PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Sistemas de reconocimiento de textos e impresos hispánicos de la Edad Moderna. La creación de unos modelos de HTR para la transcripción automatizada de documentos en gótica y redonda (s. XV-XVII)

Stefano Bazzaco (Università di Verona)
Ana Milagros Jiménez Ruiz (Universidad de Zaragoza)
Mónica Martín Molares (Universidade da Coruña)
Ángela Torralba Ruberte (Universidad de Zaragoza)

Abstract

El trabajo presenta los recientes logros en el campo del reconocimiento de textos llevado a cabo en 2021 gracias a la colaboración entre los siguientes proyectos: Progetto Mambrino (Univ. de Verona), BIDISO (Univ. de A Coruña) y COMEDIC (Univ. de Zaragoza). En concreto, en la primera parte del artículo se describe el estado de la cuestión de los sistemas de transcripción automática en relación con los textos impresos de la Edad Moderna, se relatan las primeras experiencias llevadas a cabo con la plataforma Transkribus (READ Coop) y los resultados preliminares obtenidos. En la segunda parte se presentan dos modelos de HTR que consienten la transcripción automática de textos en letra gótica y redonda de la Edad Moderna (siglos XV-XVII). En dos apéndices finales se describen según las normas tipobibliográficas actuales los documentos empleados para la creación de ambos modelos.
Palabras clave: Humanidades Digitales; Transkribus (READ Coop); HTR (Handwritten Text Recognition); impresos de la Edad Moderna; Siglos de Oro

The work presents the recent achievements in the field of text recognition carried out in 2021 thanks to the collaboration between the following projects: Progetto Mambrino (Univ. of Verona), BIDISO (Univ. of A Coruña) and COMEDIC (Univ. of Zaragoza). Specifically, the first part of the article describes the state of the art of automatic transcription systems in relation to the recognition of printed texts of the Modern Age, the first experiences carried out with the Transkribus platform (READ Coop) and the preliminary results obtained. In the second part, we present two HTR models that allow the automatic transcription of early printed texts in gothic and round scripts of the Modern Age (15th-17th centuries). In two final appendices, the documents used for the creation of both models are described according to current typobibliographical standards.
Keywords: Digital Humanities; Transkribus (READ Coop); HTR (Handwritten Text Recognition); early printed documents; Siglos de Oro

Stefano Bazzaco, Ana Milagros Jiménez Ruiz, Mónica Martín Molares, Ángela Torralba Ruberte, «Sistemas de reconocimiento de textos e impresos hispánicos de la Edad Moderna. La creación de unos modelos de HTR para la transcripción automatizada de documentos en gótica y redonda (s. XV-XVII)», *Historias Fingidas*, Número Especial 1 (2022) Humanidades Digitales y estudios literarios hispánicos, pp. 67-125.
DOI: https://doi.org/10.13136/2284-2667/1190 - ISSN: 2284-2667.

Premisa[*]

En el primer capítulo del libro *From Gutenberg to Google*, Peter Shillingsburg, uno de los precursores y más importantes estudiosos del fenómeno de la migración de textos a un entorno digital, deja constancia de un problema que afecta a cualquier transmisión de prácticas de escritura en la web: la falta de fiabilidad de los contenidos informativos que se encuentran en línea. En esas provechosas páginas, de sumo interés para cualquier especialista en Humanidades, se hallan claras evidencias de un constante choque entre la tensión del filólogo hacia la reconstrucción exhaustiva del texto como acto informativo y, por otra parte, el aspecto que adquiere un texto, cualquiera que sea su forma de fijación, dentro del heterogéneo y volátil espacio del *World Wide Web*.

Para abordar la cuestión, el autor considera que cualquier editor tiene una responsabilidad compleja, fundada en la necesidad de declarar todos los aspectos relativos al trabajo editorial y a la interacción del nuevo texto con sus precedentes evolutivos, es decir, sus múltiples concreciones físicas a lo largo del tiempo. Por esta razón, la nueva edición debería guardar información fiable con respecto a dónde ha sido encontrado, qué procedimientos editoriales se le aplicaron, cuáles son las diferencias entre la obra editada y los materiales fuente, teniendo en cuenta constantemente que ningún acto editorial es una operación neutral, sino deliberadamente electiva (2006, 19 y ss.).

En opinión de Schillingsburg, todas estas cuestiones adquieren aún más importancia en la época digital, inicialmente gobernada por un clima de entusiasmo general que ha ocultado en parte las fisuras intrínsecas del proceso de migración de los textos a la red. Al respecto, según el estudioso,

[*] El presente trabajo se desarrolla en el marco de los siguientes proyectos: *Proyecto PRIN 2017 Mapping Chivalry. Spanish Romances of Chivalry from Renaissance to 21st Century: a Digital Approach* (2017JA5XAR), investigador principal Anna Bognolo, Università di Verona (2017-2023); Progetto di Eccellenza «Le Digital Humanities applicate alle lingue e letterature straniere», Università di Verona (2018-2022); Proyecto de Investigación *Catálogo de Obras Medievales Impresas en Castellano (COMEDIC)*, PID2019-104989GB-I00, financiado por MCIN/AEI/10.13039/501100011033, que se inscribe en el grupo investigador Clarisel y cuenta con la participación económica del Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón; proyecto de I+D+i *Biblioteca Digital Siglo de Oro 6* (BIDISO 6), referencia: PID2019-105673GB-I00 financiado por MCIN/AEI/10.13039/501100011033/.

los resultados obtenidos a principios del nuevo milenio por el Proyecto Gutenberg demuestran una inconsistencia de fondo, tanto que le llevan a preguntarse:

> does anyone believe that a Project Gutenberg electronic text could be relied upon to be accurate? Do these productions state accurately what the source text was? Do they describe the bibliographic features of the source text? Did the 'editors' pick as a source text one that has any sort of authority or historical importance? Did they indicate in any way how the editing or transcribing of scanning involved changed the text? (2006, 21).

Constatando que los textos del Proyecto Gutenberg que ha consultado son inservibles para su trabajo filológico, Schillingsburg declara que el trasvase de un documento a un entorno digital conlleva las mismas responsabilidades que afectan a su edición en formato analógico, como tener conciencia del texto que se está editando, de los métodos que se emplean para su transmisión y de los diferentes estadios que llevan a su nueva representación, como la colación y el *proofreading*. En ausencia de tales responsabilidades, contaríamos tan solo con ediciones imprecisas y poco útiles para cualquier propósito de investigación académica (2006, 22)[1].

Es evidente que, si extendemos la mirada a proyectos afines al Project Gutenberg, los mismos inconvenientes afectan a la mayoría de los repositorios textuales de largo alcance presentes en la red, puesto que en ellos la reproducción de obras literarias está sometida generalmente a dudosas prácticas ecdóticas. Como resultado, Internet ha llegado a configurarse como un cosmos interminable de *fast food libraries*, donde las ediciones correctas, preciadas como joyas, son muy difíciles de detectar y consultar (Italia, 2020, 38-39).

[1] «My only point at this time is that bringing a text from an old book or manuscript into the twenty-first century takes more than a computer – with or without a scanner or digital camera. It takes thoughtfulness about text, an exercise of care and good judgment about methods, and an old-fashioned devotion to sight collation and proofreading that tends to dampen enthusiasm. In the absence of these responsibilities, what will we have? Noisy texts, without any doubt. Misleading texts, very likely. Text useless for scholarly purposes, of course» (2006, 22). Estas mismas preocupaciones se encuentran tratadas de forma más general en relación con el concepto de «infoesfera» en Floridi (2014).

Son muchas las cuestiones interesantes sugeridas por Schillingsburg. Por un lado, sus reflexiones abarcan aspectos todavía por resolver, como el papel del filólogo dentro del contexto de la migración masiva de documentos que se publicaron impresos o manuscritos a la red. Por otro, sus palabras apuntan a la necesidad de rehabilitar el espacio virtual con la recreación de materiales textuales fiables, correctos y bien identificables por medio de metadatos que informen sobre los actores del proceso editorial y el texto proporcionado, por ejemplo, indicando el nombre del autor, el nombre del editor, cuál es el ejemplar transcrito, etc.

Es interesante que, con respecto a las técnicas que determinan la forma del objeto digitalizado y que han influido en su escasa fiabilidad hasta ahora, Schillingsburg insista en la exigencia de declarar cuál es la tecnología que ha sido empleada para su digitalización[2]. Al respecto, es improbable que el estudioso americano se refiera aquí a la conversión en formato imagen de los libros, ya que en sí no puede representar una fuente de riesgo: por supuesto, las primeras experimentaciones en este campo llevaron a la proliferación de imágenes facsímil de mala calidad, pero sería algo forzado sugerir que esto sea el problema principal que afecta a los textos en la web. Por otra parte, parece más lógico pensar que Shillingsburg se refiera aquí a los inconvenientes derivados de la transcripción masiva con sistemas defectuosos de reconocimiento de textos, una tendencia que ha llevado a resultados nefastos en varios ámbitos de la edición digital.

Esta desconfianza hacia la transcripción automatizada no sorprende: la crítica ha sugerido en varias ocasiones que la dificultad en localizar ediciones literarias fiables en la red procede en gran medida del uso impropio que se hace de las herramientas de transcripción automática. En efecto, los avances en este campo, que en el siglo pasado contaba solamente con los sistemas de reconocimiento óptico de caracteres (*Optical Character Recognition*, de aquí en adelante OCR) y que ahora ha ido enriqueciéndose con nuevos sistemas de reconocimiento de textos manuscritos basados en redes neuronales (*Handwritten Text Recognition*, de

[2] Entendemos aquí digitalización en su sentido extenso como acto de conversión de un objeto analógico al entorno digital.

aquí en adelante HTR), han permitido una migración masiva del patrimonio textual analógico a la web; pero ¿de qué manera se dio la colonización del espacio digital por medio de estas herramientas?

