PHILOLOGIA HISPALENSIS

ESTUDIOS LITERARIOS

2025 | VOL. XXXIX 2

PHILOLOGIA HISPALENSIS

AÑO 2025
VOL. XXXIX/2

ESTUDIOS LITERARIOS

FACULTAD DE FILOLOGÍA
UNIVERSIDAD DE SEVILLA

Evaluación de originales: Los originales se someten a una evaluación ciega, un proceso anónimo de revisión por pares, siendo enviados a evaluadores externos y también examinados por los miembros del Consejo de Redacción y/o los especialistas del Consejo Asesor de la Revista.

Periodicidad: Anual en formato tradicional y en formato electrónico.

Publicación en Internet: <https://editorial.us.es/es/revistas/philologia-hispalensis>, <https://revistascientificas.us.es/index.php/PH>.

Bases de datos: *Philologia Hispalensis* se encuentra indexada en CARHUS Plus+2018, CIRC (grupo A), DIALNET, DOAJ, Dulcinea, Index Islamicus, Latindex 2.0 (100% de los criterios cumplidos), MIAR (ICDS 2022 = 10), MLA, REDIB, SCOPUS, ERIHPLUS y ANVUR (Clase A). Asimismo, cuenta con el sello de calidad de la FECYT (8ª edición, 2023, renovado en 2024 y válido hasta 2025) en los campos de conocimiento Lingüística y Literatura dentro de la modalidad Humanidades.

Envío de originales y suscripciones: Las colaboraciones deben enviarse a través de <https://revistascientificas.us.es/index.php/PH>.

Dirección de contacto: Secretariado de la Revista Philologia Hispalensis, Facultad de Filología, Universidad de Sevilla, C/ Palos de la Frontera, s/n, 41004 Sevilla; o bien al correo electrónico <philhisp@us.es>.

Intercambios o canjes (Bibliotecas Universiarias): Solicítense a Editorial Universidad de Sevilla o al Secretariado de la revista <philhisp@us.es>.

Financiación: Revista financiada por la Universidad de Sevilla dentro de las ayudas del VII PPIT-US y del Decanato de la Facultad de Filología.
Portada: referencias.maquetacion@gmail.com
Depósito Legal: SE-354-1986
ISSN: 1132 - 0265 / eISSN 2253-8321
Maquetación: referencias.maquetacion@gmail.com
Imprime: Podiprint
Distribuye: Editorial Universidad de Sevilla, Porvenir, 27, 41013 Sevilla

Revisores del volumen 39, número 2 (2025). Estudios Literarios

Han actuado como revisores anónimos para uno o más artículos de este número, tanto los aceptados como los rechazados, los siguientes investigadores:

Alvite Díez, María Luisa (Universidad de León)
Calarco, Gabriel Alejandro (CONICET, Argentina)
Camprubí Vinyals, Adriana (Universitat de Barcelona)
de Mora Valcárcel, Carmen (Universidad de Sevilla)
Fernández Riva, Gustavo (Heidelberg Universität, Alemania)
Gatica Cote, Paulo Antonio (Universidad Complutense de Madrid)
González Martínez, Déborah (Universidade de Santiago de Compostela)
González Miranda, Emilio (Universidade de Santiago de Compostela)
González Montes, Antonio (Universidad Nacional Mayor de San Marcos / Academia Peruana de la Lengua, Perú)
González Sanz, Marina (Universidad de Sevilla)
Martí Junquera, Sadurní (Universitat de Girona)
Martín Junquera, Imelda (Universidad de León)
Moreno Lago, Eva María (Universidad de Sevilla)
Morrás Ruiz-Falcó, María (Universitat Pompeu Fabra)
Pérez Agustín, María Mercedes (Universidad Complutense de Madrid)
Pérez Ben, Lorena (Universidade de Santiago de Compostela)
Plata Parga, Fernando (Colgate University, EEUU)
Roeder, Torsten (German National Academy Leopoldina, Alemania)
Sánchez Dueñas, Blas (Universidad de Córdoba)
Simó Torres, Meritxell (Universitat de Barcelona)
Vaamonde Dos Santos, Gael (Universidad de Granada)
Vargas Díaz-Toledo, Aurelio (Universidad Complutense de Madrid)

ÍNDICE

Reseñas de libros

ESTUDIOS LITERARIOS

UN MODELO HTR PARA INCUNABLES CASTELLANOS

AN HTR MODEL FOR SPANISH INCUNABULA

José Manuel Fradejas Rueda
Universidad de Valladolid
josemanuel.fradejas@uva.es
0000-0001-8603-6765

Mario Cossío Olavide
Université de Bretagne Occidentale
mario.cossioolavide@univ-brest.fr
0000-0002-1447-389

Recibido: 10-12-2024 | Aceptado: 18-02-2025

Cómo citar: Fradejas Rueda, J. M. y Cossío Olavide, M. (2025). Un modelo HTR para incunables castellanos. *Philologia Hispalensis*, *39*(2), 131-177. https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Resumen

Este artículo estudia la aplicación de modelos de reconocimiento automático de texto (HTR) a incunables castellanos. En la primera sección, realizamos un repaso metodológico sobre las características y capacidades actuales de las plataformas de HTR disponibles, acompañado de una discusión metodológica sobre los distintos sistemas de transcripción disponibles y una explicación del flujo de trabajo para entrenar un modelo HTR en la plataforma Transkribus. En la segunda parte, describimos el entrenamiento y validación del modelo HTR Spanish Gothic Incunabula (HSMS), desarrollado para transcribir incunables castellanos con una tasa de error inferior al 1%.

Palabras clave: incunables españoles, reconocimiento automático de texto (HTR), facsímil digital, Transkribus, edición digital, humanidades digitales.

Abstract

This article examines the application of Handwritten Text Recognition (HTR) models to Castilian incunabula. The first section provides a methodological review of the features and current capabilities of available HTR platforms, along with a discussion of various transcription systems and an explanation of the workflow for training an HTR model using Transkribus. The second part describes the training and validation of the Spanish Gothic

Incunabula (HSMS) HTR model, developed to transcribe Castilian incunabula with an error rate of less than 1%.

Keywords: Spanish incunabula, Handwritten Text Recognition, digital facsimile, Transkribus, digital edition, digital humanities.

1. Introducción

En la segunda mitad de 1472, con toda probabilidad en Segovia, salió del taller de Juan Párix de Heidelberg, un impresor de origen alemán que parece haber desarrollado su arte en Roma, el primer libro de letra de molde de la lengua castellana: el *Sinodal de Aguilafuente* (Reyes Gómez, 2004). No solo fue el primer libro impreso en castellano, también fue el primer libro impreso en la península ibérica. A partir de ese momento, la industria editorial hispánica no se detendrá, dando inicio a un fenómeno que se extiende a toda la península ibérica (y a todas sus lenguas, incluyendo numerosas ediciones en latín).

Los libros que se imprimieron a lo largo del siglo xv, hasta el 31 de diciembre de 1500 son llamados *incunables*. Posteriormente, los especialistas, al tener en cuenta que durante las dos primeras décadas del quinientos no hay innovación técnica digna de mención, hablan del periodo *postincunable*, que se alarga hasta 1520.

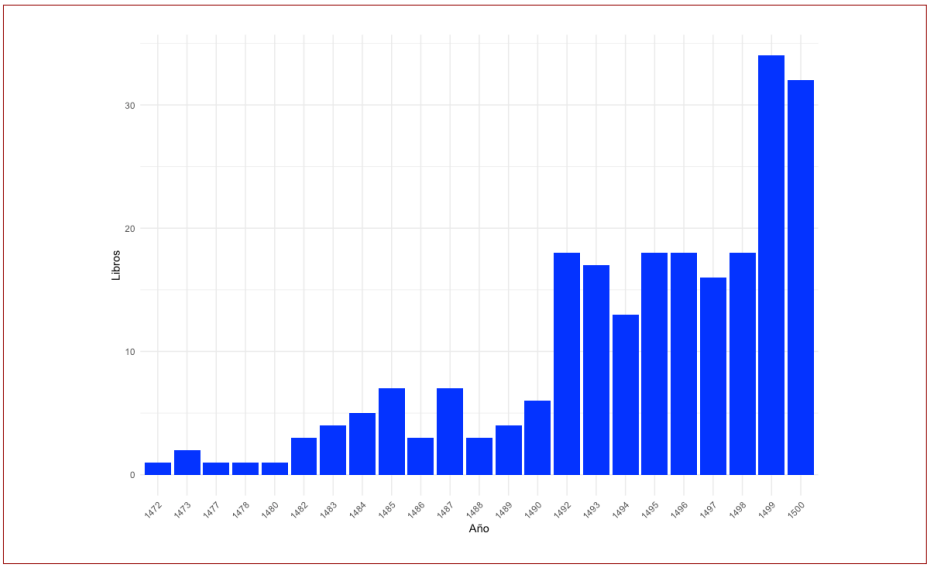
De estos 232 libros impresos, actualmente solo una mínima parte son accesibles en formato electrónico. No nos referimos a que sean accesibles sus reproducciones digitales (*digital copies*), copias de baja resolución (alrededor de 150 dpi), accesibles en línea y descargables en formato PDF; o incluso, a las copias de alta resolución no descargables de algunas colecciones, como las Reales Bibliotecas o la Biblioteca de Catalunya, que pueden ser minadas y descargadas en lote (*batch download*) empleando *scripts* diseñados para sortear las restricciones impuestas por los repositorios. Nos referimos a la existencia de facsímiles digitales (*digital facsimiles*) como los entienden Donaldson (1997), Ciula (2009) y Fafinski (2022): sustitutos digitales (*digital surrogates*) que contienen una representación multinivel (*multilayered*) de un documento histórico (páginas, espacios, folios, cuadernillos, pero también información sobre el pergamino y el papel, las filigranas, la encuadernación, la manuscritura o tipografía, los propietarios y la historia del documento), incluyendo su contenido textual, codificado usando un sistema que permita hacerlo legible por ordenadores (*machine readeable*)[1]. Es cierto que algunas bibliotecas digitales,

[1] La precisión ofrecida por esta nueva terminología, desarrollada y empleada fundamentalmente en las filologías inglesas y germánicas, nos ha llevado a replantearnos la validez de seguir llamando *facsímil digital*/*fac-similé numérique*/*facsimile digitale* a lo que no es sino una *reproducción digital* de un texto, un uso aún muy extendido en las filologías románicas (en las hispánicas, véase BETA (1997); en la francesa, Camps (2021); y en la italiana, Pierazzo (2015: 93-98) y Mancinelli y Pierazzo (2020: 53-56)). No abandonamos el uso, de mayor aceptación, de los términos *facsímil*, *facsimilar* para designar las ediciones que reproducen mecánicamente (físicamente) un manuscrito o una edición antigua

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Figura 1

Número de libro impresos en castellano durante el periodo incunable



Nota. Fuente: Elaboración propia con información recolectada de PhiloBiblon.

como el proyecto *Biblioteca Digital Hispánica* (BDH) de la Biblioteca Nacional de España (BNE), ofrecen reproducciones digitales de un elevadísimo número de incunables, y que muchos de ellos han sido objeto de un proceso muy rudimentario de reconocimiento automático del texto (ATR), pero el resultado es un texto ilegible (Figura 4) e incompleto desde la perspectiva que hemos descrito.

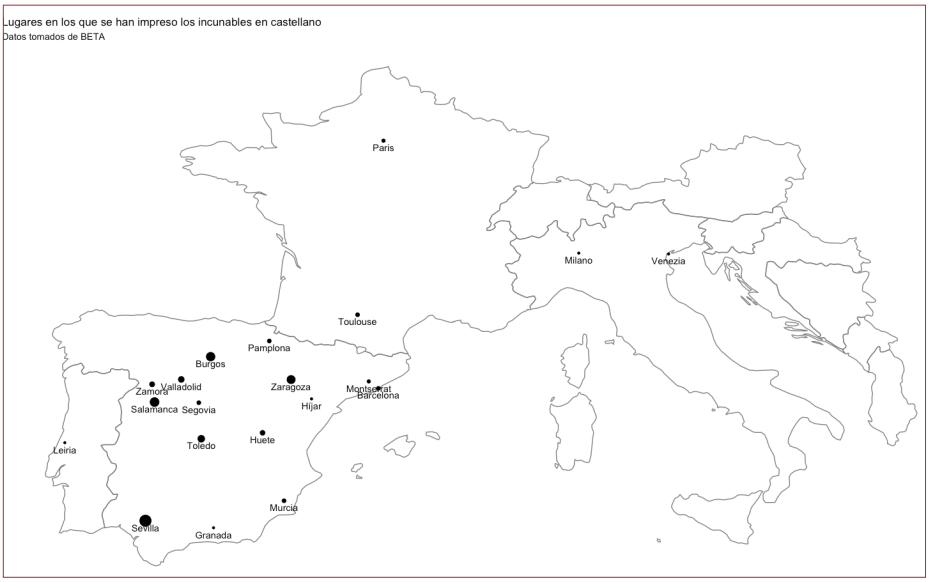
Según la información recopilada por la *Bibliografía Española de Textos Antiguos* (BETA) del proyecto *PhiloBiblon*, durante el periodo incunable se imprimieron 232 libros, cuya lengua es mayoritariamente el castellano (Figura 1). Pero no todos estos libros salieron de talleres radicados en los reinos peninsulares (Figura 2), pues varios de ellos vieron la luz en imprentas establecidas en Europa: Venecia (1), Milán (1), París (2) y Toulouse (3). Los grandes centros de producción peninsulares fueron Sevilla, Salamanca y Burgos, en la corona de Castilla, y Zaragoza, en el reino de Aragón (Figura 3).

(*edición facsimilar del manuscrito…*, *reproducción facsimilar de la* editio princeps…), ofreciendo un objeto físico que intenta imitar las características del original; tampoco renunciamos a su uso para designar el tipo de transcripción que reproduce lo más fielmente posible el texto de un manuscrito o impreso (*transcripción facsimilar*).

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Figura 2

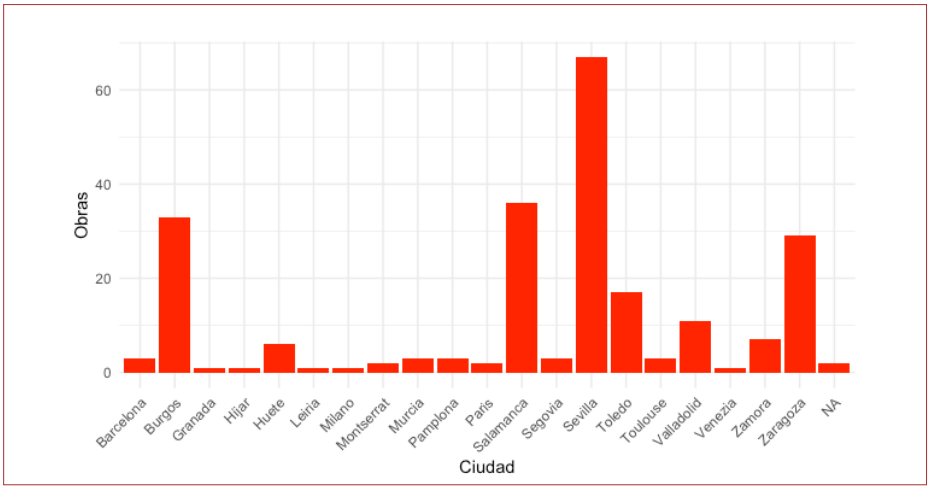
Ciudades impresoras de libros en castellano durante el periodo incunable



Nota. Fuente: Elaboración propia con información recolectada de PhiloBiblon.

Figura 3

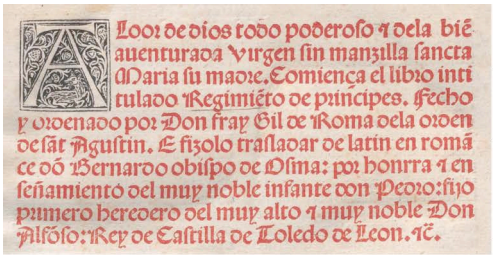
Número de libros en castellano impresos en cada ciudad durante el periodo incunable



Nota. Fuente: Elaboración propia con información recolectada de PhiloBiblon.

Figura 4

Íncipit del Regimiento de príncipes de Egidio Romano (RGP), f. 1r, y versión OCR



Loor de dios todo poderoso ⁊ dela bie
auenturada Virgen ſin manzilla ſancta
Maria ſu madre. Comiença el libro inti
tulado Regimiẽto de principes. Fecho
y ordenado por Don fray Gil de Roma dela orden
de ſãt Agustin. E fizolo traſladar de latin en roma
ce oõ Bernardo obiſpo de Oſina: por honrra ⁊ en
ſeñamiento del muy noble infante don Pedro: fijo
primero heredero del muy alto ⁊ muy noble Don
Alfõſo: Rey de Caſtilla de Toledo de Leon. ⁊c.

5S| Eoo: OC OÍOS rooo poôerofo ^ ocU bié ^O^
aucnmraoa Virgen fin mantilla fanera ^M| fiOaria
fu maoicXomienca el libio inti «all tnlaoo
iRegimiéto oc piincipcs, ^eebo r oioenaoo poi
2)on frar ©il oe iKonta ocla oioen oe lar ^guftin, £
fí$olo trafiaoar oe latín en roma ce oó ïSernaroo
obífpo oe Ófina: poi bonrra 4 en fenamíento oel
mnr noble infante oon peoioifíp pnmero
bereoero oel muf alto 4 mup noble 2>on
fllfófonRcr oe Caftilla oe Xoledo oe Eeon.4c\

Nota. Los códigos alfanuméricos de tres caracteres indican los impresos usados para entrenar el modelo HTR de incunables, cuyas referencias precisas aparecen en las tablas 7 y 8 del anexo 1. Fuente: *Biblioteca Digital Hispánica*, Biblioteca Nacional de España.

