/2025].

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: Universidad de Sevilla.

JANSSEN, Maarten (2012): «NeoTag: A POS Tagger for Grammatical Neologism Detection», en *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*. Estambul: European Language Resources Association (ELRA), pp. 2118-2124. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1098_Paper.pdf [Consulta: 17/09/2025].

JANSSEN, Maarten (2014): *TEITOK – a Tokenized TEI environment*. http://www.teitok.org/ [Consulta: 17/09/2025].

JANSSEN, Maarten (2016): «TEITOK: Text-Faithful Annotated Corpora», en *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož: European Language Resources Association (ELRA), pp. 4037-4043. https://aclanthology.org/L16-1637.pdf [Consulta: 17/09/2025].

JANSSEN, Maarten y Gael VAAMONDE (2020): «Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK», en Rosario Álvarez Blanco y Ernesto Xosé González Seoane (eds.): *Calen barbas, falen cartas: A escrita en galego na Idade Moderna*. Santiago de Compostela: Consello da Cultura Galega, pp. 271-292.

Letradas = ALMEIDA CABREJAS, Belén (coord.): *Letradas. Corpus de textos escritos por mujeres en España (1400-1900)*. https://corpora.uah.es/letradas/ [Consulta: 17/09/2025].

LLOBERES, Marina, Irene CASTELLÓN y Lluis PADRÓ (2010): «Spanish FreeLing Dependency Grammar», en *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta: European Language Resources Association (ELRA), pp. 693-699. http://www.lrec-conf.org/proceedings/lrec2010/pdf/562_Paper.pdf [Consulta: 17/09/2025].

Lozana Digital = DÍAZ BRAVO, Rocío y Gael VAAMONDE (dirs.): *LD. Lozana Digital*. http://corpora.ugr.es/lozana [Consulta: 17/09/2025].

MARCET RODRÍGUEZ, Vicente J. y María de las Nieves SÁNCHEZ GONZÁLEZ DE HERRERO (2016): «La documentación medieval de Miranda de Ebro: Presentación del corpus y rasgos lingüísticos», en Johannes Kabatek (ed.), *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín, Boston: De Gruyter, pp. 157-174. https://doi.org/10.1515/9783110462357-008 [Consulta: 17/09/2025].

MARTÍN AIZPURU, Leyre (2016): «Algunos recursos informáticos al servicio de la edición de textos: la edición en XML-TEI», en Chiara Albertin y Santiago del Rey Quesada (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 y María de las Nieves SÁNCHEZ GONZÁLEZ DE HERRERO (2019): «El estudio de la documentación alfonsí: un proyecto abierto», en Déborah González y Helena Bermúdez Sabel: *Humanidades digitales: miradas hacia la Edad Media*. Berlín: De Gruyter, pp. 111-130. https://doi.org/10.1515/9783110585421-009 [Consulta: 17/09/2025].

MELE = FERREIRA, Maria do Rosário y José Carlos Ribeiro MIRANDA (dirs.): *Da memória escrita à leitura do espaço: Pedro de Barcelos e a identidade cultural do Norte de Portugal*. Integrado en *Escritorio Galego-Portugués Antigo*. https://corpora.uah.es/egpa/mele [Consulta: 17/09/2025].

MIGUEL FRANCO, Ruth y Pedro SÁNCHEZ-PRIETO BORJA (2016): «CODEA: A "Primary" Corpus of Spanish Historical Documents», *Variants*, 12-13, pp. 211-230. https://doi.org/10.4000/variants.364 [Consulta: 17/09/2025].

MORAL DEL HOYO, M.ª Carmen (2019): «Volver a (y revolver) los orígenes del castellano: el Corpus Histórico del Español Norteño (CORHEN)», en Mónica Castillo Lluch y Elena Diez del Corral Areta (eds.): *Reescribiendo la historia de la lengua española a partir de la edición de documentos*. Bern: Peter Lang, pp. 361-390.

MORENO MORENO, María Águeda (2018): «Subcorpus documental administrativo del antiguo concejo de Baeza (Jaén): el corpus COHSANRE», *Scriptum digital*, 7, pp. 67-94. https://raco.cat/index.php/scriptumdigital/article/view/343466 [Consulta: 17/09/2025].

NOSTOI = PICHEL, Ricardo y Clara PASCUAL-ARGENTE (dirs.): *NOSTOI: Corpus de textos troyanos ibéricos (ss. XIII-XVI)*. https://corpora.uah.es/nostoi/ [Consulta: 17/09/2025].

ODE = CALDERÓN CAMPOS, Miguel y María Teresa GARCÍA-GODOY (2010-2019): *Oralia Diacrónica del Español*. http://corpora.ugr.es/ode/ [Consulta: 17/09/2025].

OLAIA, Inês (2023): «What's in a signature? Assessing the use of the royal signature by the Queens of Portugal in the late Middle Ages», *RiMe. Rivista dell'Istituto di Storia dell'Europa Mediterranea*, 12/I n.s. (número especial), pp. 91-114. https://doi.org/10.7410/1607 [Consulta 15/07/2024].