A pesar de que existen estudios que aseguran cierta fiabilidad de los textos OCRizados presentes en la red (por ejemplo, Kichuk, 2019)[3], es incuestionable que la conversión de los libros antiguos en objetos digitales ha dado pie a una proliferación de documentos electrónicos de dudosa calidad, de los cuales desconocemos con frecuencia características fundamentales como la procedencia, las fuentes y las normas de transcripción adoptadas. En otras palabras, ha ido consolidándose en varios contextos la tendencia a emplear herramientas de transcripción automatizada de forma no supervisada con resultados preocupantes, tanto en la publicación de *ebooks* y ediciones comerciales, que se distribuyen repletos de errores, como en la indexación de los objetos informativos, que cuentan frecuentemente con unos metadatos derivados de un proceso de extracción con OCR descontrolado y asistemático. A todo ello, hay que añadir que la figura del humanista, inicialmente escéptico hacia los resultados generados por los sistemas de transcripción automatizada, ha quedado inevitablemente fuera de la ecuación, hasta ser un mero intérprete y no un actor del proceso de migración digital de los textos, que siguen multiplicándose en la web de modo desordenado y exponencial.

El presente trabajo trata la posibilidad de invertir esta tendencia y ofrecer al humanista un punto de acceso válido para la difusión masiva de textos fiables en la red, jugando en el mismo campo tecnológico que ha favorecido su exclusión del proceso, es decir, el de la transcripción automatizada.

En particular, en estas páginas se presentan los primeros resultados obtenidos por un proyecto colaborativo de transcripción de impresos hispánicos instituido en 2021 por medio de la colaboración entre los siguientes proyectos de investigación: Progetto Mambrino (Università di Verona); BIDISO (Universidade da Coruña); COMEDIC (Universidad de Zaragoza).

En la primera parte del artículo, se ofrece una introducción a los

[3] Sin embargo, hay que reparar en el hecho de que el estudio de Kichuk se ocupa de la distribución masiva de textos en la red para fines no científicos.

sistemas de reconocimiento de caracteres, destacando los principales aspectos relativos a su evolución, su estado actual y los desafíos futuros que suponen con respecto al estudio de los impresos antiguos. La segunda parte está dedicada a la descripción del proyecto de colaboración, del cual se detallan los objetivos y los primeros logros en el ámbito de la transcripción automatizada de impresos de la Edad Moderna. En concreto, se consideran los siguientes aspectos relativos al proyecto: la publicación de dos modelos de transcripción automatizada para los impresos hispánicos en gótica y redonda, estrenados a finales de 2021 y disponibles en acceso abierto dentro de la plataforma Transkribus (READ Coop SCE); las posibilidades de explotación de las transcripciones obtenidas; los principales canales de difusión de los resultados del proyecto; la progresiva alimentación de los modelos de reconocimiento publicados.

La sección conclusiva está dedicada a la presentación de los dos modelos extendidos de HTR *SpanishGothic* (Apéndice 1) y *SpanishRedonda* (Apéndice 2). De ambos se ofrece una ficha de síntesis que proporciona información acerca de las características principales, las instituciones académicas y los estudiosos participantes en su creación, la última versión publicada del modelo y las indicaciones para citarla. Sigue una descripción detallada de las obras que se emplearon para el entrenamiento de la máquina y que constituyen el *dataset* para la creación de los dos modelos: de cada obra se proporcionan las principales informaciones gráficas y bibliográficas con la intención de asistir a los especialistas en la comprensión, utilización y alimentación de los recursos descritos[4].

[4] La *Premisa*, los epígrafes 1 y 2 y las *Conclusiones* son obra de Stefano Bazzaco (Univ. di Verona); el *Apéndice 1* de Ana Milagros Jiménez Ruiz y Ángela Torralba Ruberte (Univ. de Zaragoza); el *Apéndice 2* de Mónica Martín Molares (UDC).

1. Los sistemas de reconocimiento de textos y los impresos antiguos: estado de la cuestión

1.1. Breve historia del reconocimiento de textos

La historia de los sistemas de reconocimiento de textos es amplia y está marcada por precisos saltos tecnológicos. Los estudiosos coinciden en que los primeros pasos en este campo se deben buscar en la invención del *Retina scanner*, allá por 1870, por parte de Carey, es decir, en la creación de un dispositivo que permitía la lectura de imágenes simulando la acción del ojo humano. Este aparato estimuló la experimentación en el campo del reconocimiento automático, lo que produjo la creación de medios electrónicos que pudiesen sustentar la lectura de los ciegos. A partir de él, rápidamente aparecieron el Optófono, un aparato ideado por Fournier d'Albe en la primera década del siglo XX capaz de vocalizar las palabras impresas en una pantalla, y la *Reading Machine* de Tauschek, estrenada por primera vez en 1928 y considerada a todos los efectos el antepasado de los actuales sistemas de transcripción automatizada. Se trata, en efecto, de un aparato que busca la coincidencia entre caracteres impresos y unas representaciones modélicas de los mismos colocadas en un disco rodante: al encontrar una correspondencia, la máquina imprime ese carácter en una nueva hoja y continúa con el reconocimiento del carácter sucesivo.

A partir de este momento, la línea evolutiva de los sistemas de reconocimiento de textos coincide esencialmente con la historia de las herramientas de OCR. En concreto, hablamos de sistemas de OCR en sentido estricto solamente a partir de los años 50, cuando este campo empieza a relacionarse con los intereses de empresas comerciales que por primera vez vislumbraron la posibilidad de ejercer un control generalizado sobre enormes cantidades de datos textuales. La primera generación de *hardware* OCR, es decir, de artefactos físicos que consentían transcribir de forma automatizada documentos impresos, como el IBM 1418[5], se produciría solamente en la década posterior, con la creación de unos

[5] Producido por la Endicott, este dispositivo se lanzó el 12 de septiembre de 1960. Para más detalles, remitimos a la web de IBM, y en especial al siguiente enlace: <https://www.ibm.com/ibm/history/exhibits/endicott/endicott_chronology1960.html> (cons. 21/02/2022).

prototipos con funcionalidades muy limitadas porque consentían la interpretación de un *set* concreto de letras (Narang *et al*., 2020, 5119-5121). En este período, paralelamente, aparecían también unos especiales *typefaces* conocidos como OCR A y OCR B, que eran sistemas gráficos específicamente dibujados para ser interpretados por medio de las tecnologías muy limitadas del período y que constituyeron un primer avance en el área de la transcripción no supervisada.

Llegamos pues a los años 70, un momento determinante para la evolución de los sistemas de reconocimiento de texto porque, junto con la aparición de una segunda generación de *hardwares* OCR que consentían la transcripción de documentos *multifont*, es decir, de mpresos que mezclaban distintos sistemas gráficos, se registran los primeros atisbos de una evolución en el campo del reconocimiento de textos manuscritos. En principio, la máquina pudo interpretar solamente números o letras aisladas y poco complejas como los códigos postales. Necesariamente, debemos fijar aquí el nacimiento de esta subárea del reconocimiento de textos, cuya historia en parte sigue solapándose con la de los OCR hasta adquirir en años recientes un estatuto propio y registrar una creciente y rápida expansión.

Con la reducción del coste del *hardware* y la consiguiente distribución de los *personal computers*, asistimos a la aparición de los primeros paquetes *software* de OCR, que constituyen un verdadero avance porque remiten el problema de la transcripción automatizada a la comunidad de usuarios. Como acaece con frecuencia en el contexto del desarrollo de herramientas digitales, el hecho de que los usuarios tengan acceso a una nueva tecnología constituye un punto de inflexión en la evolución de la misma, puesto que ejercicios e intuiciones particulares son el motor de nuevas experimentaciones. Como directa consecuencia, desde la mitad de los 70, las prestaciones de los sistemas de OCR se incrementan de forma notable: los *softwares* de reconocimiento de textos, sobre todo impresos, llegan a descifrar conjuntos de caracteres muy distintos, mientras que paralelamente se empieza a prestar atención a la interpretación de documentos complejos, por ejemplo, los textos multilengua.

La última etapa evolutiva de los *softwares* de reconocimiento, que podríamos colocar desde los primeros años del 2000 a la época actual, está

caracterizada por los avances más notables. Al mismo tiempo que aparecen proyectos de digitalización de largo alcance y se consolidan los intereses de grandes empresas privadas (Terras, 2010), los sistemas de transcripción automatizada siguen perfeccionándose, sustentados por la ilusión de que en un futuro ya próximo se llegue a transformar todo el patrimonio textual analógico en texto electrónico que la máquina pueda medir y manejar. Con la introducción de nuevos procedimientos de la inteligencia artificial basados en arquitecturas *Long Short Term Memory* (LSTM), como el *deep learning* y las redes neurales, los *softwares* de reconocimiento llegan a interpretar distintas grafías cada vez más complejas (*complex scripts*), incluso textos no occidentales, impresos antiguos y documentos manuscritos.

Sin embargo, si por un lado podríamos decir que el reconocimiento de textos impresos de la actualidad (posteriores a 1930) se considera un problema solucionado, con los impresos antiguos y los manuscritos la situación está lejos de resolverse. Los resultados más alentadores se están dando solo en los últimos años, con el florilegio de plataformas de transcripción automatizada de HTR que constituyen una verdadera revolución para convertir los contenidos textuales de bibliotecas y archivos al espacio virtual de la web. Es interesante notar cómo el nombre de estas herramientas guarda un cambio notable: el interés ha pasado de la interpretación de caracteres aislados (*Optical Character Recognition*) a la búsqueda de *patterns* recurrentes en porciones o líneas de texto (*Handwritten Text Recognition*), prometiendo resultados de reconocimiento que hace 10 años no habríamos podido ni imaginar.

1.2. Los sistemas de OCR/HTR y los estudios humanísticos

Describir la relación entre los humanistas y los sistemas de reconocimiento de texto equivale a trazar una historia de promesas desatendidas.

En la época del desarrollo de los primeros *softwares* de OCR que, como vimos, coincidió con la eclosión de grandes proyectos de digitalización del patrimonio textual, los sistemas de reconocimiento

fueron bien aceptados por parte de la comunidad científica, con filólogos y expertos de documentación a la cabeza; sin embargo, a la luz de unos primeros resultados no propiamente significativos, la ilusión pronto se convirtió en frustración. De hecho, los humanistas empezaron a percibir los sistemas de reconocimiento como instrumentos no fiables para la investigación porque proporcionaban transcripciones repletas de errores, lo cual tuvo como consecuencia una neta distinción entre *clean transcription*, es decir, la transcripción manual realizada con métodos tradicionales, y *dirty OCR*, o sea, los textos generados de forma automatizada.