Algunas instituciones, como la misma BNE, han intentado solventar este problema a través de programas de colaboración ciudadana (*crowdsourcing*). La BNE lanzó *BNE Comunidad*, una iniciativa compuesta de «retos» para abordar las transcripciones, entre otros textos, de impresos antiguos[2]. El proyecto comenzó con el pliego «Aquí comienzan dos romances…»[3]; y ha continuado con una serie de cinco pliegos poéticos[4] y con la *Égloga interlocutoria graciosa y por gentil estilo nuevamente trovada* atribuida a Diego de Ávila e impresa por Estanislao Polono entre 1502 y 1504[5]. Recientemente, el programa ha finalizado la transcripción de la edición de 1542 del *Jardín de nobles donzellas* de fray Martín de Córdoba (R/9717)[6]. En la mayoría de los casos, estas tareas colaborativas tienen como objetivo ofrecer ediciones legibles de los textos elegidos, para lo que ofrecen unas muy simples normas de transcripción (Figura 5)[7].

[2] Otros retos incluyen la corrección del OCR de periódicos decimonónicos como *El Español* (Londres), *La luz del porvenir* o el *Periódico de damas*.

[3] https://comunidad.bne.es/proyectos/transcripcion-aqui-comienzan-dos-romances/

[4] https://comunidad.bne.es/proyectos/donde-se-siguen-los-romances/

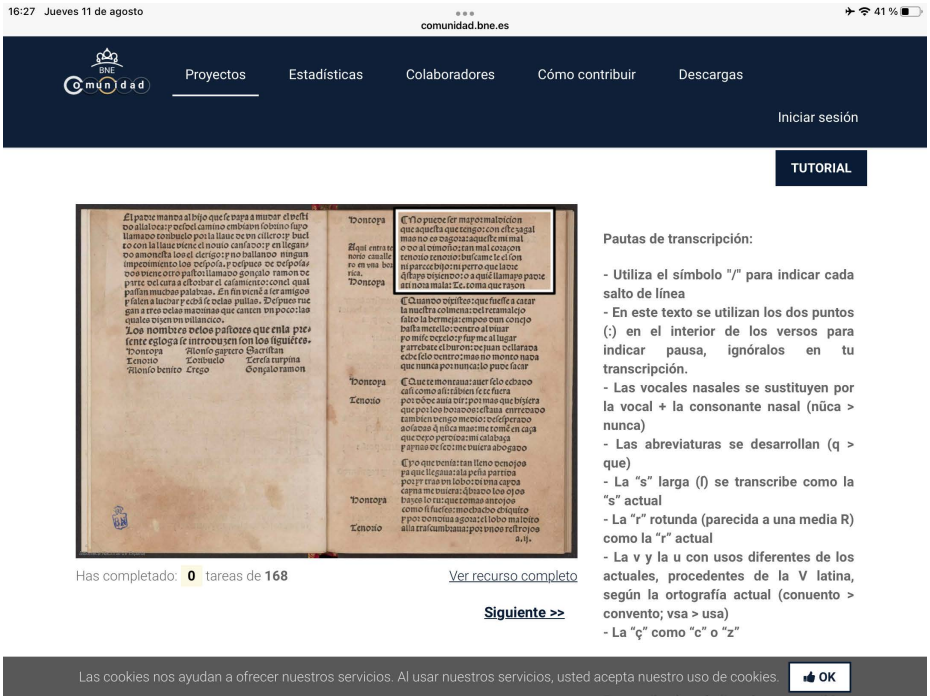
[5] https://comunidad.bne.es/proyectos/una-egloga-interlocutoria-graciosa-y-por-gentil-estilo-trovada-del-siglo-XVI/

[6] https://comunidad.bne.es/proyectos/jardin-nobles-doncellas/

[7] Uno de los primeros proyectos que abordaron la transcripción comunitaria de textos empleando medios electrónicos fue el proyecto *Transcribe Bentham* (2010-), que se centra en la transcripción manual de los escritos del filósofo inglés Jeremy Bentham (1784-1832). El objetivo del proyecto era convertir en texto legible por los ordenadores los manuscritos de este filósofo británico. Así, según Causer *et al.* (2018: 471, tablas 1 y 2), durante un periodo de dos años se transcribieron cerca de 11 000 documentos, lo que supuso un total de casi 1,2 millones de palabras.

Figura 5

Instrucciones para la transcripción colaborativa de la Égloga interlocutoria graciosa y por gentil estilo trovada



Nota. Fuente: Comunidad BNE.

Este tipo de trabajos tiene escaso valor científico, aunque cumple sobradamente un objetivo diferente, el de la difusión cultural. Precisamente por ello, son esfuerzos que se limitan a textos breves: el *Jardín de nobles donzellas* es un folleto en cuarto de cincuenta hojas, abordable en muy poco tiempo y que puede ser transcrito por pocas personas. Por el contrario, asumir la transcripción de un texto extenso y complejo, como el *Regimiento de príncipes* de Egidio Romano, de 250 folios, es una tarea inabordable para este tipo de iniciativas[8].

Existen otras fuentes que permiten acceder a las versiones electrónicas de algunos de los textos atesorados y transmitidos por los incunables castellanos. Uno de los proyectos más conocidos fue el *Archivo digital de manuscritos y textos españoles* (ADMYTE) (Faulhaber y Marcos-Marín, 1990) que, en su primer disco (1991) ofreció la transcripción, de acuerdo con las normas del proyecto del *Dictionary of Old*

[8] La BNE conserva cuatro ejemplares del *Regimiento de príncipes*: INC/242, INC/606, INC/901 e INC/1511.

Spanish Language (Buelow y Mackenzie, 1977), de sesenta y un textos procedentes de libros antiguos[9]. Junto con las transcripciones se ofrecieron las reproducciones, en imágenes BPM de 1-bit (bitonales), de las obras transcritas[10]. Sin embargo, el uso de ADMYTE en la actualidad es prácticamente nulo, ya que se distribuyó en discos ópticos (CD-ROM) creados con un *software* que funcionaba en el sistema operativo Windows 3.x, que dejó de tener mantenimiento a finales de 2001, diez años después de su lanzamiento.[11]

A pesar de este panorama, el nivel de desarrollo tecnológico actual permite la transcripción rápida de manuscritos medievales e impresos incunables y postincunables, alcanzando bajísimas tasas de error y con validez científica suficiente como para incorporarse a los corpus lingüísticos de referencia (Bazzaco, 2024). Uno de ellos, el *Old Spanish Textual Archive* (Gago Jover y Pueyo Mena, 2018a, 2018b, 2020) —en adelante, OSTA—, es una continuación del proyecto lexicográfico del *Dictionary of Old Spanish* (Nitti, 1978) que nació en 1970 en la Universidad de Wisconsin. Sin embargo, la versión 1.0 de OSTA solo recoge el contenido de 55 incunables, lematizados y analizados morfológicamente (Figura 6), de los más de 230 recogidos en BETA(mientras que la versión 2.0 añade 38 nuevos incunables, llegando a un total de 93).

---

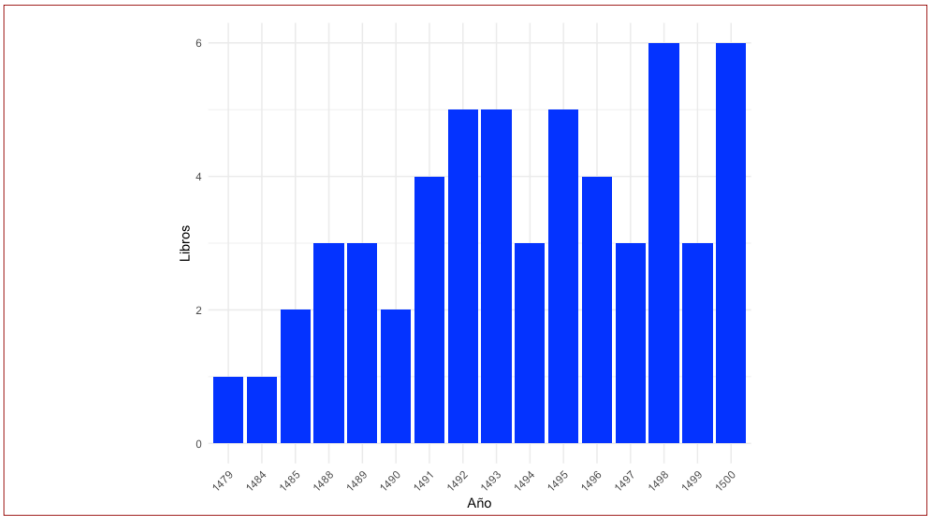
9 El cómputo de libros es ligeramente diferente, puesto que algunas obras se encuentran en un mismo volumen, como cinco textos de Bernardo Gordonio (Sevilla: Meinardo Ungut y Estanislao Polono, 1495) conservados en el BNE INC/2438. Además, el periodo considerado por ADMYTE es mucho más amplio que el de los incunables, pues la obra más moderna que se incluyó fue impresa en 1531, un decenio más allá del límite de los llamados postincunables.

10 Sobre el proceso de digitalización, véase Marcos Marín (1994: 184-189). Sorprendentemente, algunas digitalizaciones bitonales usadas para ADMYTE siguen disponibles en la página de la BDH, como los INC/1405 (*Leyes por la brevedad y orden de los pleitos*) INC/2157 (*De imitatione Christi* de Jean de Gerson).

11 Actualmente, la única forma de acceder a los contenidos de este tipo de bases de datos —a las que podemos sumar la edición electrónica del *Diccionario crítico-etimológico castellano e hispánico* de Joan Corominas y José Antonio Pascual (publicado por Gredos en 2012) y la *Patrologia Latina Database*, que reúne los 221 volúmenes editados por J.-P. Migne en el siglo XIX (re-editado por Chadwyck-Healey en 1995; disponible también bajo un modelo de suscripción en ProQuest)— es transfiriendo los contenidos de los CD-ROM a un formato de imagen de disco óptico (ISO), virtualizando los sistemas operativos compatibles (Windows 3.x, Windows 95 o Windows 98) en emuladores anidados en sistemas actuales, como UTM (Mac) o VirtualBox (Mac y PC), creando unidades virtuales de discos compactos y ejecutando los contenidos. Sin embargo, la progresiva desaparición de lectores de CD-ROM dificulta la transferencia al formato ISO y también fuerza a virtualizar unidades de disco óptico dentro de los sistemas operativos virtualizados (para lo que son necesarios códecs que no siempre son compatibles con el funcionamiento de los discos originales: la *Patrologia Latina* pide que el usuario introduzca y retire los discos ópticos en orden para realizar una búsqueda global en el corpus, pero en ciertos emuladores, los discos virtuales solo pueden ser creados antes de iniciar el sistema operativo virtualizado y no se pueden cambiar durante la ejecución). Además, aunque la información de las bases de datos puede ser consultada y, en algunos casos, exportada, el uso de sistemas de protección de archivos impide la exportación de otra información, como las imágenes en formato BPM usadas por ADMYTE.

Figura 6

Número de libros impresos en castellano durante el periodo incunable incluidos en OSTA



Nota. Fuente: Elaboración propia con información recolectada de OSTA.

¿Es factible la transcripción rápida y altamente fiable de textos antiguos, de textos de valor histórico? Hoy en día la respuesta es sí, debido al reciente desarrollo y disponibilidad de sistemas de reconocimiento automático de textos (*automatic text recognition*, ATR). El ATR, cuyo objetivo es extraer el texto de imágenes digitalizadas —fotografías digitales o escaneos de los originales—, combina los mundos de la visión artificial (*computer vision*) y del procesamiento de lenguas naturales (NLP por sus siglas en inglés). La tarea básica de un sistema ATR, diseñado como un problema de aprendizaje automático supervisado (*supervised machine learning*), es extraer los rasgos pertinentes (*features*) a partir de las imágenes de una palabra, o de toda una línea de texto, y transformarla en una secuencia de caracteres. De esta manera, el primer problema que ha de resolver es el de la segmentación de la imagen para determinar las zonas que contienen texto (*layout analysis*). El segundo es ver cuál es la probabilidad de que el carácter que ve se corresponda con un carácter o una secuencia de caracteres (*n-gram*) dentro del modelo aprendido.

Dentro de los sistemas ATR existen dos campos relacionados, el reconocimiento automático de la letra manuscrita (*handwritten text recognition*, HTR) y el reconocimiento óptico de caracteres (*optical character recognition*, OCR). Este último se ha empleado fundamentalmente para textos impresos modernos pues, como lo demuestra la Figura 4, no funciona con impresos antiguos debido a que los caracteres no pueden ser aislados y procesados con facilidad por los sistemas de OCR al uso (Mancinelli, 2016: 256). Por oposición, los modelos HTR, desarrollados sobre redes

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

neuronales convolucionales (*convolutional neural networks*), procesan los textos manuscritos o impresos (que pueden o no tener consistencia gráfica, es decir, ser regulares) en forma de líneas secuencializadas, a partir de las cuales el modelo de lenguaje (*language model*) infiere un conjunto de reglas estadísticas que le permite reconocer futuras secuencias gráficas (Strauß *et al.*, 2017: 7). Los recientes desarrollos en el HTR y su aplicación exitosa a textos impresos de los siglos XV y XVI (Bazzaco, 2020) ofrecen una solución a la problemática que hemos descrito hasta aquí.

2. El flujo de trabajo de un modelo HTR

El flujo de trabajo para la creación de un modelo HTR tiene ocho etapas (Figura 7). En puridad, los dos primeros pasos son anteriores al proceso de creación del modelo, pero los hemos incluido en el flujo debido a que son determinantes para obtener un modelo HTR viable. Emplear imágenes de baja resolución (cualquier número debajo de los 150 dpi), significa que los detalles son poco legibles y que las líneas y los trazos de la manuscritura o la letra impresa comienzan a desdibujarse. Esto resultará en un modelo con una tasa de error elevada y, por tanto, inservible (no ocurre lo mismo en el sentido contrario: transcribir una imagen de baja resolución con un modelo HTR entrenado con imágenes de alta resolución y con una alta tasa de confiabilidad generalmente tiene buenos resultados).

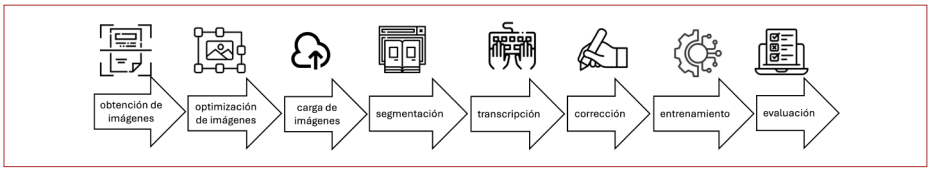
Sin embargo, un problema recurrente en nuestra experiencia es que es imposible controlar la calidad de las imágenes publicadas por las instituciones que preservan las colecciones históricas. Actualmente, muchas bibliotecas, como la BNE o la Real Biblioteca del Monasterio de El Escorial, han comenzado a restringir la consulta de los originales debido a consideraciones de conservación, favoreciendo, en cambio, el uso de las reproducciones digitales por parte de los investigadores. Aunque por lo general, las calidades que ofrecen estas bibliotecas y archivos en las reproducciones accesibles en línea sin coste suelen ser suficientes, el tamaño de los originales puede crear problemas, como hemos confirmado con manuscritos digitalizados por la BNE, la Bibliothèque nationale de France (BnF) y la Biblioteca Histórica de la Universidad de Salamanca (USAL). Así, una resolución de 150 dpi en las imágenes de un manuscrito o impreso en gran folio, pero con una letra muy pequeña y menuda, como la *Crónica de 1344* de los BNE mss. 10814 y 10815, el *De preconiis Hispaniae* del USAL ms. 1821 o el *Confesional* del Tostado del Biblioteca de Catalunya Inc. 84-8, hacen ilegible la reproducción digital[12]. En otros casos, las reproducciones digitales disponibles se basan en reproducciones fotográficas anteriores (microfilms bitonales: blanco y negro) de muy baja resolución, como el

[12] La BDH difunde pública y gratuitamente documentos en formato PDF con una resolución de 150 dpi, mientras que las imágenes máster, almacenadas en formato TIFF (sin compresión, con profundidad de color de 24 bits y resolución de 400 dpi), son ofrecidas con un precio de 5 € por imagen.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Figura 7

Flujo de trabajo para la creación de un modelo HTR



Nota. Fuente: Elaboración propia.

Libro de cetrería de Evangelista (BNE ms. 21549), el *Cancionero de Salamanca* (USAL ms. 2139) o la *Primera Partida* (BnF ms. Espagnol 440)[13].

Tras el preprocesamiento de las imágenes, se puede continuar al tercer paso del flujo de trabajo, cargar las imágenes en los servidores del servicio de HTR[14]; dependiendo de la plataforma, se aceptarán imágenes en formatos JPG, PNG o incluso archivos en PDF, ya sea a través de carga directa o utilizando un servidor FTP.

[13] La BnF, por ejemplo, ha comenzado a sustituir sus viejas reproducciones digitales, realizadas a partir de microfilms, por nuevas reproducciones digitales profesionales. Así, los mss. Espagnol 36 (*Libro del caballero Zifar*) y Espagnol 216 (*Libro de la montería* de Alfonso XI) cuentan ahora con copias en alta resolución. Pero en vez de eliminar el acceso a las primeras reproducciones digitales, la biblioteca las ha mantenido junto a las nuevas versiones. Este tipo de buena práctica archivística hace que los especialistas puedan realizar un seguimiento de la evolución de los manuscritos en tiempos modernos. El año pasado, la comparación entre un microfilm de un ejemplar de la *Séptima Partida* realizado a mediados del siglo xx y una reproducción digital realizada en 2023, permitió establecer que un fragmento de este manuscrito, que forma parte de la biblioteca de la Real Colegiata de San Isidoro de León, había sido sustraído de la colección en los años sesenta y eventualmente vendido a la Biblioteca Real de Bélgica (Fradejas Rueda, 2023).