OSPINA GIRALDO, Liliana Estefanía y María Claudia GONZÁLEZ-RÁTIVA (2021): «Los corpus SEAH», en María Claudia González-Rátiva y Liliana Estefanía Ospina Giraldo (coords.): *El español tardocolonial en Antioquia (1701-1816). Corpus y análisis en documentación histórica*. Antioquia: Universidad de Antioquia, pp. 21-34.

Panépica Digital = RODRÍGUEZ MOLINA, Javier y Gael VAAMONDE (dirs.): *Panépica Digital: Corpus de la primitiva épica hispánica*. http://corpora.ugr.es/cid [Consulta: 17/09/2025].

PASCUAL-ARGENTE, Clara y Ricardo PICHEL (2024): «NOSTOI: Un entorno digital colaborativo para la edición y estudio de los textos troyanos ibéricos», *Medievalia*, 27 (1), pp. 353-388. https://doi.org/10.5565/rev/medievalia.663 [Consulta 15/09/2024].

PICHEL, Ricardo y Francisco Xavier VARELA BARREIRO (2014): «Edición de textos da Galiza medieval e moderna. Algúns proxectos en marcha», en Leticia Eirín y Xoán López-Viñas (eds.): *Lingua, texto, diacronía: estudos de lingüística histórica*. A Coruña: Universidade da Coruña, Departamento de Galego-Portugués, Francés e Lingüística, pp. 291-318.

PICHEL, Ricardo y Miguel GARCÍA-FERNÁNDEZ (2024): «O Escritorio Galego-Portugués Antigo (EGPA)», en I Xeira CLARIAH-GAL (Santiago de Compostela, 2 de mayo de 2024). https://ilg.usc.gal/sites/default/files/poster_egpa_final_ilg.pdf [Consulta: 17/09/2025].

PICHEL, Ricardo; Miguel GARCÍA-FERNÁNDEZ y Víctor CABALLERO GÓMEZ (2025): «O Escritorio Galego-Portugués Antigo (EGPA): unha nova ferramenta colaborativa de edición dixital en acceso aberto», *Madrygal. Revista de Estudios Gallegos*, 28 (en prensa).

PONS RODRÍGUEZ, Lola, Eva BRAVO GARCÍA, Blanca GARRIDO MARTÍN y Álvaro S. OCTAVIO DE TOLEDO Y HUERTA (2014): «La edición de textos de quejas: propuestas preliminares en torno a un corpus histórico discursivo», *Scriptum digital*, 3, pp. 183-200. https://raco.cat/index.php/scriptumdigital/article/view/316399 [Consulta: 17/09/2025].

P.S. POST SCRIPTUM = CLUL (ed.) (2014): *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. http://teitok.clul.ul.pt/postscriptum/ [Consulta: 17/09/2025].

RED CHARTA (2013): *Criterios de edición de documentos hispánicos (orígenes-siglo XIX) de la Red Internacional CHARTA*. https://corpora.uah.es/index.php?action=recursos [Consulta: 17/09/2025].

RODRÍGUEZ MOLINA, Javier y Gael VAAMONDE (2025): «Panépica Digital: Integrando marcación textual y anotación lingüística para el estudio de la épica hispánica medieval», en Mario Cossío Olavide, José Manuel Fradejas Rueda y Ricardo Pichel (eds.): *Filología digital hispánica. Aplicaciones a la lengua y literatura medieval*. Berlín: Walter de Gruyter (en prensa).

ROMERO, Verónica, Alejandro Héctor TOSELLI y Enrique VIDAL (2012): *Multimodal Interactive Handwritten Text Transcription*. New Jersey: World Scientific.

SÁNCHEZ GONZÁLEZ DE HERRERO, María de las Nieves, Juan Pedro SÁNCHEZ MÉNDEZ, Ingmar SÖHRMAN y María Jesús TORRENS ÁLVAREZ (2013): «La Red CHARTA: objetivos y método», en Emili Casanova Herrero y Cesáreo Calvo Rigual (eds.): *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas*, vol. VII. Berlín: De Gruyter, pp. 263-274.

SÁNCHEZ SIERRA, Diego (2021): «Aproximación a la configuración léxica del español en el suroeste de los Estados Unidos (1733-1900)», *Cuadernos del Instituto de Historia de la Lengua*, 14, pp. 147-185.

SÁNCHEZ-PRIETO BORJA, Pedro (1998): *Cómo editar los textos medievales: criterios para su presentación gráfica*. Madrid: Arco.

SÁNCHEZ-PRIETO BORJA, Pedro (2009a): «Hacia un estándar en la edición de las fuentes documentales», en Cristina Castillo Martínez y José Luis Ramírez Luengo (eds.): *Lecturas y textos en el siglo XXI. Nuevos caminos en la edición textual*. Lugo: Axac, pp. 125-143.