La crítica subraya cómo este prejuicio está en la base de una renuncia sustancial a la utilización de los sistemas de reconocimiento de textos por parte de los humanistas (Smith-Cordell, 2018, 10-11). En efecto, se trató de una sospecha de fondo difícil de extirpar que persistió hacia los años iniciales del nuevo milenio, a pesar de que en el ámbito informático se dieran progresivos avances tecnológicos en este campo de investigación.

Una primera vuelta a los sistemas de reconocimiento en el ámbito humanístico se dio solamente 20 años más tarde por medio de la creación de grandes repositorios de textos digitalizados en formato imagen, principalmente en el ámbito del proyecto *Google Book Search*, lanzado por la empresa de Mountain View en 2004 con ocasión de la Feria del libro de Frankfurt[6]. Se trata de una forma mediada porque en esta ocasión se emplearon sistemas de OCR más competitivos por parte de los técnicos de Google, sobre todo derivados de un constante refinamiento de los resultados obtenidos con la plataforma de acceso abierto *Tesseract* (desarrollada por Hewlett-Packard entre 1984-94). Sin embargo, la aplicación de esta herramienta se limitaba a la disposición de un estrato OCR oculto, favorable para la búsqueda de palabras clave internas al repositorio. De este modo, el humanista podía tranquilamente desconocer de dónde procedían los resultados de búsqueda, pero implícitamente en sus investigaciones documentales estaba ya sirviéndose de un sistema de reconocimiento de textos, y, quizás movido por los fascinantes resultados que obtenía de la explotación de materiales recolectados en formato digital, parecía también olvidarse de su escasa fiabilidad.

[6] Para una historia del proyecto lanzado por Google remito a los imprescindibles trabajos de Roncaglia (2009; 2010).

Fortalecidos por el interés de perfeccionar las funciones de búsqueda de palabras clave, los sistemas de transcripción automatizada pudieron entonces pasar por una rehabilitación, convirtiéndose en un verdadero punto de referencia para los estudiosos del texto que experimentaban en esos mismos años una inédita atracción por los trabajos de análisis cuantitativo (Moretti, 2005; 2022). Los tiempos eran maduros para la remoción del prejuicio inicial y la vuelta a la experimentación en el campo de la transcripción automatizada con la idea de que pudiera asegurar logros de sumo relieve en distintos campos de las Humanidades. El panorama reciente de los sistemas de OCR/HTR, que sigue enriqueciéndose día a día, es justamente el resultado del cambio de percepción que acabamos de señalar y de un sorprendente florecimiento tecnológico posterior, capaz de sustentar nuevos proyectos de digitalización como el que describimos en estas páginas.

1.3. La transcripción automática de impresos antiguos: estado de la cuestión

Trazar un estado de la cuestión de lo que ha llegado a ser en la actualidad el reconocimiento de textos impresos no es una tarea simple. En general, esto se debe a dos clases de problema.

De entrada, el primer asunto es que la bibliografía relacionada generalmente con este ámbito de estudio es de naturaleza muy variada. De hecho, los estudios sobre OCR/HTR tratan distintas áreas del conocimiento, que van desde la informática pura hasta las ciencias de la documentación y los estudios históricos y literarios. Por consiguiente, la producción de artículos referidos a este campo de indagación es bastante heterogénea: hay artículos técnicos, que relatan el desarrollo de una tecnología determinada en los campos del preprocesamiento de imágenes, la segmentación de imágenes (o *Layout Analysis*), el reconocimiento de documentos complejos (textos no occidentales o multilingües); artículos científicos que tratan casos de aplicación de herramientas de transcripción automática a un corpus de estudio concreto dentro de proyectos editoriales de largo alcance; artículos de carácter más generalista que dan cuenta de proyectos locales de digitalización y de pequeños resultados

obtenidos en una esfera de aplicación muy limitada. Evidentemente, encontrar referencias concretas dentro de un conjunto de estudios tan amplio implica ciertas complicaciones.

En segundo lugar, son pocos los trabajos de investigación que intentan ofrecer una mirada más extensa, que abarque los últimos quince años de actividades en el campo del reconocimiento de textos impresos. Los estudios más sugerentes al respecto vienen de humanistas digitales que, guiados por el objetivo de tratar un caso de estudio concreto, se dedican a reconstruir parte de la tradición bibliográfica relativa a las herramientas digitales adoptadas. Sin embargo, en muchas ocasiones, estos trabajos tienen una finalidad específica, acabando por ofrecer una panorámica muy limitada acerca de otras subáreas de investigación. El ejemplo más relevante al respecto son los artículos producidos por el grupo de investigación alemán que se ha formado bajo el magisterio de Christian Reul y Uwe Springmann en Würzburg[7]. Estos trabajos, a pesar de fundarse en una perspectiva crítica adecuada, acaban por centrarse únicamente en el reconocimiento de textos impresos en *Fraktur*, una grafía empleada por los periódicos alemanes a principios del siglo XX, y no plantean una visión de conjunto.

Para encontrar una investigación que intente tratar de forma exhaustiva el reconocimiento de documentos impresos hay que volver al imprescindible volumen *Electronic Textual Editing* de 2006, editado por Burnard, O'Keeffe y Unsworth con el patrocinio de la MLA (*Modern Language Association*). Efectivamente, la obra, que constituye un hito fundamental dentro de los estudios de Humanidades Digitales, contiene una sección «Practices and procedures», donde aparece un artículo de Gifford Fenton y Duggan (2006, 241-253) que intenta trazar un cuadro general de lo que ha llegado a ser en ese momento histórico el reconocimiento de textos en relación con la filología de los documentos manuscritos e impresos. Este artículo nos servirá de guía para resaltar los avances que se han dado recientemente en este campo.

En la introducción, la autora presenta su experiencia acerca de la

[7] A este grupo se debe el desarrollo de *software* de reconocimiento de textos como OCRopy/OCRopus, Calamari, OCR4All. Para un listado de las publicaciones remito a la bibliografía que se encuentra en Reul *et al.* (2018, 6).

conversión de documentos impresos del repositorio JSTOR, mientras en las siguientes secciones principales aborda los fundamentos del procesamiento con OCR y segmentación automática (*Layout Analysis* o *Zoning*), las características de las fuentes impresas y errores más frecuentes, y los factores decisivos en la adopción de sistemas de OCR en un proyecto de investigación.

Al tratar los fundamentos del procesamiento de textos con sistemas de OCR, Gifford Fenton aclara que la transcripción automática de documentos impresos estaba prometiendo significativos avances, hasta convertirse en lugar común para proyectos de digitalización de largo alcance. La estudiosa considera entonces el flujo de trabajo tradicional de cualquier sistema de OCR, centrándose en tres asuntos principales: la digitalización en formato imagen, los problemas derivados de la segmentación no supervisada y la detección de un orden de lectura de las zonas segmentadas.

Con respecto a las imágenes digitalizadas, se puede apreciar cómo ya en esa época los estudiosos estaban convencidos de que los resultados del reconocimiento con OCR dependían en gran medida de la calidad de los materiales escaneados. En la época actual, gracias a los avances que se han dado en la gestión de imágenes, que en su mayoría se difunden en formatos de alta calidad que alcanzan por lo menos los 300 dpi, también las prestaciones de los sistemas de transcripción automática se han intensificado. Por otra parte, si se considera la segmentación e interpretación de la página digitalizada, notamos que no todos los problemas se han resuelto. Gifford Fenton señala la cuestión de la siguiente manera: «while [...] zoning task would be simple for a human reader with the ability to interpret semantic clues, it can present a variety of challenges for a machine» (2006, 247-248). Es por ello por lo que, a pesar de contar en época reciente con la aplicación de redes neuronales convolucionales (típicas del entrenamiento profundo), el campo sigue presentando inconvenientes. Esto es evidente para cualquier estudioso que se acerque a las herramientas de OCR/HTR disponibles en la actualidad, puesto que las manchas y los desgastes presentes en la fuente digitalizada siguen siendo descifrados por parte del ordenador como porciones de texto y las zonas segmentadas no siempre son interpretadas

siguiendo el orden de lectura correcto.

Al respecto, los avances en el campo de la inteligencia artificial más interesantes residen en la posibilidad de entrenar la máquina sobre modelos de *layouts* preconcebidos, lo cual proporciona resultados cada vez más fiables; pero la impresión general es que faltan todavía unos años para que la cuestión esté resuelta de forma definitiva.

Al analizar las causas de errores más frecuentes en el contexto de la transcripción automatizada, Gifford Fenton destaca varios elementos de las fuentes empleadas que inciden de forma negativa en el reconocimiento: debilitamiento de la tinta, tamaño de las letras, elementos gráficos (a veces ubicados en transparencia bajo el texto), columnas de texto muy pegadas. Sin embargo, contrariamente a lo que se señaló en el caso de la segmentación de la página digitalizada, se trata de problemas resueltos en la época actual. Por lo que atañe a la presencia de elementos gráficos, en el caso de los impresos antiguos normalmente es preferible contar con imágenes de las fuentes en colores. La razón reside en que la detección del *layout* de la página está basada en la densidad de los píxeles y, por tanto, es más probable que los elementos gráficos que no son de interés y que constituyen parte del ruido que obstaculiza el reconocimiento sean excluidos del proceso. El tamaño de las letras y la cercanía de las columnas, por otra parte, no ocasionan problemas relevantes.

Finalmente, al tratar la adopción de sistemas de reconocimiento de textos dentro de proyectos de Humanidades, Gifford Fenton se centra en ocho factores decisivos: seleccionar un enfoque para promover los objetivos del proyecto; definir las características de los materiales fuente; adoptar medidas de control de calidad de los textos transcritos; definir la extensión del proyecto (*scalability*); externalizar procedimientos a un sujeto tercero; considerar el gasto de tiempo que conlleva la producción de texto electrónico; considerar la duración total del proyecto; y poner atención en los costes.