[14] Entre las herramientas de HTR disponibles actualmente, en teoría, la plataforma eScriptorium se puede instalar y ejecutar en un ordenador personal —«In fact, you can also deploy eScriptorium on your personal machine, simulating a local server» (Chagué y Clérice, 2023: 1). La realidad, sin embargo, es que es necesario poseer una compleja infraestructura de servidores, unidades de almacenamiento y tarjetas gráficas donde se puedan desplegar el aplicativo, almacenar los modelos y las imágenes y realizar las computaciones necesarias para las tareas de HTR, sin considerar los costes de mantenimiento. Clérice *et al.*, (2023: 2) indican que el proyecto *Consortium pour la reconnaissance d'écritures manuscrites de matériaux anciens* (CREMMA) «was created to fund a regional server [...]. The CREMMA funding consisted of a grant for the initial cost of the infrastructure». Más accesible nos parece un sistema comercial como Transkribus (Kahle *et al.*, 2017), que asume los costos de la infraestructura y el desarrollo tecnológico en su sistema de pago, y ofrece a los usuarios un servicio de uso intuitivo y sostenible (Nockels *et al.*, 2024 y Terras *et al.*, 2025). Esta dualidad de alternativas en el mundo del HTR es la misma que ocurre con los sistemas operativos, divididos entre el *software* gratuito y el *software* de pago. Quizá la analogía más apta sean los sistemas basados en el *kernel* Linux y los sistemas basados en Windows. Los primeros, a pesar de ofrecer su código en acceso abierto, como eScriptorium, no son sencillos de instalar (Kiessling *et al.*, 2019; Chagué y Clérice, 2023); mientras que los segundos, a pesar de ser programas de código cerrado y propietario, tienen una curva de aprendizaje corta y son amigables con un usuario no experto. Aunque escapa al propósito de este artículo, este repaso revela un problema conceptual más serio en el campo de las humanidades digitales, que merece ser discutido. Por lo pronto, asumimos la conclusión de Haugen (2006: 342): «technology should be the handmaid of philology (and not the other way around)».

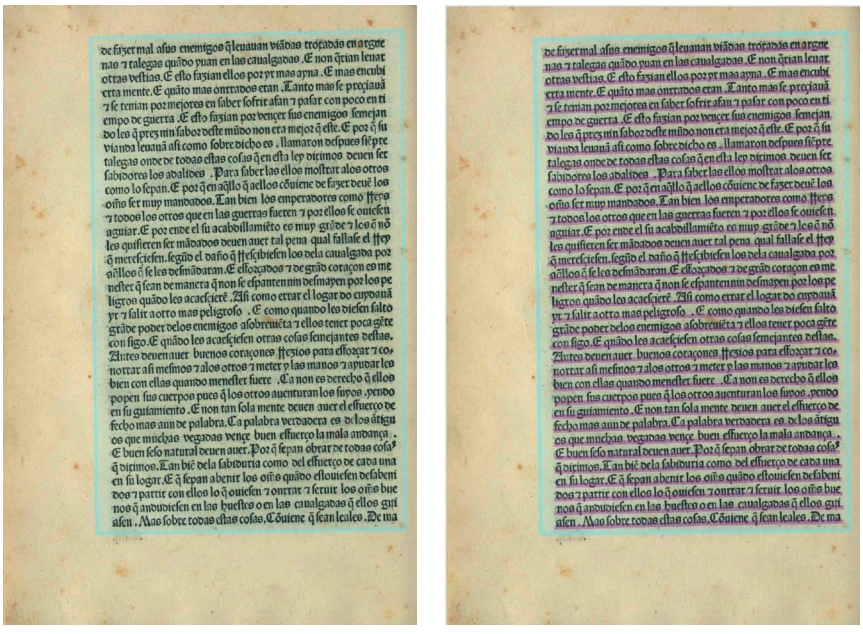
ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

2.1. Segmentación

El cuarto paso del flujo de trabajo es la segmentación (*layout analysis*), es decir, hacer que el software reconozca las zonas de texto que hay en una imagen. A lo largo de este proceso, se analizan las imágenes y se establecen las diferentes zonas de texto que contiene. Dicho de otro modo, se analiza la *mise-en-page* de la página impresa para determinar dónde se encuentra cada zona y cuál puede ser su orden de lectura cuando hay disposiciones complejas. El proceso se puede dividir en dos fases: en la primera se determinan los bloques de texto, y en la segunda se establecen las líneas de texto dentro de cada zona, aunque también es factible hacerlo todo de una vez. En algunos casos, el análisis es simple y rápido, como en el caso de páginas a una sola columna, como ocurre en el *Doctrinal de los cavalleros* (C87) de Alfonso de Cartagena (figura 8). En otros casos hay maquetas complejas, como los *Cinco libros de Séneca* (CLS) traducidos por Cartagena (figura 9), donde la página tiene el título corriente (*Dela prouidençia de dios.*) en el margen superior y la signatura (*i iij*) en el inferior, además de la mancha central compuesta en un tipo de mayor tamaño (que incluye la rúbrica —*Capitulo .vj.*— y una inicial decorada) y alrededor de esta una serie de glosas aclaratorias de Cartagena.

Figura 8

Segmentación sencilla: una sola zona de texto (*izquierda*) *y segmentación por líneas* (*derecha*). *C87, fol. C4v*

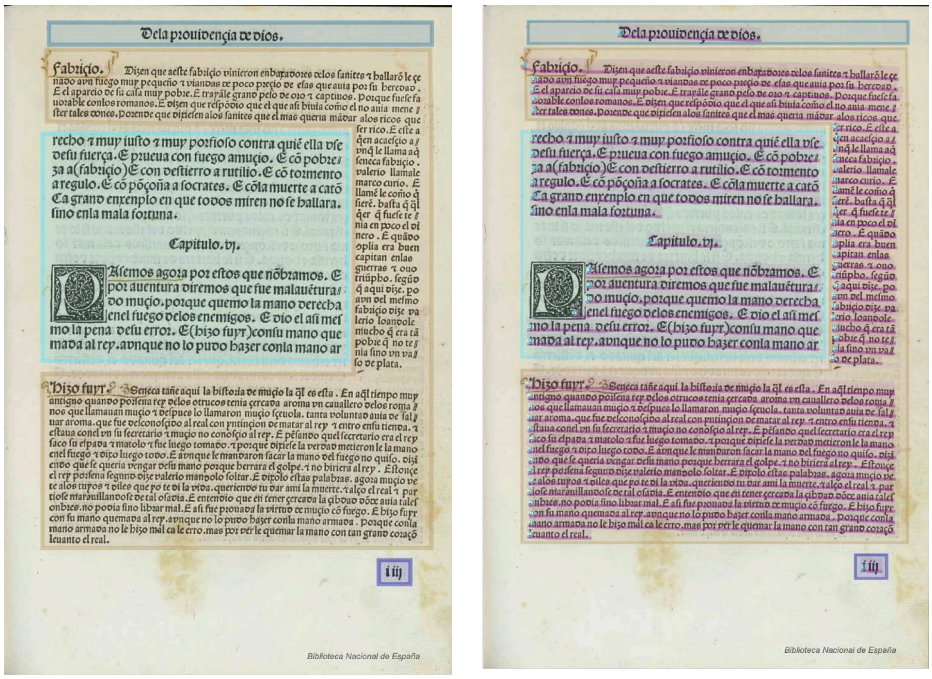


Nota. Fuente: Análisis propio.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Figura 9

Segmentación compleja: zona de texto —mancha central, glosas marginales, título corriente y signatura— (izquierda) y segmentación por líneas (derecha). CLS, f. i3r



Nota. Fuente: Análisis propio.

2.2. Transcripción

Una vez que se ha segmentado el texto, se ha de proceder a la transcripción de un número limitado de páginas para crear un texto de verdad base (*ground truth* o *golden corpus*) con el que se realizará el entrenamiento del modelo. Si la segmentación es importante, mucho más lo es la transcripción.

Antes de realizar la transcripción es necesario determinar cuál es el objetivo final del modelo, dado que esto determinará el tipo de transcripción que elegiremos. Si lo que se pretende es facilitar el acceso al texto, se puede recurrir a transcripciones normalizantes o regularizantes en las que se toma como criterio el de las llamadas ediciones escolares (Fradejas Rueda, 1991: 48) o incluso el sistema de presentación crítica desarrollado por la red Charta (Sánchez-Prieto Borja, 2011)[15]. Si el objetivo es

[15] Esta es la filosofía de transcripción que se ha seguido en los modelos HTR desarrollados en Transkribus para los incunables poéticos (Spanish Gothic Poetic Incunabula, Model ID 128193) y los

ir más allá del mero acceso al texto, e incorporarlo a un corpus lingüístico o usarlo como material para preparar una edición académica (científica) digital, entonces solo es factible un tipo de transcripción: la paleográfica.

En algunos casos se puede hacer una primera transcripción a partir de un modelo HTR general que se aproxime a lo que el proyecto se propone. Si no hay ningún modelo que se pueda tomar como base, entonces la única solución es transcribir manualmente un buen número de páginas. La opinión generalizada, pero no establecida científicamente, es que se han de transcribir como mínimo entre 5 000 y 15 000 palabras, es decir, entre 25 y 75 páginas de texto. Este número depende de varios factores: la cantidad de palabras por línea, el uso de abreviaciones de los copistas o cajistas, el número de líneas y columnas por página. En nuestra experiencia, el desarrollo inicial de un fichero *ground truth* para un manuscrito castellano (caligrafías góticas de los siglos XIV y XV) puede necesitar entre 25 y 40 folios, mientras que los impresos suelen situarse entre 15 y 20 folios.

2.3. Corrección

Una vez que se ha obtenido una primera transcripción, ya sea por medio de un modelo HTR preexistente, ya sea por medio de una transcripción *ex novo*, y antes del entrenamiento del modelo HTR, hay que corregir con sumo cuidado el texto. Uno o dos errores pueden pasar totalmente desapercibidos en el fichero *ground truth* (¡y lo harán!). Esto es especialmente importante cuando la transcripción se ha obtenido empleando otro modelo HTR, pues la máquina no ha leído realmente los caracteres que representa el texto, sino que predice la probabilidad de que pueda aparecer en una combinación de caracteres o una palabra específica en base a una tipología manuscrita o impresa diferente a la que estamos transcribiendo. El uso de modelos para entrenar nuevos modelos HTR (o reentrenar los mismos) debe tener en cuenta que los reentrenamientos no supervisados y corregidos suelen producir errores recursivos: el modelo se entrena en una transcripción errónea, que se utiliza para realizar una nueva transcripción errónea, que se añade al fichero *ground truth* y se utiliza para reentrenar el modelo, creando una cadena repetitiva de entrenamientos que acumulan y amplifican los errores del modelo original.

2.4. Entrenamiento

El entrenamiento del modelo es el corazón de cualquier sistema HTR. Un modelo es un archivo que se crea, mejora y entrena con un fichero de entrenamiento (*train*

impresos castellanos góticos de los siglos XV y XVI (SpanishGothic_XV-XVI_extended_v1.2, Model ID 45941), que contempla obras en prosa, de los que hablaremos más adelante.

set) al que se le ha concedido la categoría de *ground truth*. Este fichero consiste en varias páginas de transcripciones (paso 2), alineadas línea a línea con las imágenes que se cargaron (paso 3) y se segmentaron (paso 4). Estas transcripciones han de ser revisadas manualmente (paso 5). La transcripción corregida se considera la verdad (*truth*), y la inteligencia artificial la utiliza para aprender a distinguir el signo (carácter) correcto.

Durante el proceso de aprendizaje, el sistema apartará unas pocas de páginas del fichero *ground truth*, elegidas por el usuario, para usarlas como juego de validación (*validation set*). Lo que hará la máquina, tras el entrenamiento, es una transcripción de las páginas extraídas como juego de validación y las comparará con la transcripción del fichero *ground truth* y analizará la tasa de éxito.

2.5. Evaluación

La efectividad del modelo HTR entrenado se medirá a la luz de la tasa de errores. La forma de evaluación más común es la tasa de errores de caracteres (*character error rate*, CER). Esta métrica compara el número total de caracteres en el juego de entrenamiento (N), incluidos los espacios, con la suma de la adición (I), de la sustitución (S) y del borrado (D) de caracteres necesarios para obtener el resultado del fichero *ground truth*. Esto significa que un espacio mal colocado, la confusión de una letra por otra (*n* por ñ; *u* por *n*), cuenta como un error y forma parte del CER.

$$\text{CER} = \frac{\text{I} + \text{S} + \text{D}}{\text{N}}$$

La otra métrica que se utiliza es la tasa de error de palabras (*word error rate*, WER). Esta calcula el número mínimo de adiciones (I), borrados (D) y sustituciones (S) de una palabra para convertirla en la palabra correcta según el juego de validación.

$$\text{WER} = \frac{\text{I} + \text{S} + \text{D}}{\text{N}}$$

En realidad, ambas métricas son la misma, solo que una establece el error desde el punto de vista de los caracteres (CER) y la otra desde el de palabras (WER). Por este motivo, la palabra se ha de entender como una secuencia de caracteres alfanuméricos entre dos espacios en blanco o signos de puntuación, en el texto transcrito en la Figura 4 solo hay tres palabras transcritas correctamente: *Virgen*, *noble* e *infante*, lo que implica una tasa de error de 95,16 %.

Una vez obtenido un modelo cuya tasa de éxito (CER y WER) se considere óptima, se puede llevar a cabo la transcripción del ejemplar usado para el entrenamiento u otros con idéntica tipología manuscrita o impresa.

3. Creación del modelo Spanish Gothic Incunabula (HSMS)

3.1. Antecedentes

Dentro de los modelos públicos disponibles en Transkribus que pueden ser empleados para la conversión a texto legible de incunables impresos en español, se encuentran los modelos Spanish Gothic Poetic Incunabula, desarrollado por Enrique Ripoll con una reducida tasa de CER de 0,51 %[16], y SpanishGothic_XV-XVI_extended_v1.2, desarrollado por un equipo encabezado por Stefano Bazzaco con un reducido (pero algo *peor*) CER de 0,91 %[17].

En la Figura 10 se ofrece una imagen del íncipit del *Regimiento de príncipes* de Egidio Romano y, debajo de ella, la propuesta de transcripción que han presentado el modelo Spanish Gothic Poetic Incunabula (izquierda) y modelo SpanishGothic_XV-XVI (derecha). Aunque este último modelo tiene un CER ligeramente más elevado (de 0,91%), el resultado de su transcripción es muchísimo mejor. En la Figura 11 se han marcado los problemas detectados en ambas transcripciones. En el caso de la solución del modelo SpanishGothic_XV-XVI solo hay dos *fallos* dignos de mención: la marca de corte a final de la línea 5, que no procede, e ignorar por completo el «et cetera» de la última línea. En el caso de la propuesta de Spanish Gothic Poetic Incunabula, hay nueve errores de transcripción (marcados en amarillo: *Recho* > *Fecho*, *on* > *Don*, Sil > *Gil*, *Osmat* > *Osma*, *el* > *e*, *Pon* > *Don*, *sey* > *Rey*[18], *Toleco* > *Toledo* y *lleon* > *Leon*); tres problemas de corte de palabra al final de la línea (en verde, líneas 3, 6 y 7), aunque este es fácilmente explicable: la poesía no corta palabras al final de la línea, con lo que el modelo ignora esta posibilidad, lo cual, se podría interpretar como un sesgo del modelo. En azul están marcados los problemas de división de palabras dentro de las líneas (*Pedrorfijo* > *Pedro: fijo* y *Alfonsorsey* > *Alfonso: Rey*), creados por la presencia de un signo de puntuación, en ambos casos los dos puntos (:), que también puede tratarse de un problema del diseño del modelo, que, por lo que parece, fue entrenado para ignorar la puntuación[19].

[16] Desarrollado para el proyecto «Poesía, ecdótica e imprenta» (PID2021-123699NB-I00). No hay información acerca de cuántos ni de cuáles ejemplares fueron usados para su entrenamiento, tan solo que han tenido en cuenta 47 930 palabras distribuidas a lo largo de 11 449 líneas (versos, más bien).

[17] Este modelo está entrenado sobre 20 impresos, de los que cuatro son incunables, cinco postincunables y el resto fueron publicados entre 1526 y 1563, tanto en talleres peninsulares como europeos (Lisboa, Roma y Venecia). El fichero *ground truth* está construido por 220 904 palabras y 25 531 líneas impresas.

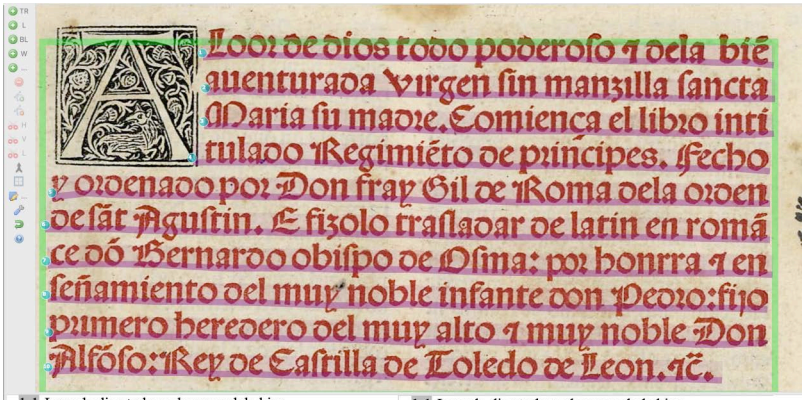
[18] Este error no se ha marcado con amarillo porque se combina con otro error marcado en azul.

[19] Cabe la posibilidad de que los textos poéticos utilizados para el entrenamiento carezcan de puntuación, extremo que no podemos comprobar puesto que no se informa de qué títulos y ejemplares se tuvieron en cuenta para su creación.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Figura 10

Transcripción automática del incipit del Regimiento de príncipes (RGP). Izquierda: hipótesis del modelo Spanish Gothic Poetic Incunabula; derecha: hipótesis del modelo SpanishGothic_XV-XVI



1-1	Loor de dios todo poderoso e dela bien	1-1	Loor de dios todo poderoso y de la bien
1-2	auenturada virgen sin manzilla sancta	1-2	auenturada virgen sin manzilla sancta
1-3	Maria su madre Comiença el libro inti	1-3	Maria su madre. Comiença el libro inti-
1-4	tulado Regimiento de principes. Recho	1-4	tulado Regimiento de principes. Fecho
1-5	y ordenado por on fray Sil de Roma dela orden	1-5	y ordenado por Don fray Gil de Roma de la orden-
1-6	de sant Agustin E fizolo trasladar de latin en roman	1-6	de sant Agustin. E fizolo trasladar de latin en roman-
1-7	ce don Bernardo obispo de Osmat por honrra el en	1-7	ce don Bernardo obispo de Osma: por honrra y en-
1-8	señamiento del muy noble infante don Pedrorfijo	1-8	señamiento del muy noble infante don Pedro: fijo
1-9	primero heredero del muy alto e muy noble Pon	1-9	primero heredero del muy alto & muy noble Don
1-10	Alfonsorsey de Castilla de Toleco de lleon e.	1-10	Alfonso: Rey de Castilla de Toledo de Leon.