SÁNCHEZ-PRIETO BORJA, Pedro (2009b): «El Corpus de Documentos Españoles Anteriores a 1700 (CODEA)», en Andrés Enrique-Arias (ed.): *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*, Madrid-Frankfurt am Main: Iberoamericana-Veuvert, pp. 25-38. https://doi.org/10.31819/9783865278685-003 [Consulta: 17/09/2025].

SÁNCHEZ-PRIETO BORJA, Pedro (2011): *La edición de textos españoles medievales y clásicos: criterios de presentación gráfica*. San Millán de la Cogolla: Cilengua, Fundación San Millán de la Cogolla.

SÁNCHEZ-PRIETO BORJA, Pedro (2012a): «La red CHARTA: proyecto global de edición de documentos hispánicos», en María Jesús Torrens Álvarez y Pedro Sánchez-Prieto Borja (eds.): *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Berna: Peter Lang, pp. 17-44.

SÁNCHEZ-PRIETO BORJA, Pedro (2012b): «Desarrollo y explotación del "Corpus de Documentos Españoles Anteriores a 1700" (CODEA)», *Scriptum digital*, 1, pp. 5-35. https://raco.cat/index.php/scriptumdigital/article/view/316410 [Consulta: 17/09/2025].

SÁNCHEZ-PRIETO BORJA, Pedro (2018): «El corpus ALDICAM-CM: geografía lingüística diacrónica de la Comunidad de Madrid», *Chimera: Romance Corpora and Linguistic Studies*, 5 (1), pp. 69-75. https://doi.org/10.15366/chimera2018.5.1.004 [Consulta: 17/09/2025].

SÁNCHEZ-PRIETO BORJA, Pedro y Hiroto UEDA (2018): *Inventario léxico del corpus CODEA (Castilla la Vieja)*. https://h-ueda.sakura.ne.jp/lyneal/ilc-cv.htm [Consulta: 17/09/2025].

SEAH = SEMILLERO ESPAÑOL HISTÓRICO DE ANTIOQUIA: *Corpus SEAH* [en fase de desarrollo].

SPENCE, Paul (2014): «Siete retos en edición digital para las fuentes documentales», *Scriptum digital*, 3, pp. 153-181. https://raco.cat/index.php/scriptumdigital/article/view/316398 [Consulta: 17/09/2025].

SPENCE, Paul, Carmen ISASI, Elena PIERAZZO e Irene VICENTE (2012): «Cruzando la brecha: marcación digital con criterios filológicos», en María Jesús Torrens Álvarez y Pedro Sánchez-Prieto Borja (eds.): *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Berna: Peter Lang, pp. 465-483.

TABERNERO SALA, Cristina (2020): «Contribución al estudio del español norteño a partir de un corpus de declaraciones en procesos judiciales (siglos XVI-XIX)», *Scriptum digital*, 9, pp. 87-115. https://raco.cat/index.php/scriptumdigital/article/view/377289 [Consulta: 17/09/2025].

TORRENS ÁLVAREZ, María Jesús (2019): «El hibridismo latinorromance de fueros y documentos de finales del s. XII y comienzos del XIII», en Diana Esteba Ramos *et al*. (eds.): Quan sabias e quam maestras: *disquisiciones de lengua española*. Málaga: Universidad de Málaga, pp. 101-112.

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

TORRUELLA, Joan (2020): «Un corpus documental para el estudio de las interferencias lingüísticas entre el español y el catalán», en Rosario Álvarez Blanco y Ernesto Xosé González Seoane (eds.): *Calen barbas, falen cartas: A escrita en galego na Idade Moderna*. Santiago de Compostela: Consello da Cultura Galega, pp. 225-252.

TORRUELLA, Joan y Gloria CLAVERÍA (2019): «Corpus de documentos castellanos redactados en Cataluña (siglos XVIII y XIX): inicio de un proyecto», en Mónica Castillo Lluch y Elena Diez del Corral Areta (eds.): *Reescribiendo la historia de la lengua española a partir de la edición de documentos*. Bern: Peter Lang, pp. 43-60.

VAAMONDE, Gael (2015): «P. S. Post Scriptum: dos corpus diacrónicos de escritura cotidiana», *Procesamiento del Lenguaje Natural*, 55, pp. 57-64.

VAAMONDE, Gael (2018): «Escritura epistolar, edición digital y anotación de corpus», *Cuadernos del Instituto de Historia de la Lengua*, 11, pp. 139-164.

VILLEGAS, Mauricio, Joan Andreu SÁNCHEZ y Enrique VIDAL (2015): «Optical Modelling and Language Modelling Trade-off for Handwritten Text Recognition», en *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. Túnez: IEEE, pp. 831-835. https://doi.org/10.1109/ICDAR.2015.7333878.

ZavalDiCor = MARY TROJANI, Cécil (dir.): *Corpus digital de correspondencias de la familia Zavala* [en fase de desarrollo].