Evidentemente, muchos de estos asuntos son de interés, pero no aplicables al contexto contemporáneo. Con el reciente desarrollo de herramientas de OCR/HTR de acceso abierto, los costes de la transcripción automática no parecen ser una cuestión determinante a la hora de planear un proyecto digital; lo que sí es fundamental es el control

final de los textos transcritos. El argumento de interés entre los señalados por Gifford Fenton es el tercero, que trata el control de la calidad, es decir, las operaciones posteriores al reconocimiento. Al respecto, se están llevando a cabo importantes experimentaciones en distintas áreas de las Humanidades Digitales. En el campo de la automatización se están integrando diccionarios y sistemas de procesamiento del lenguaje natural que agilizan el trabajo del filólogo que revisa los textos. Por otro lado, en el ámbito de la colaboración, se están promoviendo recientemente planes de transcripción en grupo, como el *Transcribathon* de Europeana <https://www.transcribathon.com/en/> (cons. 15/05/2022) y la revisión múltiple en *crowdsourcing*. Se trata de unas prácticas de largo alcance que miran por la formación de estudiantes y colaboradores para la producción de datos fiables y certificados, generalmente realizados bajo la supervisión de especialistas de la materia que validan el trabajo y gestionan todo el flujo de producción. El resultado es que de estos proyectos pronto se obtendrá una cantidad ingente de transcripciones certificadas, que podrán constituir el entrenamiento de *softwares* de reconocimiento de textos progresivamente más precisos y fiables.

1.4. De la función instrumental a la revolución heurística: ¿cuál puede ser el futuro de los sistemas de reconocimiento de textos?

Si miramos retrospectivamente a la evolución de los sistemas de reconocimiento de textos y partimos de una mirada menos ilusionada de la que tenían los humanistas de los 80, más que de sueños frustrados parece que deberíamos hablar de una actitud sustentada por premisas erróneas y caracterizada por cierta impaciencia. Es habitual que, para adaptarse a las metodologías tradicionales de las disciplinas humanísticas, una nueva tecnología digital necesite en principio un tiempo de acomodamiento y de reflexión crítica que regularice su utilización indiscriminada (Orlandi, 1994, 7 y ss.)[8].

[8] Al respecto, propongo retomar las palabras de Tito Orlandi, padre fundador de la escuela romana de Humanidades Digitales, quien afirma: «È accaduto a varie riprese che siano state prodotte delle macchine, nuove e potenti, ma con scopi limitati e praticamente semplici, e che soltanto dopo esse

Por ejemplo, esto pasó con la codificación en lenguaje XML de las ediciones académicas digitales, donde el modelo proporcionado por la TEI (*Text Encoding Initiative*) se asentó como estándar solamente después de un tiempo muy largo de reflexión crítica que abarcó las últimas dos décadas del siglo XX. A pesar de que ahora está padeciendo legítimas críticas por parte de algunos detractores, no hay duda de que se trata de un estándar altamente productivo, que sigue representando un punto de referencia esencial en el campo de la publicación digital por haber sido contextualizado y variadamente explotado por tantos estudiosos de Humanidades desde su aparición.

Lo mismo está ocurriendo con los estudios cuantitativos de la literatura, que ahora, quince años después de la publicación de los primeros trabajos de lectura distante realizados por Moretti (2005), están demostrando su solidez y fertilidad, abriendo la vía a nuevos caminos interpretativos, sobre todo para la estilometría (Hernández Lorenzo, 2019; Calvo-Tello, 2021) y, dentro de ella, la atribución de autoría (García-Reidy, 2019; Bazzaco, 2022).

No se dieron las mismas condiciones en el campo del reconocimiento de textos. Quizás se deba a un entusiasmo inicial por parte de los humanistas, que veían en la nueva tecnología la posibilidad de simplificar trabajos engorrosos como la transcripción manual y la metadatación, y a una sucesiva frustración derivada del empleo de una tecnología, en esos años poco madura, que cometía muchos errores. La consecuencia natural de esta situación fue el prematuro abandono de las experimentaciones de transcripción automatizada de manuscritos e impresos antiguos, calificándolas como poco rentables para el filólogo. Sin embargo, si consideramos que recientemente el campo del reconocimiento de textos ha experimentado unos notables avances, ¿podemos rectificar el prejuicio que caracterizaba esta área de estudio de las Humanidades Digitales? o, en otras palabras, ¿es legítimo suponer que la adopción de sistemas de reconocimiento de textos podría constituir un recurso de interés para la reorganización del trabajo filológico? y, junto a

abbiano generato una riflessione teorica, che dunque ha seguito e non preceduto l'innovazione tecnologica. […] Tuttavia sono state proprio le riflessioni teoriche che hanno mostrato il vero significato dei risultati che si potevano ottenere con le macchine» (1994, 8).

ello, ¿podemos imaginar posibles aplicaciones de herramientas de transcripción automática para acrecentar la presencia de documentos fiables en la red, invirtiendo la tendencia actual?

Si para contestar a estas preguntas nos limitamos a una percepción que se consolidó hace más de 30 años, corremos el riesgo de valorar de forma errónea el problema. Los sistemas de reconocimiento de textos, en efecto, han llegado en tiempos recientes a prometer unos resultados de transcripción automatizada muy alentadores, que se acercan al 1% de error para los impresos antiguos y al 5% para los manuscritos: se hace necesaria una rehabilitación de este campo de trabajo que, en la estela de los avances tecnológicos, enseñe los nuevos caminos que pueden abrirse para la labor filológica.

Originariamente el reconocimiento de textos se fundamenta en la idea de aligerar y acelerar de forma notable el proceso de transcripción manual. Ya señalamos en otra ocasión (Bazzaco, 2020, 539 y ss.) cómo esta función, que llamaríamos de tipo *instrumental*, estaba ya en la base de los planes de digitalización del patrimonio cultural que surgieron durante los años 80 del siglo XX. Melissa Terras (2010) recuerda cómo los primeros proyectos de escaneo de fuentes documentales tenían como premisa la transformación de los archivos de imágenes en un texto electrónico; lo cual nos lleva a pensar que el mismo proceso de digitalización estaba en su nacimiento íntimamente relacionado con el desarrollo de sistemas de OCR fiables que permitiesen volcar los datos textuales extraídos en un formato que la máquina pudiera medir y manejar (*machine readable form*).

Es cierto que esta es la función primaria que le asignamos a los sistemas de transcripción automatizada: prometer una rápida transformación de las imágenes digitales en textos electrónicos explotables, que pueden ser reutilizados después en varios campos de la investigación en Humanidades, como la edición digital, la extracción de corpora y lemas, los procedimientos de análisis cuantitativo, etc.

Proyectos de Humanidades Digitales que emplean sistemas de transcripción automática de forma masiva son, por ejemplo,

TranscribeBentham (2013-2017)[9], TrAIN (*Tracing Authorship In Noise*, 2018)[10], *Entangled Histories: Ordinances of the Low Countries*[11], CREMMA (*Consortium pour la Reconnaissance d'Écriture Manuscrite des Matériaux Anciens*, 2020-2021)[12], o, en relación con los estudios hispánicos, el prestigioso proyecto ETSO (Estilometría aplicada al Teatro del Siglo de Oro, 2018) dirigido por Germán Vega y Álvaro Cuéllar[13]. En todos estos casos, sin embargo, el reconocimiento de textos es visto como un medio para apresurar y aligerar el trabajo manual, no como un momento fundamental dentro del flujo ecdótico y editorial.

Aun considerando que la función instrumental sigue siendo la razón esencial para la adopción de sistemas de OCR/HTR en el interior de un proyecto de edición de textos, es necesario avanzar en la reflexión teórica y apuntar los posibles caminos para el futuro empleo de estas herramientas que aclaren hasta dónde se puede llegar.

Como punto de partida, conservan su validez las reflexiones de Raul Mordenti sobre el procedimiento ecdótico en ambiente informático. El estudioso, en efecto, habla en su obra de transcripción, que propone considerar como un momento más del acto de codificación del documento, o sea el primer eslabón de un proceso editorial que pretende fijar gráficamente el texto; si por un lado la maquetación del texto electrónico es «momento crucialissimo» porque corresponde a la «immissione nella macchina dell'informazione da cui dipenderanno tutti i successivi trattamenti e manipolazioni», por otra parte se califica la transcripción como «momento forte», «decisivo» porque corresponde al procedimiento «più costoso in termini di tempo/uomo», capaz de configurar todos los pasos posteriores (2001, 29). Cuando transcribimos en formato máquina un texto estamos en el nivel de una primera codificación, es decir, la transformación del texto contenido en los

[9] Transcribe Bentham <https://blogs.ucl.ac.uk/transcribe-bentham/> (cons. 15/05/2022). Véase Causer-Terras (2014).
[10] TrAIN <http://www.etrap.eu/research/tracing-authorship-in-noise-train/> (cons. 15/05/2022). Véase Franzini *et al.* (2018).
[11] Entangles Histories <https://lab.kb.nl/dataset/entangled-histories-ordinances-low-countries> (cons. 15/05/2022).
[12] CREMMA <https://www.dim-map.fr/projets-soutenus/cremma/> (cons. 15/05/2022).
[13] ETSO <https://etso.es/> (cons. 15/05/2022).

materiales escaneados en una secuencia de bits. A esta primera codificación, sigue una segunda codificación de las informaciones más relevantes que el documento transmite, es decir, la modelización de los elementos que el editor quiere conservar, sean estos aspectos semánticos, materiales, lingüísticos, etc. del texto fuente.

Con respecto al proceso ecdótico, los sistemas de reconocimiento de textos permiten automatizar –o, en otros términos, delegar a la máquina– el primer tipo de codificación, es decir, la «digitización» (del inglés *to digit*, o sea «teclear») del texto fuente, su versión en un «magnetoescripto» (Mordenti, 2001, 49). Al respecto, apreciamos cómo en la ecdótica tradicional los dos momentos se dan de forma sincrónica, puesto que al acto de transcripción manual se integra la necesidad de explicitar cuáles son las características principales del texto que se edita[14]. Sin embargo, dentro del espacio digital, que impone ordenar y formalizar de modo secuencial las operaciones (tanto que *divide et impera* ha llegado a ser el mantra de los filólogos digitales), la conversión no supervisada de imágenes en texto electrónico y su modelización corresponden a dos procedimientos distintos y sucesivos: primero se transcribe un documento, que en un segundo momento se maqueta de acuerdo con estándares compartidos. La interpretación se limita, pues, a la segunda codificación, mientras que la primera correspondería a un acto aparentemente neutral, que quizás podemos asemejar en el contexto analógico a la realización de una edición diplomática, donde a partir de unos criterios fijos de transcripción, que evidentemente se establecen con prioridad respecto al trabajo, se proporciona un texto que es una reproducción fiel de lo que aparece en la fuente.