Nota. Fuente: Elaboración propia.

Figura 11

Errores de transcripción, modelos Spanish Gothic Poetic Incunabula (izquierda) y SpanishGothic_XV-XVI (derecha)

1-1	Loor de dios todo poderoso e dela bien	1-1	Loor de dios todo poderoso y de la bien
1-2	auenturada virgen sin manzilla sancta	1-2	auenturada virgen sin manzilla sancta
1-3	Maria su madre Comiença el libro inti	1-3	Maria su madre. Comiença el libro inti-
1-4	tulado Regimiento de principes. Recho	1-4	tulado Regimiento de principes. Fecho
1-5	y ordenado por on fray Sil de Roma dela orden	1-5	y ordenado por Don fray Gil de Roma de la orden-
1-6	de sant Agustin E fizolo trasladar de latin en roman	1-6	de sant Agustin. E fizolo trasladar de latin en roman-
1-7	ce don Bernardo obispo de Osmat por honrra el en	1-7	ce don Bernardo obispo de Osma: por honrra y en-
1-8	señamiento del muy noble infante don Pedrorfijo	1-8	señamiento del muy noble infante don Pedro: fijo
1-9	primero heredero del muy alto e muy noble Pon	1-9	primero heredero del muy alto & muy noble Don
1-10	Alfonsorsey de Castilla de Toleco de lleon e.	1-10	Alfonso: Rey de Castilla de Toledo de Leon.

Nota. Fuente: Elaboración propia.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Dijimos que un error compartido por ambos modelos es el tratamiento del «et cetera» de la última línea. Este tipo de problema, ignorar o desarrollar erróneamente una secuencia gráfica completa, es común a todos los modelos HTR si cuando se creó el fichero *ground truth* no se incluyó texto con la forma, logrando que el modelo aprendiese a desarrollarla[20].

No es un error, sino una decisión editorial, el del tratamiento de la nota tironiana. En Spanish Gothic Poetic Incunabula, la nota tironiana se ha desarrollado consistentemente como *e* (líneas 1, 7, 9, 10), mientras que en SpanishGothic_XV-XVI aparece en dos casos como *y*, que es la solución mayoritaria y ocasionalmente aparece como *&*. Otra decisión editorial tomada por los autores de los modelos es cómo resolver la unión de las preposiciones y artículos escritos en un solo tramo: en el caso de Spanish Gothic Poetic Incunabula no se han separado cuando aparecen unidos (*dela*, líneas 1 y 5), mientras que en SpanishGothic_XV-XVI se tomó la decisión de separarlos.

En definitiva, de los dos modelos públicos disponibles para transcribir incunables castellanos, el que ofrece una tasa de acierto mejor es el SpanishGothic_XV-XVI. En el brevísimo pasaje elegido hay un único error (de división al final de línea), lo que da un WER de 1,25%. En cambio, en el modelo Spanish Gothic Poetic Incunabula el WER, en el mejor de los casos, es de 11,25 %.

3.2. La transcripción

El modelo SpanishGothic_XV-XVI es excelente si lo único que se pretende es realizar análisis literarios o incluso estilométricos, pero no es válido para el análisis lingüístico, pues oculta muchos rasgos en el desarrollo de las abreviaturas. Ambos comparten un problema, desarrollando sistemáticamente la nota tironiana bien como *e* (Spanish Gothic Poetic Incunabula), bien como *y* (SpanishGothic_XV-XVI), en una época en las que sus soluciones gráficas están contendiendo y aún no se han estabilizado[21].

[20] En los primeros modelos HTR creados en el proyecto *7PartidasDigital* para transcribir la segunda edición de las *Siete Partidas* (Sevilla: Cuatro Compañeros Alemanes, 1491) se detectó que la abreviatura *mrs* (m<a>r<avedie>s, m<a>r<avedi>s) se transcribía consistentemente como *mes*. Este error se debía a que el fichero *ground truth* de los modelos para impresos no recogía ningún caso de *mrs* y *mes* fue la solución que el modelo encontró estadísticamente más probable. En un reentrenamiento se incluyeron folios representativos, en los que sí aparecía la forma *mrs*, y el error desapareció. Por oposición, un modelo (Model ID 58784) desarrollado para uno de los manuscritos del Archivo Capitular de Toledo, ms. 43-11, que contiene la *Primera Partida*, y cuyo fichero *ground truth* recogía y resolvía varias apariciones de la abreviación *mrs*, resultó capaz de desarrollar correctamente la abreviación desde el primer intento.

[21] En las adiciones y concordancias que Alonso Díaz de Montalvo añadió al texto de la *Siete Partidas* (Sevilla: Meinardo Ungut y Estanislao Polono, 1491 (SPO)) hay 1784 casos de conjunción copulativa; de ellos, el 89,29 % son ocurrencias del signo tironiano, el 7,45 % de *e* y el 3,64 % de *y*.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Ciertamente, para el análisis lingüístico, y también para las ediciones críticas, se necesitan transcripciones con un altísimo grado de precisión y que traten de retener el máximo de información codicológica y paleográfica («*micro-features*», Guéville y Wrisley, 2024: 11). Aquí se plantea el problema de los niveles de transcripción, el cual se puede remontar a algunos tipos de ediciones de inicios del siglo XX, como la del *Libro de buen amor* de Ducamin (1901) y la posterior reseña de Menéndez Pidal (1901), pero que se ha reavivado desde el mismo momento en que los ordenadores pasaron a ser una parte importante de las herramientas de la filología.

En el proyecto Menota (*Medieval Nordic Text Archive*) se establecieron tres niveles de transcripción (Haugen, 2004: 78-79): *facsimilar* («each character, whether it is an ordinary character or an abbreviation mark is faithfully copied, and mayor allographs are recognized [...]. Word and line divisions are clearly marked, so that for every line of the manuscript there is a faithfully reproduced in the facsimile transcription»); *diplomática* («there is usually less allographical variation [...]. More important, however, is the fact that abbreviations are expanded»); y *normalizada* («[t]he normalized orthography is based in standard dictionaries and grammars»).

Robinson y Solopova (1993: 22-23) establecieron cuatro niveles de transcripción cuando abordaron la edición digital del prólogo del cuento «The Wife of Bath» de los *Canterbury Tales* de Geoffrey Chaucer: *gráfica* («every mark in the manuscript, every space, is represented in the transcription, even to the point of decomposition of letter-forms into discrete marks»); *grafética* («every distinct letter-type is distinguished as: *r* short is transcribed apart from *r* round and *r* long descender, etc.»); *grafémica* («every manuscript spelling is preserved (as: 'she', 'sche') without distinction of separate letter-forms as in a graphetic transcription"); y *regularizada* («all manuscript spellings are regularized to a particular norm, perhaps the spelling of a manuscript considered authoritative»).

Camps (2017: 31-32) solo considera dos niveles de transcripción: la *alográfica* («pas de normalisation des allographes, ni de résolution des abréviations ou de la ponctuation et des accents médiévaux ; lettrines imprimées aux dimensions qu'elles occupent dans le manuscrit, en nombre de ligne de réglure ; pas de normalisation de la segmentation, ni d'emploi des diacritiques des règles de Meyer-Roques ; indication des ajouts ou suppression par des artifices typographiques») y la *grafemática* («Les allographes sont normalisés, et les abréviations résolues, selon les principes exposés en introduction, les suppressions des copistes retirées et leurs ajouts intégrés ; une ponctuation éditoriale se substitue à la ponctuation médiévale, et les majuscules sont alignées sur l'usage moderne. Les diacritiques des règles de Meyer-Roques sont utilisés»).

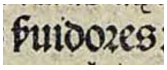
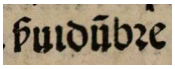
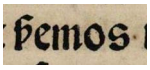
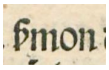
Por su parte, Guéville y Wrisley (2024: 5) hablan de tres niveles de transcripción: *normalizada*; *semidiplomática* («shows how special letter-forms can be preserved as they are written in the text, making the difference between *u*/*v* or *s*/*ſ*, preserving the original capitalisation or spacing as much as possible. This method expands

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

the abbreviations but usually indicates their purposeful expansion»); y *diplomática* («seeks to preserve as much information as possible from the original manuscript [...] this kind of diplomatic transcription identifies written characters, linking them to Unicode»).

Tabla 1

Ejemplos de brevígrafo o glifo para ser- en incunables castellanos

	SVH, fol. 8r1[4]
	SVH, fol. 8r1[5]
	CTY, fol. 32v2[16]
	CTY, fol. 43v2[22]
	CAU, fol. 10r2[1]

Nota. CTY: *Crónica Troyana* (Pamplona: Arnaldo Guillén de Brocar, 1500), según el BNE INC/733. CAU: Guido de Cauliaco, *Cirugía* (Sevilla: Meinardo Ungut y Estanislao Polono, 1498) según el BNE INC/196. Para el resto de las siglas, véase la Tabla 7.

Tanto Guéville y Wrisley (2024) como Gille Levenson (2023a), los primeros con Transkribus y el segundo con eScriptorium, tratan de representar los distintos alógrafos que presentan sus fuentes para la construcción de modelos HTR empleando caracteres del juego de caracteres Unicode y las extensiones desarrolladas por la *Medieval Font Initiative* (MUFI)[22]. Todos ellos se dan cuenta de los problemas que ese sistema supone, pues no todos los caracteres necesarios para la transcripción se encuentran dentro del dominio público del UTF-8 y es necesario recurrir al llamado dominio privado[23]. Gille Levenson (2023a: 8) menciona específicamente el

[22] https://mufi.info/q.php?p=mufi

[23] Además, Haugen (2006: 343) llamó la atención hace décadas sobre un problema adicional del que hablaremos más adelante, la interoperabilidad, resultante de la multiplicación de los juegos de caracteres (fuentes) empleados para transcribir manuscritos medievales e impresos tempranos, cada uno adaptado a la necesidad de un proyecto o un idioma específico, pero no compatibles entre sí: «Modern font technology has simplified the process of designing special characters (e.g. Fontographer, FontLab) due to the widespread use of font design applications a number of non-compatible fonts have been produced since the late 1980's. [...] This has led to much unnecessary and sometimes disruptive work on file conversions».

brevígrafo para *ser-*, del que pueden verse varios ejemplos en la Tabla 1. Estos investigadores muestran que son pocos los casos de brevígrafos *peculiares*. Así, Guéville y Wrisley (2024: 11) afirman, sin ofrecer ejemplos, que el número de abreviaturas y grafías únicas es muy limitado, mientras que Gille Levenson (2023a: 8) argumenta que el brevígrafo mostrado en la Tabla 1 «only appears in the manuscript Q, about a dozen times»[24].

La Tabla 2 resume, con una línea extraída del incunable del *Regimiento de príncipes*, qué sé entiende por transcripción facsimilar-alográfica-diplomática, por una parte y grafemática-paleográfica-semidiplomática, por la otra, resumiendo las propuestas de Haugen (2004), Gille Levenson (2023a) y Guéville y Wrisley (2024), ya que todos trabajan con el sistema UTF-8 MUFI, algo que no estuvo al alcance de Robinson y Solopova (1993).

Tabla 2

Muestra de transcripción facsimilar-alográfica-diplomática y paleográfica-grafemática-semidiplomática, RGP, f. 228r1 (sig. F4r), línea 19

Haugen/Gille Levenson/Guéville y Wrisley	
Facsimilar-alográfica-diplomática	empaꝺoꝛ pompilio:q̄nꝺo yua ala faziẽ
Paleográfica-grafemática-semidiplomática	emper*a*ꝺoꝛ pompilio: q*ua*nꝺo yua ala fazie*n*

Nota. Fuente: Elaboración propia.

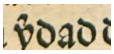
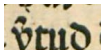
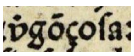
El sistema gráfico de MUFI está muy bien desarrollado para las lenguas germánicas medievales y se puede utilizar correctamente con el latín, pero no con las lenguas romances (al menos con el español). Ya hemos visto que no hay un tipo particular para la abreviatura de *ser-*. En el caso de la *v* de la forma abreviada *ver~vir-* en v<er>dad, v<ir>tud, v<er>dolagas o v<er>gonçosa, se ha de construir combinando dos caracteres, la *v* con una especie de coma sobre ella (*combining hook above*), pero no todos los impresores (ni copistas) usaron el mismo sistema; algunos emplearon un punto (*combining dot above*) sobre la *v* (Tabla 3).

[24] En OSTA se documentan 10 771 casos de *ser-* abreviado con ese mismo brevígrafo a lo largo de 293 textos manuscritos y hay 301 ocurrencias en 39 impresos. Es cierto que la frecuencia de aparición es relativamente baja: el incunable que presenta más casos, 34, es la versión castellana de *Fiammetta* (Salamanca: Juan de Porras[?], 1497), mientras que en la *princeps* de las *Siete Partidas* (Sevilla: Meinardo Ungut y Estanislao Polono, 1491 (SPO)) solo se documentan dos casos de entre 5543 posibles.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Tabla 3

Brevígrafos de ver- y vir-

	v̊ðað	GEN, f. 69r1, línea 27
	v̊tuð	GEN, f. 60r1, línea 37
	v̇ðolagaſ	GOR, f. 39v1, línea 9
	v̇gõçoſa	GOR, f. 163r2, línea 23

Nota. Fuente: GOR: Bernardo Gordonio, *Lilio de medicina* (Sevilla: Meinardo Ungut y Estanislao Polono, 1495) Elaboración propia.

Variaciones gráficas como esta suponen que, para que los investigadores puedan intercambiar datos y modelos HTR, lo primero que habría que hacer es ponerse de acuerdo en cómo codificar cada uno de los posibles signos que pueden aparecer en sus *originales*. Un caso peculiar al que nos hemos enfrentado es el del dígrafo que aparece en los libros impresos por Fadrique de Basilea para representar la vibrante múltiple en inicial absoluta (Tabla 4). No la contempla MUFI, pues no se encuentra dentro de las *r* usuales dentro del ámbito anglo-germánico medieval[25]. Uno de los trucos a los que se suele recurrir para poder representar lo que vemos, aunque no lo que el documento dice, puede ser usar otro glifo desarrollado por MUFI, el *latin small letter middle-Weslh ll*[26]. Aparentemente se trata del mismo símbolo, pero son dos eles unidas por un trazo superior, no una erre doble (aunque esta práctica no es aconsejable, pues pertenece al dominio privado de MUFI).

Lo que nosotros vemos como un problema, Bermúdez Sabel (2022: 15) lo considera una ventaja a la hora de realizar una edición digital, «la possibilité de concevoir nos propres polices, ce qui permet de créer des glyphes lorsque l'on considère que les polices existantes ne contiennent pas de lettres ou de combinaisons de caractères (des abréviations et des signes diacritiques) dont la forme reflète adéquatement la forme de l'original». Esta estrategia impediría la interoperabilidad de los materiales y su explotación posterior estaría muy limitada puesto que sus ficheros podrían ser ilegibles a muy corto plazo, que lo es lo que sucedió con las fuentes diseñadas por Robinson (1989: 99) para su estudio de sagas islandesas[27].

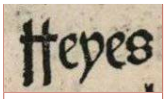
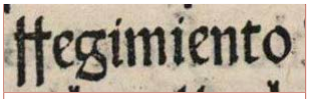
[25] ¿Convendría en estos casos hablar de archigrafema, relacionándolo con el archifonema de la fonología (Catach, 1990: 53), para resumir en un único símbolo todas las posibilidades gráficas de un grafema? MUFI (https://mufi.info/q.php?p=mufi/chars/char/R) recoge 30 formas para la *r*, pero no incluye nuestra particular *r* doble.

[26] La entidad que habría que introducir desde el teclado sería «&llwelsh»; o el código decimal «ỻ». Para más información, véase https://mufi.info/q.php?p=mufi/chars/unichar/7931

[27] «A font was designed in which the letter forms mimicked, as closely as possible, the characteristic letter forms of the scribes. The font was supplemented by various control codes, indicating

Tabla 4

La r en inicial absoluta en impresos de Fadrique de Basilea

			
ℓaſtro	ℓomanas	ℓeyes	ℓegimiento

Nota. Fuente: Elaboración propia.

Otro problema que concita intentar representar las *micro-features* paleográficas (o tipográficas) es que, al hacerlo, el riesgo de errores en un modelo HTR también aumenta exponencialmente. La Figura 12 es un pequeño pasaje del modelo creado por Gille Levenson (2023b)[28] para el incunable del *Regimiento de príncipes* de Egidio Romano[29]. En este brevísimo pasaje se descubren varias inconsistencias.

En primer lugar, resalta el distinto tratamiento que se da a la forma abreviada de q<ue>. Lo normal es que sea *q̃* (líneas 2, 4, 6, 8 y 13), pero hay tres casos en los que la lineta abreviativa no está indicada (marcados en rojo, en las líneas 10 y 12). En segundo lugar, es evidente el errático tratamiento de la puntuación. Todos los signos (puntos y dos puntos) aparecen siempre *unidos* a la palabra anterior (marcados en azul), por eso es extraño que los dos puntos de la línea 4 estén separados de la palabra que le precede, lo mismo sucede en la línea 15, mientras que en la línea 5 están separados de la palabra precedente y *unidos* a la siguiente. Lo mismo sucede con el punto y seguido de la línea 8. El tercer problema que se detecta es la errática unión y separación de palabras. No es que se hayan de considerar «las décimas de milímetros de más o de menos» (Sánchez-Prieto Borja, 1998: 100) que puedan separar una palabra de otra, sino ver si las palabras están separadas o no. La distancia que hay entre *q̃* y *ſin* y 7 y *ſin* (línea 2), entre *legiõ* y *d̈* (línea 6), entre *metio* y *ſe* (línea 9), entre *q̃* y *ſi* y *ſu* (línea 11) o entre *p̈mera* y *mẽte* (línea 17) es mayor, en todos los casos, que la que hay entre *cõ* y *el* (línea 13), pero estas dos palabras se han transcrito separadas, mientras que todas las demás están unidas.