Por lo que atañe a la segunda codificación, que llamamos interpretativa (o, lo que es lo mismo, crítica), los sistemas de reconocimiento de textos todavía no ofrecen una solución viable de

[14] Con respecto a este asunto, ténganse en cuenta las palabras de Mordenti: «[è] possible vedere come il concetto di edizione critica contenesse in sé e unificasse molte cose assai diverse fra loro. Quando noi trascriviamo un testo noi compiamo un'operazione di ricodifica che in realtà sovrappone e mescola nel gesto del trascrivere (che ci appare, del tutto erroneamente, semplice) diverse funzioni» (2001, 47). Entre las funciones mencionadas se encuentran: la conservación, la reproducción, la corrección, el acercamiento al lector; en el caso del texto digital se suma a estas funciones la necesidad de re-codificación en formato informático para explotar después el poder de cálculo de la máquina.

automatización del trabajo, porque la maquetación tiene que pasar por una *selección* de las características textuales explícitas. Por otra parte, en el caso de la codificación del texto en formato máquina, los recientes logros en el campo de la transcripción automatizada pueden representar un avance considerable: en primer lugar, porque permiten obtener un texto fiable, que respeta con regularidad todos los signos gráficos que aparecen en la fuente; en segundo lugar porque, simplificando la transcripción manual, dejan al editor centrarse principalmente en el acto de codificación interpretativa; y, en tercer lugar, porque potencian exponencialmente la posibilidad de contrarrestar la difusión de contenidos poco fiables en la red, agilizando de modo considerable la creación de ediciones digitales que, por ser muy apegadas al texto fuente, son también más respetuosas.

Basándose en estos presupuestos, la transcripción automatizada podría afectar al mismo procedimiento ecdótico, porque permite transcribir varios testimonios de un texto en un tiempo en que normalmente la transcripción manual no llegaría ni a ofrecer la transcripción completa de un ejemplar. De tal manera, podemos imaginar que en un futuro muy cercano existirán métodos que faciliten la corrección de los testimonios transcritos y la colación inmediata de todos ellos, de cara a una publicación digital que consienta navegar a través de las variantes textuales[15]. Se trata, en otras palabras, de concebir el proceso ecdótico de otra forma, rebajando la necesidad de un texto único reconstruido y centrando la atención en la tradición textual de una obra, que es el fruto de sus distintas concreciones a lo largo del tiempo[16].

A manera de ejemplo, y para sintetizar cuanto acabamos de decir, piénsese en un proyecto que supone la edición de un texto bastante largo que cuenta con más de cinco ediciones. En el contexto analógico, necesariamente sometido al paradigma de la página, el editor seleccionaría un ejemplar fiable, lo transcribiría, cotejaría las variantes presentes en otras ediciones y ejemplares, proporcionaría un texto reconstruido que tuviera

[15] Pienso, por ejemplo, en la visualización sinóptica de variantes que ofrece la herramienta EVT2, segunda versión de la herramienta Edition Visualization Technology proporcionada por el grupo de la universidad de Turín dirigido por Roberto Rosselli del Turco. Véase Rosselli Del Turco *et al.* (2019).

[16] Al respecto, es constante la publicación de trabajos que insisten en una inédita *primacía del documento* en el contexto de la filología digital como, por ejemplo, Mordenti (2001, 31 y ss.), Pierazzo (2015) y Allés Torrent (2017, 69 y ss.).

en cuenta (utópicamente) todas las variantes y, finalmente, relegaría al aparato las lecciones alternativas encontradas. Al revés, el medio digital promete unos cambios metodológicos significativos con respecto al mismo proceso, porque no impone la selección de un testimonio ni la fijación de un texto artificialmente reconstruido. El editor, en tales condiciones, puede considerar conjuntamente todas las ediciones (incluso todos los ejemplares supervivientes), transcribir los testimonios de forma no supervisada con herramientas de OCR/HTR, corregir el texto (por medio también de *specific domain dictionaries*), detectar semi-automáticamente las variantes en las transcripciones obtenidas, publicar una edición que de modo simultáneo y dinámico permita navegar entre las distintas cristalizaciones del documento.

De tal manera la remediación digital condiciona la misma heurística del trabajo editorial, porque no se fundamenta en la intención de solucionar viejos problemas (función instrumental), sino en la posibilidad de abrir nuevas vías para la reproducción de los textos analógicos en un entorno digital. No hay duda de que los sistemas de reconocimiento de textos, por las razones indicadas, podrían convertirse en herramientas de interés para el trabajo ecdótico y tener un papel determinante en la migración web de documentos impresos.

Esto es el presupuesto principal que ha llevado a la constitución del proyecto colaborativo que tratamos a continuación.

2. Modelos de HTR para el reconocimiento de impresos hispánicos de la Edad Moderna

2.1. Primeras experimentaciones en el campo de la transcripción automatizada de impresos de la Edad Moderna

El Progetto Mambrino nació en 2003 para llevar a cabo una exploración de las continuaciones italianas de los ciclos caballerescos castellanos de Amadís y Palmerín. El grupo de investigación veronés se acercó al campo de la transcripción automática de impresos antiguos con la plataforma Transkribus a partir del año 2017; sin embargo, la idea de

transcribir y editar las obras de interés del proyecto entraba en los planes de sus directores desde hacía tiempo[17].

Hacia 2010 hubo un primer empuje debido a la posibilidad de financiar a dos becarias de investigación durante un año, con un proyecto de digitalización de los ejemplares del corpus conservados en bibliotecas locales, sobre todo la del ayuntamiento de Verona. En tal ocasión se pudieron crear unos recursos digitales que contenían las imágenes de las obras en alta calidad (formato RAW) y que salieron en DVD en una publicación unitaria de 20 discos. Al mismo tiempo se publicaron los recursos en un formato apto para la difusión en línea en el sitio web del proyecto[18], que entonces surgió precisamente como recolector de datos para censar y distinguir los ejemplares registrados, y para alojar las colecciones digitales que teníamos preparadas. En aquella ocasión, cada una de las becarias transcribió una obra del ciclo amadisiano a partir de las imágenes escaneadas de las fuentes, realizando una edición de ese ejemplar, especialmente valiosas siendo las primeras trascripciones manuales de estos libros de caballerías italianos[19].

Este primer intento fue muy importante porque puso de relieve los principales límites de sostenibilidad del proyecto: se trataba, en concreto, de transcribir un grupo de obras muy extensas –cada una de más de 900 hojas–, lo cual suponía unos costes muy elevados en términos económicos y, sobre todo, de tiempo.

Frente a esta dificultad, se consideró la oportunidad de automatizar parte de la tarea de transcripción gracias al empleo de herramientas de reconocimiento de textos. No obstante, para los impresos antiguos no existían todavía sistemas de OCR fiables. En principio, se llevaron a cabo algunos experimentos con el *software* ABBY FineReader, pero la aplicación ofrecía un reconocimiento por caracteres aislado, por lo tanto, aun entrenando el *software* con la transcripción manual de parte del texto, los resultados de transcripción no fueron prometedores. Debido a la

[17] Los primeros pasos del proyecto se describen en Bazzaco (2018).

[18] Web del Progetto Mambrino, Sección *Collezioni digitali* <https://www.mambrino.it/it/collezioni-digitali/biblioteca-civica-di-verona> (cons. 23/03/2022).

[19] Paola Bellomi editó la continuación al quinto libro de Amadís titulada *Il secondo libro delle prodezze di Splandiano* (Venezia, Michele Tramezzino, 1564); Federica Colombini editó la continuación al décimo libro, la *Aggiunta al Florisello (Le prodezze di don Florarlano)* (Venezia, Michele Tramezzino, 1564).

materialidad de las fuentes, es decir, impresos del Renacimiento en formato octavo, altamente manejados y, en ocasiones, muy desgastados por el paso del tiempo, junto con la presencia de efectos impropios de iluminación y contraste derivados del escaneo manual, se obtuvieron unos resultados inservibles, con transcripciones que tenían un margen de error elevado (Mancinelli, 2016; Bazzaco, 2018).

A partir del año 2016, establecimos los primeros contactos con el Proyecto Europeo READ (Recognition and Enrichment of Archival Documents). READ nacía de otro proyecto financiado por la Unión Europea llamado TranScriptorium[20], que tenía el objetivo de poner a disposición de los usuarios una refinada tecnología de HTR que permitiera la digitalización en formato texto electrónico de documentos de archivo, sobre todo manuscritos. Persiguiendo los mismos objetivos y tras la experiencia de este primer proyecto, READ (2016-2019), al amparo de otra financiación europea, produjo y difundió la plataforma de HTR Transkribus. Tal aplicación contaba desde su nacimiento con una consistente comunidad de referencia[21] y prometía resultados alentadores no solo con los textos manuscritos, sino también con los impresos antiguos, porque ambos compartían unos problemas parecidos en cuanto a variedad gráfica de las letras (*fluctuation*).

Guiados por la idea de considerar los impresos antiguos como si fueran documentos manuscritos muy regulares, llevamos a cabo unas pruebas con la plataforma Transkribus, que presentamos durante las jornadas de estudios «Transcribing. Towards an OCR for old fonts» (5-6 junio de 2018)[22]. La iniciativa, que se concluyó con un taller práctico para aprender a usar Transkribus, vio la participación del director del proyecto READ, el Dr. Günter Mühlberger, y de unos colaboradores del grupo de

[20] El proyecto TranScriptorium (2013-2015) fue llevado a cabo gracias a la colaboración entre la Universidad de Innsbruck, la University of London, la Universidad Politécnica de Valencia y otras instituciones. Para más detalles véase el siguiente enlace: <https://cordis.europa.eu/project/id/600707/it> (cons. 24/03/2022).

[21] Participaron en la fundación del proyecto más de diez instituciones, entre las cuales figuran, junto con las universidades ya citadas, la Universidad de Rostock, la Technische Universität de Wien, el University College de Londres. Para más detalles véase: <https://cordis.europa.eu/project/id/674943/it> (cons. 24/03/2022).

[22] Enlace a la iniciativa: <https://www.dlls.univr.it/?ent=iniziativa&id=7892&lang=it> (cons. 24/03/2022).

investigación veronés.

Para entonces, dentro del Progetto Mambrino ya se habían llevado a cabo los primeros entrenamientos sobre la cursiva veneciana del siglo XVI, lo cual había permitido obtener las transcripciones de un primer conjunto de textos del ciclo italiano de *Amadís de Gaula* con un margen de error muy bajo[23].

La labor de transcripción automática que se había llevado a cabo constó de las siguientes fases[24]:

1. Censo de los ejemplares digitalizados y descarga de las imágenes facsímiles.
2. Subida de las imágenes a la plataforma Transkribus.
3. Segmentación (*Layout Analysis*) de las imágenes en distintas áreas (correspondientes a caja y líneas de texto).
4. Transcripción manual (*Ground Truth production*) de una porción del texto segmentado (alrededor de 2000 palabras por cada obra).
5. Entrenamiento de la máquina y creación de un modelo de HTR individual (específico para cada obra).
6. Transcripción automática de las obras del corpus, en principio las que constituyen el Ciclo italiano de *Amadís de Gaula* (alrededor de 20 obras, entre traducciones y continuaciones originales);
7. Extracción de las transcripciones en formato DOC, TXT y XML para la modelización de cada edición según el estándar TEI.

Los resultados obtenidos en la cursiva gracias a Transkribus fueron muy alentadores, porque con alrededor de 2000 palabras transcritas manualmente para cada libro se podía generar una transcripción muy fiable con un índice de error (en Transkribus: *Character Error Rate*) que no superaba el 2%. Véase al respecto la siguiente tabla que indica los resultados obtenidos con los seis volúmenes del *Sferamundi di Grecia*, libro trece del *Amadís*:

[23] Para un inventario de los resultados obtenidos, consúltese Bazzaco (2018, 265-268).

[24] Una descripción completa del flujo de trabajo que supone Transkribus se encuentra en Mühlberger *et al.* (2019, 957 y ss.).

Libro	Localización ejemplar	Letra	Resultados (*CER*)
13/1 *Sferamundi. Prima parte*. 1558	Madrid, Biblioteca Nacional de España, 5-4978	cursiva	1.57%
13/2 *Sferamundi. Seconda parte*. 1560	Madrid, Biblioteca Nacional de España, 5-4978	cursiva	1.21%
13/3 *Sferamundi. Terza parte*. 1563	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 20)	cursiva	1.80%
13/4 *Sferamundi. Quarta parte*. 1563	München, Bayerische Staatsbibliothek, P.o.hisp. 105 k-4	cursiva	1.11%
13/5 *Sferamundi. Quinta parte*. 1565	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 22)	cursiva	1.59%
13/6 *Sferamundi. Sesta parte*. 1565	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 22)	cursiva	1.71%

Tabla 1. Primeros resultados de transcripción automática con el *Sferamundi di Grecia* (Venecia, s. XVI)

En un segundo momento, las transcripciones obtenidas, revisadas por unos especialistas y etiquetadas según el estándar XML TEI compondrían el corpus piloto de la Biblioteca Digital del Progetto Mambrino, un proyecto de edición digital de los libros de caballerías italianos que se está llevando a cabo hoy en día gracias a una financiación concedida en 2018 por el Ministerio Italiano de Universidad dentro de un proyecto más amplio llamado *Mapping Chivalry: Spanish Romances of Chivalry from Renaissance to 21st Century: a digital approach* (en el que participa un equipo que incluye cuatro unidades: de Verona, Trento, Roma y Salerno) y otra financiación que obtuvo en el mismo año el Departamento de Lenguas y Literaturas Extranjeras de la Universidad de Verona, como Departamento de Excelencia.

En el tiempo en que estábamos transcribiendo de forma automática la letra cursiva, se fortalecieron los contactos con los proyectos de investigación españoles BIDISO y COMEDIC, que propusieron poner a prueba la plataforma Transkribus con textos castellanos de su interés.

En el interior del Progetto Mambrino ya se estaba pensando en extender el reconocimiento automático a los textos españoles, gracias a la elaboración de unos modelos de reconocimiento para obras caballerescas impresas en el XVI en letra gótica. Las primeras pruebas con la gótica fueron muy alentadoras, ya que con un *dataset* muy limitado de páginas

transcritas manualmente (alrededor de 1500 palabras por obra) pudimos entrenar un modelo de HTR individual para los libros de caballerías castellanos con el que experimentamos[25]. Finalmente, una vez acabado el entrenamiento de la máquina, transcribimos las obras caballerescas en gótica con un margen de error cercano al 2%. Poco más tarde fueron los primeros intentos de transcripción automática de documentos impresos en letra redonda, que aseguraron unos resultados parecidos (Tabla 2).

Libro	**Localización ejemplar**	**Letra**	**Resultados (*CER*)**
Leandro el Bel. Toledo, Ferrer, 1563	Madrid, Biblioteca Nacional de España, R/9030	gótica	1.43%
Florando de Inglaterra. Lisboa, Gallarde, 1545	London, British Library, C62 H14	gótica	2.13%
Silves de la Selva. Sevilla, De Robertis 1549	Madrid, Biblioteca Nacional de España, R/865	gótica	1.58%
Libro de los Siete Sabios de Roma. Barcelona, Andreu, 1678	Madrid, Biblioteca Nacional de España, R/530	redonda	2.30%

Tabla 2. Primeros resultados de transcripción automática de documentos en gótica y redonda

Si se tiene en cuenta que estas pruebas se realizaron en un momento en que la plataforma Transkribus estaba todavía en su desarrollo tecnológico y contaba con un número mínimo de documentos procesados, se puede comprender cuáles son las posibilidades reales que ofrece la herramienta. En la actualidad, con la introducción de nuevos sistemas de reconocimiento (del HTR básico a los métodos HTR+/PyLaya) y una cantidad interminable de documentos procesados, podemos imaginar que los resultados serían aún más convencedores[26].

En la estela de estos primeros logros y guiados por la idea de extender el reconocimiento a un conjunto más extenso de textos hispánicos de la Edad Moderna, se generó una red de colaboración internacional entre

[25] Los colaboradores del proyecto que participaron fueron: Stefano Bazzaco (*Leandro el Bel*), Stefano Neri (*Florando de Inglaterra*), Giada Blasut (*Silves de la Selva*). Otras pruebas con la redonda fueron realizadas por Bazzaco en el mismo año.

[26] Recordamos al respecto que el reconocimiento en Transkribus, fundándose en procedimientos de *machine learning*, incrementa sus prestaciones según aumenta la cantidad de documentos procesados (Mühlberger *et al.*, 2019, 957).

investigadores de los tres proyectos, con el fin de proporcionar unos recursos que fueran de utilidad para toda la comunidad científica.

2.2. Modelos de HTR para la transcripción automática de impresos hispánicos en gótica y redonda

El proyecto de reconocimiento de impresos hispánicos de la Edad Moderna nace formalmente en 2021 de la colaboración entre tres proyectos de investigación distintos: el Progetto Mambrino (Universtà di Verona), BIDISO (Universidade da Coruña) y COMEDIC (Universidad de Zaragoza).

En un primer seminario formativo impartido por Stefano Bazzaco, se constituyó una red de colaboración que vio la participación de una docena de investigadores en un trabajo colectivo para la transcripción manual y el entrenamiento de la plataforma Transkribus en relación con distintos textos impresos en gótica y en redonda. Los investigadores involucrados en esta colaboración fueron: Giada Blasut, Federica Zoppi, Manuel Garrobo Peral (Progetto Mambrino); Ana Milagros Jiménez Ruiz, Ángela Torralba Ruberte, Nuria Aranda García, Daniela Santonocito, Gaetano Lalomia (COMEDIC); Carlota Fernández Travieso, Mónica Martín Molares (BIDISO). Los resultados del proyecto se dieron a conocer por primera vez en el Congreso Internacional «Humanidades Digitales y estudios literarios hispánicos. De los impresos de la Edad Moderna a las ediciones académicas digitales», que se celebró en la Universidad de Verona en junio de 2021.

Objetivos del proyecto colaborativo: la finalidad del proyecto, como ya se anticipaba, era poner a disposición de otros estudiosos unos modelos de HTR que fueran aplicables directamente, sin la necesidad de entrenar la máquina cada vez que se necesitaba un nuevo texto transcrito. Por esta razón, buscamos la vía para generar unos modelos de reconocimiento extendidos que pudieran abarcar un conjunto muy amplio de obras. Con respecto a los modelos de HTR individuales, que están basados en una sola obra, los modelos de HTR extendidos se fundamentan pues en obras

distintas y permiten transcribir todos los textos que presenten unas características tipográficas parecidas con un buen grado de fiabilidad.

Flujo de trabajo: para empezar, se escogieron los documentos que constituirían el *dataset* del modelo, es decir, obras de naturaleza variada que abarcasen distintas representaciones gráficas de caracteres del mismo tipo. La creación de unos modelos de HTR extendidos constó entonces de tres etapas principales: transcribir manualmente, según unos mismos criterios, un número determinado de páginas (en nuestro caso, alrededor de 20 páginas) pertenecientes a distintas obras; entrenar la máquina sobre el conjunto de transcripciones realizadas; generar un modelo de HTR único que transcribiera automáticamente otros textos que no pertenecieran al conjunto inicial. Para cumplir con este flujo de trabajo, Transkribus ha sido un recurso fundamental, porque se presenta como una plataforma con varias funciones de automatización y realmente colaborativa, ya que consiente la interacción de los usuarios de forma simplificada y asíncrona[27].