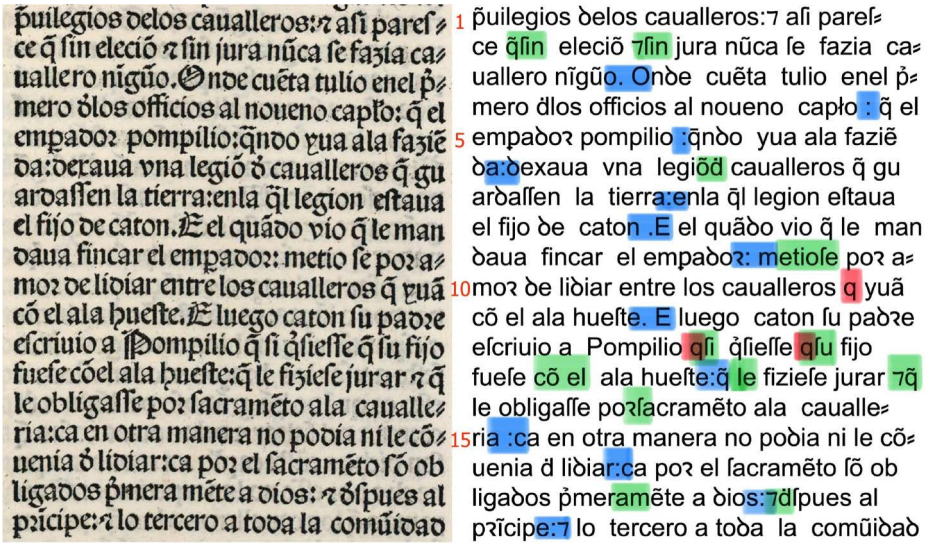
superscription, abbreviation, and the like» (Robinson, 1989: 99).

[28] El material, tanto el recorte del impreso, como la transcripción proceden del *dataset* publicado en Zenodo por Gille Levenson (2023b). La imagen procede del fichero data_v2/corpus/in_domain/Sev_Z/4_pg_18.jpg mientras que la transcripción se ha extraído de data_v2/corpus/in_domain/Sev_Z/4_pg_18.xml. El texto se ha obtenido del literal del atributo @CONTENT de la etiqueta <String> según el XMLSchema ALTO (Analized Layout and Text Object). Para la información técnica de ALTO, véase https://www.loc.gov/standards/alto/techcenter/structure.html

[29] El ejemplar utilizado por Gille Levenson es el BNE INC/901. La reproducción digital es accesible en https://bdh-rd.bne.es/viewer.vm?id=0000176298&page=749 (hasta la page=779).

Figura 12

Fragmento y transcripción alográfica



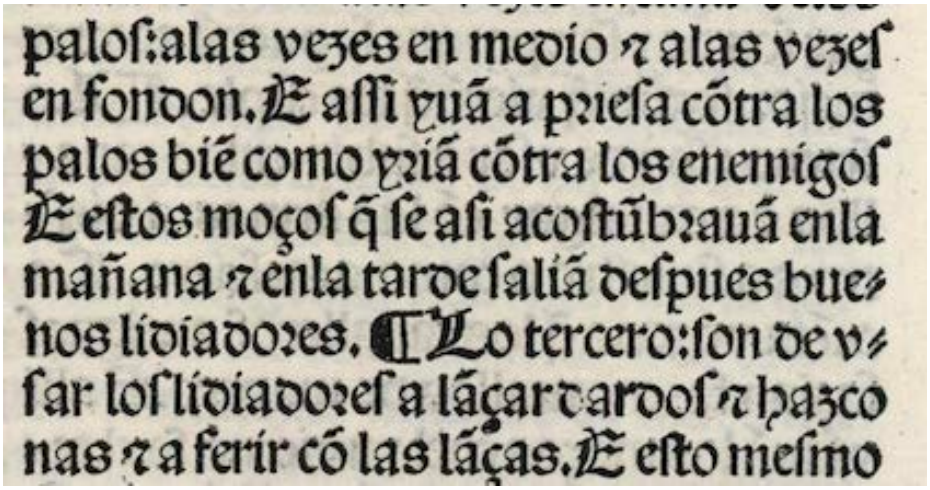
Nota. Fuente: Gille Levenson (2023b).

Aunque entendemos cuál es el objetivo de este tipo de transcripciones alográficas (Gille Levenson, 2023a), diplomáticas (Guéville y Wrisley, 2024) y facsimilares (Haugen, 2004)[30], realizadas porque cualquier otro tipo de transcripción «entails a loss of information» (Guéville y Wrisley, 2024: 6) cuando la «finalité principale [est] l'étude du système graphique, dans des perspectives relevant de la paléographie et de la linguistique de l'écrit» (Camps, 2017: 31), no compartimos la apreciación de que estas sean las formas de transcripción más adecuadas para el estudio y procesamiento de textos incunables (o incluso, manuscritos tardomedievales). Es cierto que, en algunos casos, la grafía puede ser un elemento básico para establecer la cronología de los manuscritos, como bien apuntó Sánchez-Prieto Borja (1998: 95) respecto al uso de la ꝺ uncial y la *d* recta. Así lo ha demostrado Rodríguez Díaz (2024), quien ha añadido la consideración de diferenciar la *r* de martillete y la ꝛ rotunda, que Sánchez-Prieto Borja no consideró pertinente. Sin embargo, en otros casos el mantener los alógrafos no aporta nada y puede complicar la tarea de la transcripción (y el entrenamiento de un modelo HTR). Podría pensarse que en impresos incunables la distinción entre *s* y ſ puede presentar una cierta distribución: la *s* solo

[30] Existe una cuarta posibilidad, las transcripciones digitales *ultradiplomáticas*: «nous utiliserons le terme *édition hyperdiplomatique numérique* pour désigner les éditions numériques conservatrices du point de vue de la reproduction du système d'écriture présent dans l'original et qui tentent aussi d'imiter fidèlement la mise en page de la source. Une édition numérique hyperdiplomatique exploite des techniques particulières d'hypertexte» (Bermúdez Sabel, 2022: 15).

Figura 13

RGP, f. 228r2 (sig. F4r), líneas 22-29



Nota. Fuente: *Biblioteca Digital Hispánica*.

aparece al final de palabra, mientras que la ſ ocurre a principio de palabra o en el interior (véase la Figura 12). Sin embargo, el cajista, en una misma página, puede utilizar la ſ en cualquier contorno, como puede verse en la Figura 13, en la que la ſ a final de palabra ocurre en 35% de los casos.

3.3. El HSMS

Pensamos, como Haugen (2004), Camps (2017), Gille Levenson (2023a) y Guéville y Wrisley (2024), que se ha de conservar el máximo de rasgos que puedan tener valor lingüístico, pero no creemos necesario caer en el ultradetallismo de las transcripciones facsimilares, alográficas, diplomáticas o hiperdiplomáticas que proponen, ni tampoco el extremo opuesto de la normalización usual presente en los estudios literarios, cuyos criterios se han visto en los modelos SpanishGothic_XV-XVI y Spanish Gothic Poetic Incunabula.

Nuestro punto de referencia para la forma de codificar las transcripciones del modelo HTR fue emplear un sistema extendido y versátil (adaptable). En el proyecto *7PartidasDigital*, como explicaremos más adelante, se asumieron las normas de transcripción semipaleográficas del HSMS, que cumplían ambas condiciones. Por un lado, se trata del criterio más extendido para la transcripción de textos en español antiguo (seguido del criterio de documentos desarrollado por la red Charta). Por el otro, emplear esta codificación permitirá que los usuarios del modelo HTR puedan incorporar las transcripciones de los incunables resultantes al corpus de

incunables y postincunables del *Old Spanish Textual Archive* o transformarlos con sencillez al sistema de etiquetado TEI (Fradejas Rueda, 2025) En cualquier caso, la decisión ha sido provocada por un tercer motivo: nuestra experiencia previa en el desarrollo de modelos HTR para manuscritos e impresos antiguos confirma que el sistema semipaleográfico del HSMS es compatible con Transkribus y que no genera confusiones o errores en el entrenamiento de los modelos en esta plataforma[31].

Los criterios de transcripción del HSMS fueron desarrollados por Kenneth Buelow y David Mackenzie y explicitados en el *Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language* (1977) como un sistema para codificar los manuscritos e incunables que servirían como base de datos textual para alimentar el gran proyecto lingüístico del Seminario, el *Dictionary of Old Spanish* (Mackenzie, 1994; Gago Jover y Pueyo Mena, 2018b).

Eventualmente, el HSMS comenzó a publicar estos materiales en la serie *Text and Concordances* en microfichas, desarrollada décadas antes de la aparición de las primeras reproducciones digitales, permitiendo que los investigadores de todo el mundo accedieran al contenido de los manuscritos e impresos castellanos en un momento en el que la última alternativa eran las reproducciones fotográficas o xerográficas. Con el tiempo, los criterios del HSMS se han modernizado. Así, en su primera versión se utilizaban solamente caracteres disponibles en el mapa original de caracteres ASCII (*ASCII 7-bits*), que excluía caracteres no corrientes en el inglés americano. En sucesivas revisiones del *Manual* se incorporaron caracteres del mapa expandido (*Extended ASCII*), como ç, ñ y ¶.

En cualquier caso, los textos producidos para el HSMS nunca han sido realmente ediciones, como se ha afirmado intermitentemente[32], sino transcripciones y, pre-

[31] Un cuarto beneficio del sistema de transcripción del HSMS, notado por Tenenbaum (2000-2001), es su longevidad. Al anteceder a la invención del formato de texto enriquecido y emplear el código más básico en el mundo informático (ASCII), es un sistema que permite la conservación persistente de archivos, a pesar de los «imprevisibles avances tecnológicos» (153). Las transcripciones originales del HSMS, los archivos máster, son almacenadas hasta el día de hoy en ficheros TXT. Como las proverbiales cucarachas atómicas, han sobrevivido eras informáticas enteras y la aparición y desaparición de sistemas operativos. Posiblemente sigan en pie cuando los lenguajes, formatos e iniciativas de codificación actuales sean piezas de *abandonware* que los usuarios deban virtualizar para extraer información (si no lo hicieron antes, lidiando con los problemas derivados de la exportación de información de formatos antiguos a nuevos).

[32] Lucía Megías (2002: 75), por ejemplo, habla de las «ediciones del Hispanic Seminary of Medieval Studies de Madison». En parte, esta confusión se debe a que la serie *Text and Concordances* del HSMS emplea «editado por» (*edited by*), entendido como «al cuidado de», para referirse al autor o autores de una transcripción, en vez de «transcrito por» (*transcribed by*) o «curado por» (*prepared by*). Además, algunas transcripciones iniciales de la serie, en 1970 y 1980, intentaron corregir lecturas de los manuscritos transcritos a la luz de otros testimonios supervivientes de una obra, pasando de ser transcripciones semipaleográficas (*semipaleographic transcriptions*) a ediciones paleográficas (*paleographic editions*), pues realizaban una forma rudimentaria de *collatio*. Así lo notó Orduna (1994: 219) correctamente sobre la *Corónica del rey don Pedro* preparada por Wilkins y Heanon (1985).

cisamente por ello, intentan representar la realidad física de los textos medievales mediante una codificación simple, cuyas reglas hemos adaptado para este modelo HTR de la siguiente manera:

1. Los signos de abreviación (barras, linetas, puntos o símbolos) se desarrollan entre corchetes angulares: ꝑlado > p<er>lado, ꝓmesa > p<ro>mesa, om̄es > om<n>es. El punto diacrítico sobre *ẏ* no se representa[33].
2. Los glifos que incluyen letras voladas son seguidas de un acento grave: p̓mera > p<r>i`mera; q̋l > q<u>a`l[34].
3. La lineta que suple la nasal implosiva antes de *b* y *p* se transcribe <n> (incluso cuando la forma desarrollada *m* se atestigüe; esta corrección corresponde a la edición).
4. El trazo vertical que cruza la *v*, v<er>- (incluso cuando se atestigüe la forma *vir*-).
5. La cedilla y la lineta sobre la *n* se transcriben ç y ñ. El signo tironiano y los calderones & y ¶.
6. La ruptura de una palabra al final de una línea se marca con un guion corto, aparezca o no en el original.
7. Se respetan mayúsculas, minúsculas y la separación o unión de las palabras del original. Se transcribe la puntuación unida siempre a la última letra a la izquierda e introduciendo, si no existe, un espacio a la derecha: hueste.E > hueste. E[35].
8. Para evitar la confusión con la forma de indicar intervenciones editoriales (de copistas o de transcriptores) en el sistema del HSMS, que usa el paréntesis para esta función, se introduce un espacio antes y después de los paréntesis: (fabriçio) > (fabriçio)[36].

[33] La excepción es la *h* con un trazo sobre el hastil, ħ, debido a que Antonio de Nebrija, en su *Gramática castellana* de 1492, la utiliza para marcar el segundo elemento del dígrafo de la africada palatal sorda *cħ*, pero este tipo es casi exclusivo de esta edición, el mismo Nebrija no aplicó la norma en sus demás obras publicadas. Sin embargo, esta hache, si bien no con una lineta, sino con una especie de apóstrofo, la hemos documentado en el incunable CNB como forma abreviada de la *ch*, con lo que aparece en las transcripciones como <c>h, en palabras como mu<c>ho, di<c>ho e incluso en el verbo e<c>har (f. 144r1 (sig. t2r), línea 29).

[34] Lo que pueden parecer dos puntos sobre la *q*, y que aquí se han representado como comillas rectas, se remontan a una *a* visigótica que podría asemejarse a una *w*. Este minúsculo problema, cómo representar esta *micro-feature*, es una de las cuestiones que nos hacen dudar de la validez y permanencia de los sistemas de transcripción ultraconservadores que aspiran a mostrar todos los pormenores paleográficos, porque ¿qué hacer con los rasgos expletivos o redundantes?

[35] Pueden verse varios ejemplos de puntuación unida a la palabra anterior y posterior en la figura 12.

[36] Pueden verse varios ejemplos de palabras encerradas entre paréntesis en las reproducciones de la figura 9. Una vez finalizada la transcripción, los paréntesis simples serán convertidos en dobles: ((fabriçio)).

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

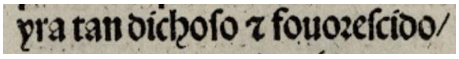
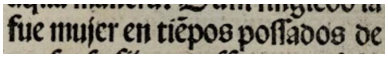
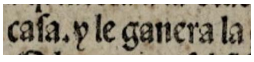
3.3.1. Nota bene

Ya que este modelo permitirá producir una primera transcripción de incunables castellanos, y no una versión final que podrá ser incorporada al OSTA, algunos puntos problemáticos deberán ser resueltos por los usuarios en la revisión final del texto transcrito. Nos referimos fundamentalmente a cuestiones relacionadas con el correcto desarrollo de las abreviaturas de los puntos 3 y 4, que deben ser establecidas según los usos del texto: si un texto desarrolla *embargos* y *embios*, el usuario deberá reemplazar la lineta con valor <n> por <m>. El glifo v̊ se ha desarrollado como v<er> y no como v<ir> para evitar que aparezcan casos de v<ir>dolaga o v<ir>dura, donde lo real es v<er>dolaga y v<er>dura (Tabla 3). Por lo tanto, los editores deberán estar atentos a casos como el de v<er>gen por v<ir>gen, o de v<er>tud (cuando el resto del texto lea *virtud*) por v<ir>tud.

También en esta etapa deberá revisarse la separación y unión de palabras, mediante intervenciones editoriales, pues estas correcciones son necesarias para la correcta lematización del texto en el OSTA. Para las intervenciones editoriales de adición o eliminación, se emplean los corchetes o paréntesis, e incluso para el marcado de erratas evidentes del impreso, que se hace por medio de borrado, marcado con paréntesis, y la adición de la forma correcta, indicada con corchetes (Tabla 5).

Tabla 5

Erratas evidentes en LIM y su corrección en la versión final

LIM, fol. 3r2, línea 2	
transcripción del modelo	yra tan dichoso & fouorescido /
transcripción final en OSTA	yra tan dichoso & f(o)[a]uorescido /
LIM, fol. 6v2, línea 20	
transcripción del modelo	fue mujer en tie<n>pos possados de
transcripción final en OSTA	fue mujer en tie<n>pos p(o)[a]ssados
LIM, fol. 51r2, línea 11	
transcripción del modelo	casa. y le ganera la
transcripción final en OSTA	casa. y le gan(e)[r]a la

Nota. Fuente: Elaboración propia.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

3.4. Creación del modelo

Nuestro trabajo con sistemas HTR y el modelo semipaleográfico del HSMS se origina con la transcripción de los ejemplares de las cuatro ediciones de las *Siete Partidas* impresas entre diciembre de 1491 y 1542 (Fradejas Rueda, 2024) en letra gótica: 1491 (Sevilla: Cuatro Compañeros Alemanes), 1501 (Venecia: Lucantonio de Giunta), 1528 (Venecia: Gregorio de Gregorii) y 1542 (Alcalá de Henares: Juan de Brocar)[37]. Esta necesidad surgió al constatar que la edición de Gregorio López de 1555 (Salamanca: Andrea de Portonariis), que se transcribió manualmente[38], presentaba unos rasgos lingüísticos más medievales que los que presentaba la *princeps* de Montalvo, de 1491 (Fradejas Rueda, 2021: 241-242, 246, Tabla 7), habida cuenta que las distintas ediciones descendían unas de otras, aunque en dos momentos diferentes, 1528 (Camero Santos, 2024) y 1555, fueron objeto de revisiones con otros códices (Fradejas Rueda, 2022). En un principio se crearon modelos HTR individuales para cada una de las ediciones usando como base el modelo SpanishGothic_XV-XVI (Bazzaco, 2020). Sin embargo, la tarea de transcribir veinte folios para cada uno de los impresos y luego corregirlos para adaptarlos al sistema del HSMS y entrenarlos nos pareció lenta y repetitiva, por lo que se decidió crear un modelo general (sin sesgo temático), que pudiera cubrir el amplio espectro de los incunables castellanos.

Para crear el modelo Spanish Gothic Incunabula se eligieron veinte libros impresos (véase la Tabla 7 en el Anexo 1) de varios talleres para enseñar al sistema la variedad de los tipos góticos empleados en las imprentas peninsulares. Además, se optó por seleccionar obras de distintos géneros y temáticas: ficción (AYL, BMP, CAR, CNB), historia (JOS, VTS), tratados políticos o doctrinales (RGP, CLS), textos legales (C87, SPO, LES), libros de medicina (CUR, GER), veterinaria (ACM) y religión (AUG, AXP, ERI). Algunas son obras originales de autores castellanos (AYL, CAR, LES) y otras son traducciones del latín (JOS, CLS, ERI), del italiano (CNB, LIM) y del catalán (ACM). Se excluyeron libros cuyo contenido fuera poético debido a sus peculiaridades de composición y maquetación.

De cada uno de los veinte ejemplares seleccionados se extrajeron veinte carillas enfrentadas (vuelto y recto, Figura 14). Esta decisión vino impuesta por el hecho de que numerosas bibliotecas ofrecen sus reproducciones digitales de esta manera. Esto, por otra parte, forzó que en aquellos casos en los que las bibliotecas proporcionaron reproducciones carilla a carilla, tuviéramos que unir las imágenes de los vueltos con las de los rectos para crear una única imagen (Figura 15).

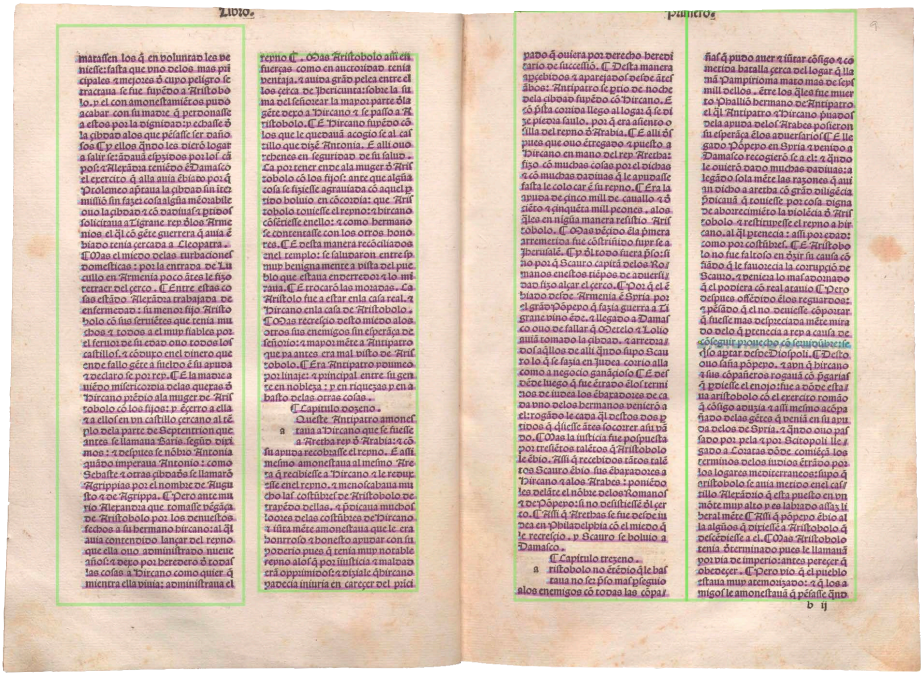
La fusión de las imágenes se logró utilizando un *script* de R con la ayuda de la librería {magick} (Ooms, 2024a). Puesto que muchas reproducciones digitales se

[37] Hay una quinta edición dentro de esta serie (Tolousse: Matthias Bonhomme, 1550), pero no la consideramos porque se imprimió en caracteres romanos o redondos.

[38] Esta transcripción fue realizada por un pequeño equipo de profesores y alumnos de la Universidad de Valladolid entre 2017 y 2019 y es accesible en Fradejas Rueda (2019).

Figura 14

Ff. 8v-9r (*sig. bv y b2r*) *de JOS*



Nota. Fuente: Elaboración propia.

distribuyen bajo el formato PDF, y dado que solo nos interesaba una pequeña selección de cada obra, se extrajeron las carillas escogidas y se convirtieron en imágenes JPG por medio de otro *script* de R escrito con la librería {pdftools} (Ooms, 2024b).

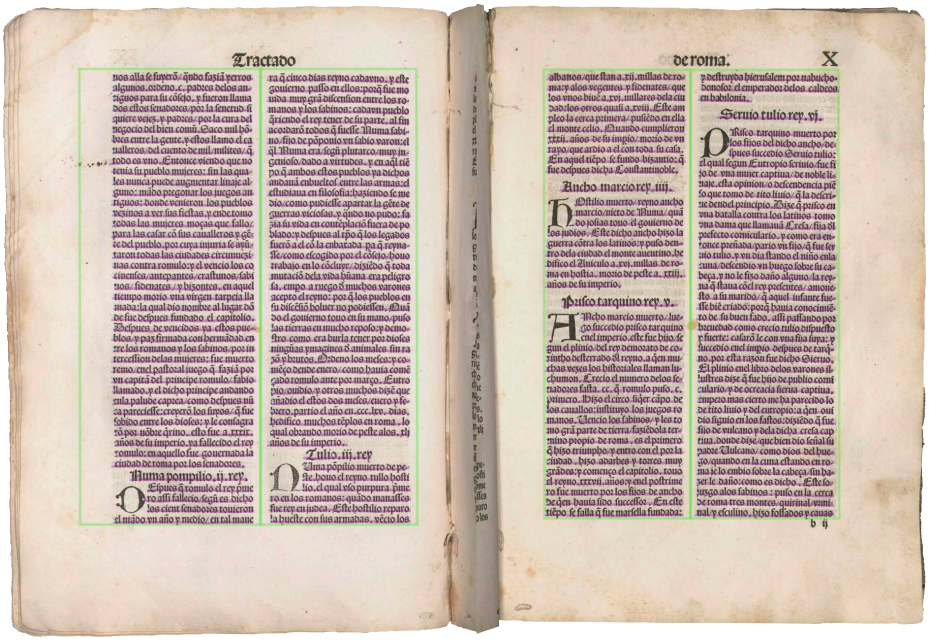
Una vez obtenidas y preparadas las imágenes de cada uno de los ejemplares que se utilizarían para la creación del modelo, se subieron a Transkribus. Se cargaron los títulos individualmente y se realizó la segmentación (*layout analysis*). Durante este proceso evitamos los títulos corrientes y signaturas porque el diseño de algunos folios hubiera requerido *retocar* en exceso la segmentación y crear zonas textuales específicas, como se muestra en la compleja maquetación recogida en la Figura 9, que en nada habrían ayudado en el entrenamiento. Además, en algunos casos las cifras romanas no están impresas con tipos góticos (Figura 15). Tampoco se tuvo en consideración las letras de recuerdo usadas por los impresores.

Una vez realizada la segmentación, se procedió a la transcripción de la selección de folios (vuelto-recto) de cada una de las obras. Para esta parte del trabajo se siguió un doble proceso. Algunas obras, como LES, RGP, AYL, C87, ERI y CNB, se transcribieron primero usando el modelo SpanishGothic_XV-XVI, puesto que ofrece

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Figura 15

Ff. 7v y 8r de VTS



Nota. Fuente: Elaboración propia.

resultados excelentes, como se ha mostrado en la comparativa de la Figura 11. Este modelo, como se ha indicado, proporciona una transcripción normalizada, por lo que el resultado fue revisado manualmente, y se marcó el desarrollo de las abreviaturas[39], se corrigieron los problemas de unión y separación de palabras cuando esta no era conforme al original y se revirtió al signo tironiano en aquellos casos en que la propuesta de Transkribus hubiera sido *y*. La segunda forma de abordar la tarea fue usando transcripciones de textos que el HSMS ya tenía en su base de datos (SPO,

[39] Como referimos antes, en el desarrollo de las abreviaturas, y para evitar la aparición de errores recursivos, se entrenó el modelo para que desarrollara toda nasal ante *p* y *b* como <n>, aun cuando el impresor tuviera una marcada preferencia por usar *m* en dicha posición en formas plenas. Así, por ejemplo, en JOS el impresor tuvo una mayor predilección por *-mp-* (1394 casos) y *-mb-* (1109 casos) frente a tan solo 12 casos de *-nb-* y 4 de *-nb-*, por lo que de los 1053 casos en los que la nasal está abreviada, en la revisión posterior a la transcripción, los transcriptores originales establecieron que el valor de la lineta era <m>, aunque en siete casos la transcribieron como <n>. En SPO, por el contrario, hay una marcada tendencia a imprimir *-np-* (4175 casos) y *-nb-* (5766) en vez de *-mp-* (63 casos) y *-mb-* (29 casos), por lo que los transcriptores establecieron que la nasal debía ser <n>, aunque quedó un caso de <m>.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

AXP, ACM, LIM, AUG, VTS y BMP), revisándolas y eliminando todas las etiquetas no pertinentes[40].

Una vez realizadas todas las transcripciones, las imágenes se reunieron en un único fichero con un total de doscientas imágenes, que correspondían a cuatrocientas carillas, procedentes de las veinte obras seleccionadas. Algunas tenían una disposición a dos columnas (SPO, CNB, LES, RGP), mientras que otras a una sola columna (AYL, C87, ERI, AXP, CAR y APL). Un caso especial es el de CLS, que tiene un complejo diseño de un bloque de textos central y glosas en los márgenes (Figura 9). De este ejemplar se seleccionaron diez hojas cuya disposición se podía *reducir* a una presentación en dos columnas.

Este documento máster está compuesto por 200 491 palabras, distribuidas a lo largo de 26 737 líneas. Se utilizaron para el *train set* 180 imágenes, con un total de 180 158 palabras y 24 060 líneas, mientras que el *validation set* se construyó con 10% de las imágenes, lo que supone 20 imágenes, que contenían 20 333 palabras distribuidas a lo largo de 2677 líneas.

Aunque se indicó a Transkribus que los ciclos (*epochs*[41]) de entrenamiento deseados eran 250, con una parada temprana (*early stopping*) de 100 ciclos, el aplicativo llegó a una tasa de CER estable después de 187 ciclos, con lo que el entrenamiento se detuvo 20 ciclos después, alcanzando 207 ciclos de entrenamiento. El entrenamiento tardó 16 horas, 32 minutos y 26 segundos. El resultado fue un modelo HTR con un CER de 0,20% en el *train set*, un CER de 0,77% en el *validation set* y un WER de 3,21%, es decir, la tasa de éxito es de 99,23% en lo que respecta a los caracteres individuales y en cuanto a las palabras correctas es de 96,79%.

3.5. Evaluación

Ante estos resultados, hicimos dos pruebas con textos sobre los que el modelo no había sido entrenado. Para ello, buscamos en la BDH dos incunables de poca extensión: los INC1243(2) e INC/2559. El INC1243(2) no indica la ciudad, el taller o el año en que fue impreso, pero según los catálogos de incunables parece haber sido impreso en Sevilla (Meinardo Ungut y Estanislao Polono) en 1492 (Martín Abad y Moyano Andrés, 2002: 104, n.º 13); se trata de unas ordenanzas dadas por los Reyes Católicos a la ciudad de Sevilla el 30 de mayo de 1492. El segundo, lo acabó de

[40] Transkribus ofrece la herramienta Text2Image (https://readcoop.eu/transkribus/docu/text2image/), por medio de la que se puede emparejar una imagen con un texto transcrito previamente. Sin embargo, nos pareció que era un proceso ligeramente más complicado que el que hemos ideado: limpiar las marcas HSMS no necesarias, y cortar y pegar en el editor de Transkribus, lo que permitía, además, detectar errores en la segmentación (líneas fantasma causadas por las iniciales decoradas).

[41] En las redes neuronales la *epoch* es un hiperparámetro que define el número de veces que un algoritmo de aprendizaje analizará y estudiará todo el set de entrenamiento.

imprimir Estanislao Polono[42] el 26 de noviembre de 1500 y es una ordenanza sobre tejidos y tejedores sevillanos (Martín Abad y Moyano Andrés, 2002: 137, n.º 81).

Estos dos ejemplares, descargados de la BNE en formato PDF, se descompusieron en ficheros individuales, eliminándose las hojas de guarda, y se exportaron a imágenes JPG con la ayuda de un *script* en R realizado con la librería {pdftools}. Una vez obtenidas las imágenes JPG y subidas a los servidores de Transkribus, se segmentó el texto y se realizó la transcripción automática con el modelo Spanish Gothic Incunabula (HSMS). Después de la transcripción, ambos textos fueron revisados manualmente y se pidió a Transkribus que comparara el texto transcrito por el sistema —Hypothesis (HTR Text)— con el texto corregido manualmente —Reference (Correct Text)— [43]. El resultado del análisis ha sido sorprendentemente mejor de lo que se esperaba con las tasas de error ofrecidas en el resumen del entrenamiento.

En el caso del INC/1243(2), que es un folleto de diez hojas, impreso a una sola columna con 9631 palabras distribuidas a lo largo de 734 líneas, el CER ha sido de 0,32 % y el WER de 1,93 %, mientras que el caso del INC/2559, que tiene ocho hojas impresas a una sola columna y con 8509 palabras a lo largo de 580 líneas, ha sido ligeramente peor, pues el CER ha sido de 0,63% y el WER de 2,68 %. En cualquier caso, estas cifras están por debajo del CER y WER iniciales del modelo (Tabla 6).

Una tercera prueba, independiente, la realizó Francisco Gago Jover con el INC/2674(1) de la BNE. Este ejemplar reúne dos características que lo convierten en un excelente banco de pruebas. Por un lado, fue impreso en el segundo semestre de 1488 y es el único superviviente del taller que Alfonso Fernández de Córdoba estableció en Híjar (Zaragoza)[44], por lo que es una imprenta no tenida en cuenta a la hora de construir el modelo. Por el otro, es un texto con una fortísima influencia de la lengua aragonesa.

Este libro, que también es muy breve, con 17 hojas a una sola columna (salvo la última, que presenta una parte a dos columnas) y 13 404 palabras a lo largo de 1184 líneas, arrojó un CER de 0,86% y un WER de 2,99 %. Ciertamente el CER es 0,09 % superior al que el entrenamiento del modelo ofreció, que era de 0,77 %. Sin embargo, el WER es ligeramente mejor, un 0,21 % menos.

[42] Meinardo Ungut murió el 12 de noviembre de 1499. Desde ese momento, el taller fue responsabilidad única de Estanislao Polono.

[43] La comparación se hace por medio de la herramienta Compute Accuracy > Compare, en la pestaña Tools del Transkribus Expert Client (v.1.28.0).

[44] De esta obra, según BETA (manid 4447), podrían existir dos ejemplares, el de la BNE y otro reportado en la Biblioteca Municipal de Zaragoza (BETA copid 8699), pero no recogido por los catálogos actuales.

Tabla 6

Comparación entre el texto propuesto y el corregido manualmente con el modelo HTR Spanish Gothic Incunabula

	Modelo	INC/1243(2)	INC/2559	INC/2674(1)
CER	0,77%	0,32%	0,63%	0,86%
WER	3,21%	1,93%	2,68%	2,99%

Nota. Fuente: Elaboración propia.

4. Conclusiones

Nuestra experiencia con los sistemas y plataformas de HTR demuestra que son una herramienta altamente fiable para la transcripción de incunables e impresos antiguos —y manuscritos, para lo que se diseñaron en un principio—. El entrenamiento de un modelo HTR es un proceso relativamente sencillo, pero requiere una minuciosa labor de selección y preparación de los materiales sobre los que se llevará a cabo: tanto de las obras elegidas, como de sus reproducciones digitales y la preparación de los ficheros *ground truth*. En nuestro caso, este conjunto estuvo compuestos por 20 páginas (10 hojas) de veinte obras de temáticas diferentes, con distintas disposiciones textuales (*mise-en-page*) y una muestra significativa de los tipos góticos empleados por las imprentas de incunables peninsulares. Esta variedad permitió que el modelo se entrenase sobre un elevado número de posibles formas abreviadas (aunque es una realidad estadística que no todas las formas posibles aparecerán en el conjunto de entrenamiento).

Queremos insistir en uno de los aspectos discutidos en el artículo: un punto básico a la hora de abordar el diseño de un modelo HTR es tener claro cuál será el objetivo final para el que se crea y aceptar que las transcripciones resultantes son representaciones que *intentan* acercarse al original, pero nunca son perfectas y requieren, además de la revisión automatizada, de una corrección supervisada (humana). Todo proceso de transcripción, incluso aquellos realizados empleando los sistemas de transcripción más conservadores, como las llamadas transcripciones (o ediciones) alográficas, facsimilares e incluso hiperdiplomáticas, suelen sacrificar *micro-features* del texto. A pesar de los alegatos ofrecidos por los usuarios y defensores de un sistema u otro, toda transcripción involucra ciertos sacrificios a la fidelidad textual, alterando el *original*, para crear un *sustituto digital*.

Los modelos HTR son sistemas probabilísticos cuyo resultado es la hipótesis de una transcripción. Por este motivo, las transcripciones que se crean con ellos *nunca* son 100% correctas; siempre hay lugares en los que el modelo yerra porque, como hemos mostrado, no conoce una forma abreviada, no tuvo material suficiente

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

durante el entrenamiento, le despista una mancha en la reproducción que se ha utilizado o las líneas están muy juntas y no logra leer una marca superescrita y la confunde con un trazo de una letra de la línea anterior.

Ya que el objetivo del modelo Spanish Gothic Incunabula (HSMS) es incorporar los textos de los incunables españoles a un corpus lingüístico (donde podrá ser usado por otros investigadores como fuente para ediciones o estudios lingüísticos), nos hemos inclinado por el uso de transcripciones semipaleográficas etiquetadas según el modelo diseñado en los años 1970 por el Hispanic Seminary of Medieval Studies. Este modelo codifica toda la información relativa a la *mise-en-page*, las erratas, errores y enmiendas de los escribas (en el caso de manuscritos), y ciertos rasgos gráficos como la presencia de letras superescritas y cómo se han desarrollado las abreviaturas. No hemos creído pertinente, sin embargo, conservar la información referente a los alógrafos que son meras variantes gráficas de una misma letra (*s* frente *ſ*; *r* frente a *ꝛ*; *d* frente ꝺ) que, aunque pueden ser elementos para la datación de los manuscritos o determinar los impresores, son difíciles de representar en el mundo digital, y de accesibilidad y permanencia mucho más compleja. Además, como se ha mostrado, la posibilidad de no ser absolutamente rígido en la marcación de ciertos alógrafos abreviativos (*ver-* y *vir-*), ofrece mayores seguridades que desarrollar equivocadamente una abreviatura.