En principio, se fijaron unos criterios de transcripción rigurosos para el conjunto de transcripciones manuales que se realizaron para la producción de los dos modelos extendidos, uno para la gótica (siglos XV-XVI) y uno para la redonda (siglos XVI-XVII). Tales criterios se regían por tres necesidades principales: preservar todos los signos gráficos que presentaba la fuente; aligerar el proceso de postproducción de los textos exportados; fomentar la solidez (*consistency*) del modelo, o sea, mantener una regularidad total en las transcripciones manuales para que la máquina pudiera aprender unos patrones recurrentes de identificación de las letras[28]. Finalmente, elegimos unos criterios de transcripción semi-diplomática que respetaran la variabilidad textual de las fuentes, pero que

[27] Entre otras señalamos las siguientes funciones: Text2Image, para importar en la plataforma textos ya transcritos; P2PaLA, para la segmentación de la página digitalizada con modelos de *layout* predefinidos; Keyword Spotting (KWS), para la extracción de palabras clave a partir de la forma gráfica de las mismas. Para más detalles véase la página de recursos <https://readcoop.eu/it/transkribus/resources/> (cons. 24/03/2022).

[28] Las convenciones de transcripción establecidas por Transkribus están disponibles en la web de Read Coop, en el siguiente enlace: <https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/> (cons. 24/03/2022).

a la vez consintieran reducir intervenciones posteriores en fase de revisión, sobre todo en el caso de las abreviaturas (Tabla 3):

Criterios de transcripción [29]
1. Signos de interpunción: se mantienen como aparecen en el texto fuente. 2. Acentos: se mantienen como aparecen en el texto fuente. 3. Signo tironiano: se trascribe como 'e' comercial (&). 4. 's' larga (ſ): se transcribe como 's' simple. 5. Abreviaturas: se desarrollan.

Tabla 3. Criterios de transcripción adoptados

Resultados: una vez acabadas las transcripciones manuales (*Ground Truth*) que habrían constituido la base del entrenamiento dentro de la plataforma, generamos los dos modelos de HTR que se describen brevemente a continuación (Tabla 4).

SpanishGothic_XV-XVI_extended [30]	**SpanishRedonda_XVI-XVII_extended** [31]
Versión actual: 1.0.0 Dataset: 16 textos, 150˙137 palabras Fiabilidad: 99.08%	Versión actual: 1.0.0 Dataset: 14 textos, 61˙938 palabras Fiabilidad: 98.93%

Tabla 4. Descripción sintética de los dos modelos de HTR publicados

Aun teniendo en cuenta que la fiabilidad de los dos modelos se basa en los textos que constituyen el *Dataset* y que, por consiguiente, en los textos externos al modelo el porcentaje de letras transcritas correctamente puede disminuir, hay que subrayar que las primeras pruebas efectuadas

[29] Los criterios están disponibles en GitHub, en los siguientes enlaces: <https://github.com/stefanobazzaco/HTR-model-SpanishGothic_XV-XVI_extended#transcription-criteria> / <https://github.com/stefanobazzaco/HTR-model-SpanishRedonda_XVI-XVII_extended#transcription-criteria> (cons. 24/03/2022).

[30] Para una descripción detallada del modelo remito al *Apéndice 1* del presente trabajo.

[31] Para una descripción detallada del modelo remito al *Apéndice 2* del presente trabajo.

han asegurado cierta consistencia con respecto a los resultados obtenidos, con índices de error muy bajos[32].

Distribución: los modelos de HTR creados están disponibles en abierto desde julio de 2021 dentro de la plataforma Transkribus en la sección *Public Models*; pueden, por lo tanto, ser empleados por cualquier usuario que tenga acceso a la plataforma.

Para la utilización de los modelos, el estudioso tiene que atender al siguiente flujo de trabajo:

a) de entrada, subir las imágenes digitalizadas de la obra de interés a la plataforma;
b) ejecutar la segmentación de las imágenes (*Layout Analysis*) de forma automatizada o manual[33];
c) lanzar el reconocimiento con uno de los modelos a disposición (duración: unos minutos por página).

Por medio de estos pasajes, el usuario podrá pasar a la exportación de las transcripciones obtenidas en varios formatos (DOC, PDF, TXT, XML)[34].

Contribución y puesta al día: los *datasets* que constituyen la base del entrenamiento de los modelos están disponibles en el repositorio Zenodo con acceso limitado[35], según el protocolo establecido por la licencia Creative Commons CC BY-NC-ND 4.0, que impide la creación de objetos derivados y su distribución con finalidades comerciales[36]. En el futuro, se piensa implementar progresivamente cada modelo de HTR por medio de

[32] Al respecto, véanse las primeras pruebas efectuadas por Giada Blasut en esta misma publicación (pp. 175-193).
[33] Se dispone también de un modelo de P2PaLA que permite la segmentación automatizada de textos en doble columna. Para más información contactar con los autores.
[34] Para una descripción exhaustiva del flujo de trabajo de la plataforma Transkribus, véase Bazzaco (2018) o bien la web de READ Coop: <https://readcoop.eu/transkribus/resources/how-to-guides/> (cons. 24/03/2022).
[35] Dataset del modelo de HTR SpanishGothic_extended_sXV-XVI: <https://zenodo.org/record/4888927#.Yj2hSerMI2w>. Dataset del modelo de HTR SpanishRedonda_extended_sXVI-XVII: <https://zenodo.org/record/4889218#.Yj2hS-rMI2w> (cons. 24/03/2022).
[36] Para más información, véase el siguiente enlace: <https://github.com/stefanobazzaco/HTR-model-SpanishRedonda_XVI-XVII_extended#copyright-statement> (cons. 24/03/2022).

la inserción de otras porciones de textos transcritos manualmente por distintos estudiosos que quieran colaborar en el proyecto colaborativo. En este sentido, Zenodo se revela como un recurso de fundamental importancia porque permite que los *datasets* implementados se pongan al día con regularidad cada semestre, incrementando las prestaciones de los modelos publicados y manteniendo un mismo identificador unívoco (DOI) para ambos recursos.

Conclusiones

El proyecto colaborativo descrito se fundamenta, evidentemente, en una constante actualización que pueda proporcionar resultados cada vez más satisfactorios para el filólogo. El punto de partida, como se ha señalado, era el de contrarrestar los efectos de la migración descontrolada de textos no especializados en la red; sin embargo, los resultados obtenidos y el flujo de trabajo establecido permiten añadir unas consideraciones ulteriores.

En concreto, observamos cómo el área del reconocimiento de texto está prometiendo beneficios de gran interés para la investigación: no hay duda de que la tecnología HTR ha alcanzado recientemente un nivel de desarrollo muy alto, permitiendo la transcripción automática de textos impresos con un nivel de fiabilidad muy elevado y apuntando a la reducción de trabajo para la publicación de ediciones científicas. Sin embargo, para invertir la tendencia de una difusión masiva de materiales textuales dudosos y alimentar la red con textos controlados, en primer lugar, es necesario reparar en el hecho de que la tecnología de por sí no es suficiente: debe estar guiada por una mirada especializada que supervise el desarrollo de los *softwares* de forma concienzuda.

Al respecto, las preocupaciones de Padre Roberto Busa, quien percibió muy pronto el desarrollo informático y computacional como un riesgo para la producción de datos sólidos y eficientes, aptos para la investigación humanística, parecen aún más fundamentales y urgentes. En opinión del jesuita, la nueva velocidad que se impuso al procesamiento del lenguaje y a la digitalización de documentos textuales no constituye de por

sí una respuesta a la degradación de los contenidos informativos de la web, sino que tiene que acompañarse con una interpretación ponderada de carácter inductivo, que busque la evidencia empírica y la producción de unos datos que puedan constituir una documentación fiable y replicable. En otras palabras,

> he foresaw that the wide availability of large collections of digitized textual data and of tools for processing them automatically would run the risk of being incorrectly exploited. Busa believed the greatest danger lay in considering Computational Linguistics (and Digital Humanities, too) not as a discipline aimed at doing things better, but rather as a tool to do things faster, both in the phase of collecting data and in that of exploiting data. He feared that the computational linguists and the digital humanists of the third millennium would cease caring for the quality of data and lose the humility to check them carefully, preferring instead to process huge masses of texts quickly and approximately, without even reading a line (Rockwell-Passarotti, 2019, 26).

En la línea de evitar que la explotación de grandes masas de datos sustituya a la recolección y al enriquecimiento de datos apropiados, los trabajos colaborativos del tipo que tratamos en estas páginas adquieren un notable interés. Estos representan una respuesta concreta a la producción de objetos digitales fiables y, a la vez, se sustentan en una visión nítida del fenómeno, que prevé que el trabajo de los humanistas esté sujeto principalmente al atento análisis de los datos y no en asuntos que atañen por la mayoría a la mejora de las máquinas. Y esto porque solo a partir de datos más consistentes y refinados, la implementación tecnológica podrá sustentar la investigación científica de forma correcta y rigurosa.

La importancia de los tres proyectos implicados y la gran experiencia en tema de edición de impresos de la Edad Moderna de los colaboradores involucrados constituye finalmente la *conditio sine qua non* para el desarrollo del proyecto y un aspecto determinante para que la colaboración proporcione resultados correctos, replicables y explotables por otros estudiosos, capaces de alimentar un conocimiento responsable, documentable y sistemático de la textualidad en el ámbito digital.