Financiación

Este trabajo forma parte de los resultados del proyecto *7PartidasDigital* (referencia PID2020-112621GB-I00/AEI/10.13039/501100011033) cuyo objetivo es la edición crítica digital de las *Siete Partidas*. El proyecto *7PartidasDigital* (https://7partidas.hypotheses.org/), que se desarrolla desde la Universidad de Valladolid, cuenta con la financiación de la Agencia Estatal de Investigación, Ministerio de Ciencia e Innovación. El trabajo también se enmarca en las actividades de la ayuda Juan de la Cierva Formación (FJC2021-047096-I) financiada por MCIN/AEI y por la Unión Europea (NextGenerationEU/PRTR).

Contribución de autoría

- Concepción: JMFR.
- Redacción del borrador: JMFR y MACO.
- Diseño y entrenamiento del modelo: JMFR.
- Revisión final del artículo: MACO.

Anexo 1

La Tabla 7 recoge los datos editoriales de los incunables empleados para el entrenamiento del modelo Spanish Gothic Incunabula y la Tabla 8 los datos bibliográficos de las copias específicas usadas, incluyendo referencias al *manid* y *copid* en BETA y a la ficha de obra del *Catalogue of Medieval Works Printed in Castilian* (Comedic) de la Universidad de Zaragoza. El asterisco (*) indica que el impreso sobrevive en un *unicum*. Todas las siglas adoptadas para los manuscritos corresponden con el sistema empleado en OSTA.

Tabla 7

Incunables empleados para el modelo Spanish Gothic Incunabula (*datos editoriales*)

Sigla	Título	Autor	Ciudad	Impresor	Fecha de impresión
C87	*Doctrinal de los cavalleros*	Alfonso de Cartagena	Burgos	Fadrique de Basilea	1487-06-20
APL*	*Vida e historia del rey Apolonio*	-	Zaragoza	Pablo Hurus	ca. 1488
ERI	*Epistoles de rabbí Samuel*	Alfonso Buenhombe (trad. atrib.)	Zaragoza	Pablo Hurus	ca.1490
SVH	*Espejo de la vida humana*	Rodrigo Sánchez de Arévalo	Zaragoza	Pablo Hurus	1491-05-13
CLS	*Cinco libros de Séneca*	Lucio Anneo Séneca; Alfonso de Cartagena (trad.)	Sevilla	Meinardo Ungut y Estanislao Polono	1491-05-28
SPO	*Siete Partidas*	Alfonso X	Sevilla	Meinardo Ungut y Estanislao Polono	1491-10-25
AYL*	*Arnalte e Lucenda*	Diego de San Pedro	Burgos	Fadrique de Basilea	1491-11-25
CAR*	*Cárcel de amor*	Diego de San Pedro	Sevilla	Cuatro Compañeros Alemanes	1492-03-03
JOS	*Siete libros de la guerra judaica. Contra Apión*	Flavio Josefo; Alfonso de Palencia (trad.)	Sevilla	Meinardo Ungut y Estanislao Polono	1492-03-27

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Sigla	Título	Autor	Ciudad	Impresor	Fecha de impresión
AUG	*Infancia Salvatoris*	Bernardo de Caravaca (atrib.)	Burgos	Juan de Burgos	ca.1493
GEN	*Tratado de la generación de la criatura*	-	Pamplona	Arnaud Guillén de Brocar	1494-10-10
RGP	*Regimiento de príncipes*	Egidio Romano	Sevilla	Meinardo Ungut y Estanislao Polono	1494-10-20
LIM	*De las mujeres ilustres*	Giovanni Boccaccio	Zaragoza	Pablo Hurus	1494-10-24
CNB*	*Las ciento novelas*	Giovanni Boccaccio	Sevilla	Meinardo Ungut y Estanislao Polono	1496-11-08
AXP	*Libro del Anticristo*	-	Burgos	Fadrique de Basilea	1497
VTS	*Viaje de la Tierra Santa*	Bernardo de Breidenbach; Martín Martínez de Ampiés (trad.)	Zaragoza	Pablo Hurus	1498-01-16
BMP*	*Baladro del sabio Merlín*	-	Burgos	Juan de Burgos	1498-02-10
CUR	*Cura de la piedra y dolor de la hijada y cólica renal*	Julián Gutiérrez de Toledo	Toledo	Pedro Hagenbach	1498-04-04
LES	*Leyes del estilo*	-	Burgos	Fadrique de Basilea	1498-07-30
ACM	*Libro de albeitería*	Manuel Díez; Martín Martínez de Ampiés (trad.)	Zaragoza	Pablo Hurus	1499-10-16

Nota. Fuente: Elaboración propia.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Tabla 8

Incunables empleados para el modelo Spanish Gothic Incunabula (datos bibliográficos)

Sigla	BETA manid	BETA copid	Ficha Comedic	Ciudad	Biblioteca	Signatura
C87	1664	2313	82	Madrid	Biblioteca Nacional de España	INC/1910
APL*	1381	-	187	Nueva York	Hispanic Society of America	Inc 146
ERI	4437	-	188	El Escorial	Real Biblioteca del Monasterio de San Lorenzo de El Escorial	b-IV-29(1)
SVH	1588	-	177	Madrid	Biblioteca Nacional de España	INC/2329(1)
CLS	1694	2242	-	Madrid	Biblioteca Nacional de España	INC/2564
SPO	1119	1031	-	México D.F.	Universidad Nacional Autónoma	KKT140 S54 1491
AYL*	2135	-	91	Madrid	Real Academia de la Historia	Inc.153
CAR*	2133	-	60	Madrid	Biblioteca Nacional de España	INC/2134
JOS	2051	-	262	Madrid	Biblioteca Nacional de España	158
AUG	2325	-	108	Madrid	Biblioteca Nacional de España	1424
GEN	2649		-	Madrid	Biblioteca Nacional de España	INC/1335
RGP	1807	1375	49	Valladolid	Biblioteca Histórica Santa Cruz	U/Bc IyR 078
LIM	1511	2278	19	Madrid	Biblioteca Nacional de España	INC/1354
CNB*	3338	-	42	Bruselas	Bibliothèque royale de Belgique	INC B 399 (RP)

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Sigla	BETA manid	BETA copid	Ficha Comedic	Ciudad	Biblioteca	Signatura
AXP	2322	1837	97	Madrid	Biblioteca Nacional de España	INC/543
VTS	1967	1508	206	Stuttgart	Württembergische Landesbibliothek	Inc.fol.3965
BMP*	1196	-	110	Oviedo	Biblioteca Central, Universidad	CEA-304
CUR	1852	1417	-	Madrid	Biblioteca Histórica Marqués de Valdecilla	INC M-29
LES	2430	1911	-	Madrid	Real Biblioteca del Palacio Real	I/108
ACM	2359	-	-	Madrid	Biblioteca Nacional de España	INC/2342

Nota. Cuando no se recoge el *copid* de BETA, el ejemplar está descrito en el manid. Fuente: Elaboración propia.

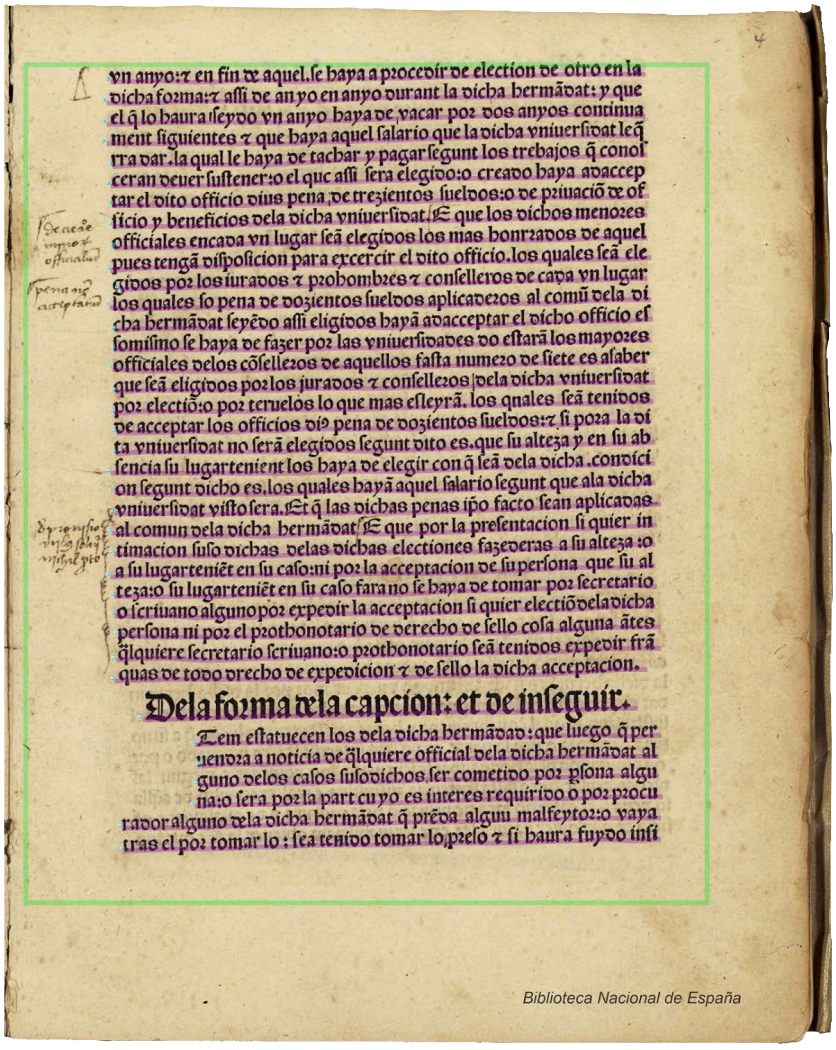
ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Anexo 2

En este anexo ofrecemos una comparación entre una transcripción inicial (hipótesis) realizada por Transkribus con el modelo Spanish Gothic Incunabula, la transcripción corregida del texto y el texto codificado con el sistema completo del HSMS para incorporarse a OSTA. El texto elegido viene del f. 4r del BNE INC/2674(1) (Figura 16).

Figura 16

F. 4r del BNE INC/2674(1) tras la segmentación



Nota. Fuente: Elaboración propia.

Ya que Transkribus guarda una copia de todas las versiones por la que ha pasado un texto desde el momento en que se carga en sus servidores, le pedimos que, por medio de la herramienta Compute Accuracy, comparara el resultado de la decodificación que hizo Transkribus y la corrección del texto.

En total, hemos detectado siete errores: tres de ellos son erratas del impresor; dos son desarrollos incorrectos, pues la forma lingüística no está en el vocabulario del set de entrenamiento del modelo y, al estar en latín, no siguen las reglas estadísticas que rigen la creación de palabras en castellano desarrolladas por Transkribus; y dos son errores de descodificación.

Las erratas del cajista que el modelo ha leído correctamente son *quc* > *que* (línea 6), *qnal* > *qual* (línea 17) y *alguu* > *algun* (línea 36)[45]. Los errores de descodificación son *electio* > *electiõ* (línea 17), debido a que la lineta ha sido oscurecida por el trazo descendente de la *g* encima[46] (mientras que en la línea 27 fue correctamente transcrita), *el* > *et* y - > . (línea 31). Los dos errores debidos al modelo son: *di*^{9} (línea 18) y *ip̄o* (línea 22). Como dijimos, ambas son abreviaturas latinas que no están en el *diccionario* de este modelo, desarrollado específicamente para textos romances (del iberorrománico central).

Hipótesis de Transkribus[47]

1. vn anyo: & en fin de aquel. se haya a procedir de election de otro en la
2. dicha forma: & assi de anyo en anyo durant la dicha herma<n>dat: y que
3. el q<ue> lo haura seydo vn anyo haya de vacar por dos anyos continua
4. ment siguientes & que haya aquel salario que la dicha vniuersidat le q<ue>-
5. rra dar. la qual le haya de tachar y pagar segunt los trebajos q<ue> conos-
6. ceran deuer sustener: o el que assi sera elegido: o creado haya adacceptar
7. tar el dito officio dius pena de trezientos sueldos: o de priuacio<n> de of-
8. sicio y beneficios dela dicha vniuersidat. E que los dichos menores
9. officiales encada vn lugar sea<n> elegidos los mas honrrados de aquel
10. pues tenga<n> disposicion para excercir el dito officio. los quales sea<n> ele-
11. gidos por los iurados & prohombres & conselleros de cada vn lugar
12. los quales so pena de dozientos sueldos aplicaderos al comu<n> dela di-
13. cha herma<n>dat seye<n>do assi eligidos haya<n> adacceptar el dicho officio es-

45 Esto no lo ha contabilizado Transkribus como error, pues coincide con la forma correcta (aunque sea, en verdad, incorrecta).

46 Hemos detectado el mismo problema en varios textos: si en la línea anterior hay una letra con un trazo descendente, superpuesta a una lineta abreviativa en la siguiente, la decodificación será errónea pues el modelo ignorará la lineta. Además, la manchas, inevitables en el papel antiguo, pueden ser interpretadas incorrectamente como linetas abreviativas (y desarrolladas según las reglas desarrolladas por el modelo HTR).

47 La numeración de líneas se ha añadido para mayor facilidad de lectura y comparación, no las introduce Transkribus ni se incorporarán en las transcripciones que se ofrecen a continuación.

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

somismo se haya de fazer por las vniuersidades do estara<n> los mayores
officiales delos co<n>selleros de aquellos fasta numero de siete es asaber
que sea<n> eligidos por los jurados & conselleros dela dicha vniuersidat
por electio: o por teruelos lo que mas esleyra<n>. los quales sea<n> tenidos
de acceptar los officios dia pena de dozientos sueldos: & si pora la di-
ta vniuersidat no sera<n> elegidos segunt dito es. que su alteza y en su ab-
sencia su lugartenient los haya de elegir con q<ue> sea<n> dela dicha. condici-
on segunt dicho es. los quales haya<n> aquel salario segunt que ala dicha
vniuersidat visto sera. Et q<ue> las dichas penas i<ist>o facto sean aplicadas.
al comun dela dicha herma<n>dat E que por la presentacion si quier in-
timacion suso dichas delas dichas ellectiones fazederas a su alteza: o
a su lugartenie<n>t en su caso: ni por la acceptacion de su persona que su al-
teza: o su lugartenie<n>t en su caso fara no se haya de tomar por secretario
o scriuano alguno por expedir la acceptacion si quier electio<n> dela dicha
persona ni por el prothonotario de derecho de sello cosa alguna a<n>tes
q<u>a`lquiere secretario scriuano: o prothonotario sea<n> tenidos expe-
dir fra<n>-
quas de todo drecho de expedicion & de sello la dicha acceptacion.
Dela forma dela capcion: el de inseguir-
Tem estatuecen los dela dicha herma<n>dad: que luego q<ue> per-
uendra a noticia de q<u>a`lquiere official dela dicha herma<n>dat al-
guno delos casos susodichos ser cometido por p<er>sona algu-
na: o sera por la part cuyo es interes requirido o por procu-
rador alguno dela dicha herma<n>dat q<ue> pre<n>da algun malfeytor: o vaya
tras el por tomar lo: sea tenido tomar lo preso & si haura fuydo insi

Texto corregido

vn anyo: & en fin de aquel. se haya a procedir de election de otro en la
dicha forma: & assi de anyo en anyo durant la dicha herma<n>dat: y que
el q<ue> lo haura seydo vn anyo haya de vacar por dos anyos continua
ment siguientes & que haya aquel salario que la dicha vniuersidat le q<ue>-
rra dar. la qual le haya de tachar y pagar segunt los trebajos q<ue> conos-
ceran deuer sustener: o el que assi sera elegido: o creado haya adaccep-
tar el dito officio dius pena de trezientos sueldos: o de priuacio<n> de of-
ficio y beneficios dela dicha vniuersidat. E que los dichos menores
officiales encada vn lugar sea<n> elegidos los mas honrrados de aquel
pues tenga<n> disposicion para excercir el dito officio. los quales sea<n> ele-
gidos por los iurados & prohombres & conselleros de cada vn lugar
los quales so pena de dozientos sueldos aplicaderos al comu<n> dela di-

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

13. cha herma<n>dat seye<n>do assi eligidos haya<n> adacceptar el dicho officio es-
14. so mismo se haya de fazer por las vniuersidades do estara<n> los mayores
15. officiales delos co<n>selleros de aquellos fasta numero de siete es asaber
16. que sea<n> eligidos por los jurados & conselleros dela dicha vniuersidat
17. por electio<n>: o por teruelos lo que mas esleyra<n>. los quales sea<n> tenidos
18. de acceptar los officios di<us> pena de dozientos sueldos: & si pora la di-
19. ta vniuersidat no sera<n> elegidos segunt dito es. que su alteza y en su ab-
20. sencia su lugartenient los haya de elegir con q<ue> sea<n> dela dicha. condici-
21. on segunt dicho es. los quales haya<n> aquel salario segunt que ala dicha
22. vniuersidat visto sera. Et q<ue> las dichas penas ip<s>o facto sean aplicadas.
23. al comun dela dicha herma<n>dat E que por la presentacion si quier in-
24. timacion suso dichas delas dichas ellectiones fazederas a su alteza: o
25. a su lugartenie<n>t en su caso: ni por la acceptacion de su persona que su al-
26. teza: o su lugartenie<n>t en su caso fara no se haya de tomar por secretario
27. o scriuano alguno por expedir la acceptacion si quier electio<n> dela dicha
28. persona ni por el prothonotario de derecho de sello cosa alguna a<n>tes
29. q<u>a`lquiere secretario scriuano: o prothonotario sea<n> tenidos expe-dir fra<n>-
30. quas de todo drecho de expedicion & de sello la dicha acceptacion.
31. Dela forma dela capcion: et de inseguir.
32. Tem estatuecen los dela dicha herma<n>dad: que luego q<ue> per-
33. uendra a noticia de q<u>a`lquiere official dela dicha herma<n>dat al-
34. guno delos casos susodichos ser cometido por p<er>sona algu-
35. na: o sera por la part cuyo es interes requirido o por procu-
36. rador alguno dela dicha herma<n>dat q<ue> pre<n>da algun malfeytor: o vaya
37. tras el por tomar lo: sea tenido tomar lo preso & si haura fuydo insi

Texto preparado para OSTA, según las normas de transcripción del HSMS[48]

[f. 3r]
{CB1.

1. vn anyo: & en fin de aquel. se haya a procedir de election de otro en la
2. dicha forma: & assi de anyo en anyo durant la dicha herma<n>dat: y que
3. el q<ue> lo haura seydo vn anyo haya de vacar por dos anyos continua
4. ment siguientes & que haya aquel salario que la dicha vniuersidat le q<ue>-

[48] Las transcripciones según las normas del HSMS no numeran las líneas. En este caso se han numerado para facilitar la comparación entre las tres transcripciones.

rra dar. la qual le haya de tachar y pagar segunt los trebajos q<ue> conos-
ceran deuer sustener: o el qu(c)[e] assi sera elegido: o creado haya adacceptar el dito officio dius pena de trezientos sueldos: o de priuacio<n> de of-
(s)[f]icio y beneficios dela dicha vniuersidat. E que los dichos menores
officiales encada vn lugar sea<n> elegidos los mas honrrados de aquel
pues tenga<n> disposicion para excercir el dito officio. los quales sea<n> ele-
gidos por los iurados & prohombres & conselleros de cada vn lugar
los quales so pena de dozientos sueldos aplicaderos al comu<n> dela dicha herma<n>dat seye<n>do assi eligidos haya<n> adacceptar el dicho officio es-
so mismo se haya de fazer por las vniuersidades do estara<n> los mayores
officiales delos co<n>selleros de aquellos fasta numero de siete es asaber
que sea<n> eligidos por los jurados & conselleros dela dicha vniuersidat
por electio<n>: o por teruelos lo que mas esleyra<n>. los q($u)[u]ales sea<n> tenidos
de acceptar los officios di<us> pena de dozientos sueldos: & si pora la dita vniuersidat no sera<n> elegidos segunt dito es. que su alteza y en su absencia su lugartenient los haya de elegir con q<ue> sea<n> dela dicha. condici-
on segunt dicho es. los quales haya<n> aquel salario segunt que ala dicha
vniuersidat visto sera. Et q<ue> las dichas penas {LAT. ip<s>o facto} sean aplicadas.
al comun dela dicha herma<n>dat E que por la presentacion si quier intimacion suso dichas delas dichas ellectiones fazederas a su alteza: o
a su lugartenie<n>t en su caso: ni por la acceptacion de su persona que su al-
teza: o su lugartenie<n>t en su caso fara no se haya de tomar por secretario
o scriuano alguno por expedir la acceptacion si quier electio<n> dela dicha
persona ni por el prothonotario de derecho de sello cosa alguna a<n>tes
q<u><<a>>lquiere secretario scriuano: o prothonotario sea<n> tenidos expedir fra<n>-
quas de todo drecho de expedicion & de sello la dicha acceptacion.
{RUB. Dela forma dela capcion: et de inseguir.}
{IN4.} [I]Tem estatuecen los dela dicha herma<n>dad: que luego q<ue> per-
uendra a noticia de q<u><<a>>lquiere official dela dicha herma<n>dat alguno delos casos susodichos ser cometido por p<er>sona algu-
na: o sera por la part cuyo es interes requirido o por procurador alguno dela dicha herma<n>dat q<ue> pre<n>da algu($n)[n] malfeytor: o vaya
tras el por tomar lo: sea tenido tomar lo preso & si haura fuydo insi}

ISSN 1132-0265
https://dx.doi.org/10.12795/PH.2025.v39.i02.06

Postscript

Durante el proceso de evaluación y aceptación de estas páginas, constatamos que algunos impresores utilizaban un tipo diferente para la <d> recta en interior y final de palabra, cuando lo normal era una <d> uncial. Al procesar alguno de estos textos en Transkribus usando el model ID 216053 —Spanish Gothic Incunabula (HSMS)—, esta <d> recta aparecía desarrollada de formas diversas y extrañas (<c>, <dl>, <ol>, <g>, <q>, <qui`>, <ql>), lo que reducía la utilidad del modelo, sin ser un error sistemático, pues el uso del tipo se restringe a un grupo reducido de impresores que compartieron el uso de tipo gótico, pero con el paso del tiempo aumenta el número de talleres que lo emplean. Por este motivo se procedió a la ampliación del modelo y su reentrenamiento. Para ello se han añadido treinta nuevas imágenes (vuelto, recto) tomadas de varias ediciones en letra gótica producidas entre 1514 y 1559:

- JER, Historia de San Jerónimo. Zaragoza: Jorge Cocci, 1514. BNE R/10893.
- TP24, Tesoro de los pobres. Burgos: Alonso de Melgar, 1524. BNE R/13136
- P57, Alcalá de Henares: Salcedo, 1557. BNE R/15431(6) [cinco hojas]
- P59, Valladolid: Sebastián Martínez, 1559. BNE R/15431(3) [cinco hojas]

El nuevo modelo se ha rebautizado Spanish Gothic Print v2 (HSMS), model ID 338253 y es público.

Referencias

Bazzaco, S. (2020). El reconocimiento automático de textos en letra gótica del Siglo de Oro: Creación de un modelo HTR basado en libros de caballerías del siglo xvi en la plataforma Transkribus. *Janus*, (9), 534-561. http://hdl.handle.net/2183/27389 https://doi.org/10.17979/janus.2020.0.09.10398

Bazzaco, S. (2024). Revolucionar el acceso al patrimonio librario: Los sistemas de HTR entre humanidades y ciencias de la información. *Philologia Hispalensis*, *38*(2), 59-77. https://doi.org/10.12795/PH.2024.v38.i02.03

Bermúdez Sabel, H. (2022). L'édition numérique au service de la philologie matérielle. Modèles de la lyrique galégo-portugaise. *Arquivo Galicia Medieval, 5*, 11-30. https://libra.unine.ch/handle/123456789/30074

Buelow, K. y Mackenzie, D. (1977). *A Manual of Manuscript Transcription for the Dictionary of Old Spanish Language*. Hispanic Seminary of Medieval Studies.

Camero Santos, E. (2024). Post-incunables e IA: la transcripción automática de un ejemplar de la edición de 1528 de las *Partidas* y su posterior tratamiento. En M. J. Lop Otín, D. Igual Luis y J. Pérez Burgueño (Eds.), *Alfonso X: el universo político y cultural de un reinado* (pp. 191-198). Universidad de Castilla-La Mancha.

Camps, J.-B. (2017). *La* Chanson d'Otinel. *Édition complète du corpus manuscrit et prolégomènes à l'édition critique*, thèse de doctorat préparée sous la direction de M. Dominique Boutet,

soutenue le 3 décembre 2016 à l'université Paris-Sorbonne. *Perspectives médiévales*, (38). https://doi.org/10.4000/peme.13004

Camps, J. B. (2021). La Philologie computationnelle à l'École des chartes. Premier bilan et perspectives. *Bibliothèque de l'École des chartes, 176*, 1-24. https://enc.hal.science/hal-03716538v1

Catach, N. (1990). Französisch: Graphetik und Graphemic. En G. Holtus, M. Metzeltin y Ch. Schmitt (Eds.), *Lexikon der Romanistischen Linguistik. Vol. I/1: Geschichte des Faches Romanistik. Methodologie* (*Das Sprachsystem*) (pp. 46-58). De Gruyter.

Causer, T., Grint, K., Sichani, A. y Terra, M. (2018). 'Making Such Bargain': Transcribe Bentham and the Quality and Cost-Effectiveness of Crowdsourced Transcription. *Digital Scholarship in the Humanities*, *33*(3), 467-487. https://doi.org/10.1093/llc/fqx064

Chagué, A. y Clérice, T. (2023). Deploying eScriptorium Online: Notes on CREMMA's Server Specifications. *A Research* (*B*)*log*. https://inria.hal.science/hal-04362085v1

Ciula, A. (2009). The Paleographical Method Under the Light of a Digital Approach. En M. Rehbein, P. Sahle y T. Schaßan (Eds.), *Kodikologie und Paläographie in Digitalen Zeitalter* (pp. 219-235). Books on Demand.

Clérice, T., Vlachou-Efstathiou, M. y Chagué, A. (2023). CREMMA Medii Aevi: Literary Manuscript Text Recognition in Latin. *Journal of Open Humanities Data*, *9*(4), 1-19. https://doi.org/10.5334/johd.97

Donaldson, P. (1997). Shakespeare and Electronic Textuality. En K. Sutherland (Ed.), *Electronic Text: Investigations in Method and Theory* (pp. 173-198). Clarendon Press. https://doi.org/10.1093/acprof:oso/9780198236634.003.0008

Ducamin, J. (Ed.). (1901). *Juan Ruiz, Arcipreste de Hita, Libro de buen amor. Texte du XIVe siècle publié pour la première fois avec les leçons des trois manuscrits connus*. Privat.

Fafinski, M. (2022). Facsimile Narratives: Researching the Past in the Age of Digital Reproduction. *Digital Scholarship in the Humanities*, *37*(1), 94-108. https://doi.org/10.1093/llc/fqab017

Faulhaber, C. B. (Dir.). (1997). *Bibliografía española de textos antiguos* [BETA]. The Bancroft Library. University of California, Berkeley. https://philobiblon.upf.edu/html/beta_en.html

Faulhaber, C. y Marcos Marín, F. (1990). ADMYTE: Archivo digital de manuscritos y textos españoles. *La Corónica*, *18*(2), 131-145.

Fradejas Rueda, J. M. (1991). *Introducción a la edición de textos medievales castellanos*. UNED.

Fradejas Rueda, J. M. (2019). López 1555. *7PartidasDigital*. https://doi.org/10.58079/agq5

Fradejas Rueda, J. M. (2021). Las *Siete Partidas*: del pergamino a la red. En M. Albert, U. Becker y E. Schmidt (Eds.), *Conceptualización y normalización de poder y señorío en la era de Alfonso X. Las* Siete Partidas *y su contribución a la constitución teórica de la monarquía* (pp. 223-264). Bonn University Press.

Fradejas Rueda, J. M. (2022). Francisco de Velasco, segundo editor de las *Siete Partidas*. *Temas Medievales*, *30*(1), 1-17.

Fradejas Rueda, J. M. (2023). *Ex cenobio Sancti Ysidori Legionensis usque ad Bibliothecam Regiam Belgicam*: De partidas, cronicones y sermones romances. *Incipit*, *43*, 15-38. https://doi.org/10.5281/zenodo.10443037

Fradejas Rueda, J. M. (2024). Las ediciones históricas de las *Siete Partidas*: Alonso Díaz de Montalvo y Francisco de Velasco. En M. J. Lop Otín, D. Igual Luis y J. Pérez

Burgueño (Eds.), *Alfonso X: el universo político y cultural de un reinado* (pp. 145-158). Universidad de Castilla-La Mancha.

Fradejas Rueda, J. M. (2025). <TEI o no TEI, esa es la cuestión/>. *Journal of the Text Encoding Initiative*, Selected Papers from the 2024 TEI Conference (en prensa).

Gago Jover, F. y Pueyo Mena, F. (2018a). El *Old Spanish Textual Archive*. Diseño y desarrollo de un corpus de textos medievales: lematización y etiquetado gramatical. *Scriptum digital, 7*, 25-35. https://raco.cat/index.php/scriptumdigital/article/view/343462

Gago Jover, F. y Pueyo Mena, F. (2018b). El *Old Spanish Textual Archive*. Diseño y desarrollo de un corpus de textos medievales: el corpus textual. *Cuadernos del Instituto Historia de la lengua*, (11), 165-209. https://doi.org/10.58576/cilengua.vi11.54

Gago Jover, F. y Pueyo Mena, F. (2020). *Old Spanish Textual Archive*. Hispanic Seminary of Medieval Studies. http://osta.oldspanishtextualarchive.org

Gille Levenson, M. (2023a). Towards a General Open Dataset and Model for Late Medieval Castilian Text Recognition (HTR/OCR). *Journal of Data Mining and Digital Humanities*. https://doi.org/10.46298/jdmdh.10416

Gille Levenson, M. (2023b). Towards a General Open Dataset and Model for Late Medieval Castilian Text Recognition (HTR/OCR). Datasets and Scripts (Version 2) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.8406222

Guéville, E. y Wrisley, D. J. (2024). Transcribing Medieval Manuscripts for Machine Learning. *Journal of Data Mining and Digital Humanities*. https://doi.org/10.46298/jdmdh.9805

Haugen, O. E. (2004). Parallel Views: Multi-Level Encoding of Medieval Nordic Primary Sources. *Literary and Linguistic Computing*, *19*(1), 73-91. https://doi.org/10.1093/llc/19.1.73

Haugen, O. E. (2006). On the Diplomatic Turn in Editorial Philology. En J. McKinnell, D. Ashurst y D. Kick (Eds.), *The Fantastic in Old Norse/Icelandic Literature. Sagas and the British Isles. Preprint Papers of the Thirteenth International Saga Conference, Durham and York 6th–12th August, 2006* (pp. 340-349). University of Durham.

Kahle, P., Colutto, S., Hackl, G. y Mühlberger, G. (2017). Transkribus. A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. En *2017 14th IAPR International Conference on Document Analysis and Recognition* (*ICDAR*) (pp. 19-24). https://doi.org/10.1109/ICDAR.2017.307

Kiessling, B., Tissot, R., Stokes, P. y Stökl Ben Ezra, D. (2019), eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops* (*ICDARW*). https://doi.org/10.1109/ICDARW.2019.10032

Lucía Megías, J. M. (2002). *Literatura románica en Internet. Los textos*. Castalia.

Mackenzie, D. (1994). Problemas de transcripción textual electrónica. En *Actas del congreso de la lengua española* (pp. 341-344). Instituto Cervantes.

Mancinelli, T. (2016). Early Printed Edition and OCR Techniques: What is the State-of-the-Art? Strategies to Be Developed from the Working-Progress Mambrino Project Work. *Historias fingidas*, (4), 255-260. https://historiasfingidas.dlls.univr.it/article/view/65/104

Mancinelli, T. y Pierazzo, E. (2020). *Che cos'è un'edizione scientifica digitale*. Carocci.

Marcos Marín, F. (1994). *Informática y humanidades*. Gredos.

Martín Abad, J. y Moyano Andrés, I. (2002). *Estanislao Polono*. Universidad de Alcalá de Henares.

Menéndez Pidal, R. (1901), Reseña del libro: «Juan Ruiz, Arcipreste de Hita, *Libro de buen amor*» [reseña del libro Juan Ruiz, Arcipreste de Hita, *Libro de buen*

amor de J. Ducamin]. *Romania, 30*(118-119), 434-440. https://www.persee.fr/doc/roma_0035-8029_1901_num_30_118_5215_t1_0434_0000_2

Nitti, J. (1978). Computers and the Old Spanish Dictionary. *Computers and the Humanities*, *12*(1-2), 43-52. https://doi.org/10.1007/BF02392915

Nockels, J., Gooding, P. y Terras, M. (2024). Are Digital Humanities Platforms Facilitating Sufficient Diversity in Research? A Study of the Transkribus Scholarship Programme. *Digital Scholarship in the Humanities*, *40*(Supplement 1) (i46–i65). https://doi.org/10.1093/llc/fqae018

Ooms, J. (2024a). magick: Advanced Graphics and Image-Processing in R. https://CRAN.R-project.org/package=magick

Ooms, J. (2024b). pdftools: Text Extraction, Rendering and Converting of PDF Documents. https://CRAN.R-project.org/package=pdftools

Orduna, G. (1994). La edición de textos históricos. En *Actas del congreso de la lengua española* (pp. 611-619). Instituto Cervantes.

Pierazzo, E. (2015). *Digital Scholarly Editing: Theories, Models, and Methods*. Routledge.

Reyes Gómez, F. (Ed.). (2004). *Sinodal de Aguilafuente*. Fundación Instituto Castellano y Leonés de la Lengua.

Robinson, P. M. W. (1989). The Collation and Textual Criticism of Icelandic Manuscripts (1): Collation. *Literary and Linguistic Computing*, *4*(2), 99-105. https://doi.org/10.1093/llc/4.2.99

Robinson, P. y Solopova, E. (1993). Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue. En N. F. Blake y P. Robinson (Eds.), *The Canterbury Project Occasional Papers* (pp. 19-52). Office for Humanities Communication. https://doi.org/10.5281/zenodo.11954056

Rodríguez Díaz, E. (2024). Elementos para fechar los códices castellanos y leoneses según los manuscritos datados (ss. XII y XIII). En Á. Romero Cambrón (Ed.), *La ley de los godos: estudios selectos* (pp. 125-229). Peter Lang.

Sánchez-Prieto Borja, P. (1998). *Cómo editar textos medievales. Criterios para su presentación gráfica*. Arco/Libros.

Sánchez-Prieto Borja, P. (2011). *La edición de textos medievales y clásicos. Criterios de presentación gráfica*. Cilengua.

Strauß, T., Weidemann, M. y Labahn, R. (2017). D7.11 Language Models. Improving Transcriptions by External Language Resources. En *Recognition and Enrichment of Archival Documents*. https://readcoop.eu/wp-content/uploads/2017/12/D7.11_final.pdf

Tenenbaum, F. (2000-2001). El sistema de transcripción del Hispanic Seminary of Medieval Studies (Madison, Wisconsin). *Incipit*, *20-21*, 153-168.

Terras, M., Anzinger, B., Gooding, P., Mühlberger, G., Nockels, J., Romein, C., Stauder, A. y Stauder, F. (2025). The Artificial Intelligence Cooperative: READ-COOP, Transkribus, and the Benefits of Shared Community Infrastructure for Automated Text Recognition [version 1; awaiting peer review]. *Open Research Europe*, *5*(16). Advance online publication. https://doi.org/10.12688/openreseurope.18747.1

FACULTAD DE FILOLOGÍA
UNIVERSIDAD DE SEVILLA