Apéndice 1. Modelo de HTR SpanishGothic_extended_sXV-XVI

Descripción Dataset:
Tipo de documentos: impresos
Nr. de palabras: 150 137
Nr. de líneas: 16 816
CER Training Set: 0.45%
CER Validation Set: 0.92%

Autores:
Stefano Bazzaco (coord.), Giada Blasut, Federica Zoppi, Nuria Aranda García, Ángela Torralba Ruberte, Ana Milagros Jiménez Ruiz, Pedro Monteiro

Versión actualmente disponible: versión 1.0.0 (julio 2021)

Cómo citar: Stefano Bazzaco (coord.), Federica Zoppi, Giada Blasut, Nuria Aranda García, Ángela Torralba Ruberte, Ana Milagros Jiménez Ruiz, & Pedro Monteiro. (2021). HTR model SpanishGothic_XV-XVI_extended DATASET (1.0.0) [Data set]. Zenodo.
DOI: <https://doi.org/10.5281/zenodo.4888927>

Enlaces:
https://github.com/stefanobazzaco/HTR-model-SpanishGothic_XV-XVI_extended
https://zenodo.org/record/4888927#.YLX6xqgzY2w
https://readcoop.eu/model/spanish-gothic-15th-16th-century/

Presentación del corpus

Los principios que rigieron la selección del corpus estuvieron condicionados por los intereses del Progetto Mambrino, COMEDIC y BIDISO. En general, perseguimos dos objetivos principales: la aplicación del modelo a un grupo de ediciones muy variado desde un punto de vista editorial (Grupo Misceláneo) y el uso de la herramienta en géneros editoriales homogéneos y consolidados en la época de la imprenta manual (Grupo de Libros de Caballerías, Grupo de Historias breves de caballería).

Desde finales del siglo XV hasta el segundo tercio del XVI, concretamente hasta 1560, el uso de tipos góticos será el hegemónico en la imprenta hispánica. Dentro de este marco temporal y tipográfico, el número de ediciones que forman nuestro corpus asciende a dieciséis (Tabla 5). A partir del género editorial de estos impresos[37], hemos establecido tres partes principales.

En primer lugar, encontramos un grupo misceláneo constituido por un conjunto de ediciones de diversa disposición bibliográfica y género literario. Estas son: el *Doctrinal de los caballeros* de Alonso de Cartagena, *La Fiameta* de Juan Boccaccio, la *Crónica del Rey Don Rodrigo* de Pedro del Corral, también conocida como *Crónica Sarracena*, el *Retablo de la Vida de Cristo* de Juan de Padilla, el «Cartujano», y una nueva edición de la *Tragicomedia de Calisto y Melibea* de Fernando de Rojas.

El segundo grupo está formado por cinco ediciones de libros de caballerías, un género literario cuyas obras, al presentar una disposición editorial homogénea, muestran errores de reconocimiento similares. Las obras escogidas de este género son: el *Lisuarte de Grecia* de Juan Díaz, el anónimo *Florando de Inglaterra*, el *Silves de la Selva* de Pedro de Luján, el *Lisuarte de Grecia* de Feliciano de Silva y el *Leandro el Bel* de Pedro de Luján.

Por último, son seis las obras del género de las historias breves de caballerías que componen el tercer grupo: tres ediciones diferentes de *El libro del conde Partinuplés* y, respectivamente, una de la *Historia de la linda Magalona*, la *Historia de la reina Sebilla* y la *Historia del rey Canamor*.

[37] Entre los muchos trabajos sobre la cuestión del «género editorial» de Víctor Infantes (especialmente, 1992), consideramos muy acertada la síntesis que presenta en Infantes (2003).

Título	Año	Lugar de impresión	Tipología	Formato	Disposición del texto	n.º de ficha en COMEDIC
Doctrinal de los caballeros	1487	Burgos. Fadrique de Basilea	Miscelánea	Folio - 168 h.	Línea tirada	82
La Fiameta	1497	Salamanca. Impresor de la Gramática de Nebrija	Miscelánea	Folio - 44 h.	Dos columnas	147
Crónica del Rey Don Rodrigo	1499	Sevilla. Meinardo Ungut y Estanislao Polono	Miscelánea	Folio - 227 h.	Dos columnas	53
Retablo de la Vida de Cristo	1510	Sevilla. Juan Cromberger	Miscelánea	Folio - 58 h.	Dos columnas	309
Tragicomedia de Calisto y Melibea	[1512 - 1515]	Roma. Marcelo Silber	Miscelánea	4º - 80 h.	Línea tirada	322
Lisuarte de Grecia	1526	Sevilla. Jacobo y Juan Cromberger	Libro de caballerías	Folio - 123 h.	Dos columnas	/
Lisuarte de Grecia	1550	Sevilla. Jácome Cromberger	Libro de caballerías	Folio - 109 h.	Dos columnas	/
Florando de Inglaterra	1545	Lisboa. Germán Gallarde	Libro de caballerías	Folio - 172 h.	Dos columnas	/
Silves de la Selva	1549	Sevilla. Dominico de Robertis	Libro de caballerías	Folio - 150 h.	Dos columnas	/
Leandro el Bel	1563	Toledo. Miguel Ferrer	Libro de caballerías	Folio - 128 h.	Dos columnas	/
El libro del conde Partinuplés	1519	Sevilla. Jacobo Cromberger	Historia breve de caballerías	4º - 95 h.	Línea tirada	106
El libro del conde Partinuplés	1558	Burgos. Herederos de Juan de Junta	Historia breve de caballerías	4º - 86 h.	Línea tirada	106
El libro del conde Partinuplés	1563	Burgos. Felipe de Junta	Historia breve de caballerías	4º - 86 h.	Línea tirada	106
Historia de la linda Magalona	1519	Sevilla. Jacobo Cromberger	Historia breve de caballerías	4º - 63 h.	Línea tirada	213
Historia de la reina Sebilla	1551	Burgos. Felipe de Junta	Historia breve de caballerías	4º - 70 h.	Línea tirada	/
Historia del rey Canamor	1527	Valencia. Jorge Costilla	Historia breve de caballerías	4º - 110 h.	Línea tirada	347

Tabla 5. Listado de las obras del corpus para el modelo *SpanishGothic*

a) Grupo misceláneo

Las ediciones que integran este grupo presentan como común denominador el haber sido impresas en los primeros decenios de implantación y consolidación de la imprenta hispánica, tres de ellas durante el periodo incunable y dos en el periodo postincunable. Durante este periodo de evolución editorial (González-Sarasa Hernáez, 2013), el formato de las ediciones determina en gran medida la configuración de los elementos que constituyen la caja de escritura –cabeceras, letras xilográficas y lombardas, grabados, foliación– y estos, a su vez, condicionarán la aplicación del *layout*. Por tanto, la exposición de nuestro caso práctico está regida por la realidad bibliográfica de las ediciones.

En lo que se refiere a los incunables, a pesar de la variedad en su género literario, tanto la edición del *Doctrinal de los Caballeros* impresa en Burgos en 1487 (el 20 de junio) por Fadrique de Basilea[38], como *La Fiameta* publicada en Salamanca en 1497 y atribuida al Impresor de la Gramática de Nebrija[39], y la *Crónica del Rey Don Rodrigo* (*Crónica Sarracina*) impresa en Sevilla en 1499 por Meinardo Ungut y Estanislao Polono[40] presentan un formato en folio. La caja de escritura del *Doctrinal* muestra el texto literario en una columna de 35 líneas –a pesar de que el título y las tablas están dispuestos a doble columna–, mientras que el texto de *La Fiameta* y la *Crónica Sarracina* se divide en dos columnas de 48 y 47 líneas, respectivamente.

La aplicación del análisis de la plana –*Layout Analysis*– ha sido exitosa en casi todas las partes de la caja tipográfica de los ejemplares del *Doctrinal* y de la *Crónica* –incluida la cabecera–. En el caso de *La Fiameta*, se ha tenido que reconstruir de forma manual en más de la mitad de las hojas, puesto que la digitalización del ejemplar presenta una baja calidad y multitud de manchas, fruto del deterioro del ejemplar. La segmentación ha funcionado solamente en la identificación de las dos columnas, pero no en la

[38] Se ha trabajado con la digitalización del ejemplar conservado por la Real Academia Española, con la signatura Inc. San Román 6. Digitalización en color.

[39] Se ha trabajado con la digitalización del ejemplar localizado en la Pierpont Morgan Library de Nueva York (Incunable Collection-Oversize: INCUNOS1, ChL 1742: PML 667). Digitalización en blanco y negro.

[40] Se ha trabajado con la digitalización del ejemplar localizado en la Hispanic Society of America (signatura Inc. 84).

separación de las líneas de cada una y, de hecho, ha habido varias hojas que han resultado irreconocibles para el modelo.

En los tres casos, se ha hallado dificultad solamente en el reconocimiento de las letras xilográficas y lombardas mayúsculas, que dan comienzo al texto, y que se han transcrito manualmente. En *La Fiameta*, no se ha reconocido la xilográfica mayúscula inicial, pero en el resto de lombardas el programa ha detectado un elemento no identificado para el que ha dejado un espacio libre –a completar por el usuario del programa. El reconocimiento de estas letras ilustradas es una de las limitaciones de Transkribus ya que, para una correcta identificación de las mismas –especialmente las xilográficas–, sería preciso emplear programas destinados al reconocimiento de grabados[41].

En el caso del *Retablo de la Vida de Cristo*, obra de gran éxito editorial por el elevado número de ediciones[42], se ha escogido la edición impresa en Sevilla en 1510 por Juan Cromberger[43]. Presenta un formato en folio y una disposición en la caja de escritura muy elaborada: además de la cabecera, el texto principal está dividido en dos columnas con 52 líneas (máx.) que incluyen grabados xilográficos, junto con los nombres de los personajes ficcionales que intervienen y que aparecen en ambos márgenes. De este modo, en total, el *Layout Analysis* había de reconocer cuatro columnas. Sin embargo, esta compleja distribución sumada a la baja calidad de la digitalización ha provocado que el programa presente problemas en la identificación de la cabecera, los grabados y las *marginalia*, que se identificaban dentro de la misma *baseline* del texto principal (Figura 1). Todos estos casos han tenido que resolverse de forma manual.

[41] A este respecto, conviene destacar los programas de OCR de ilustraciones empleados por grupos de investigación como 15cBookTrade de la Universidad de Oxford <http://15cbooktrade.ox.ac.uk/> (cons. 15/05/2022) o The Illustrated book in Lyon 1480-1600 - Equipex Biblissima, dirigido por la Dra. Barbara Tramelli.

[42] Véase la ficha 309 en Comedic: *Catálogo de obras medievales impresas en castellano hasta 1600* [en línea] <http://grupoclarisel.unizar.es/comedic/> (cons. 15/05/2022).

[43] Se ha trabajado con el ejemplar de la Biblioteca Nacional de España (signatura R/31133-3).

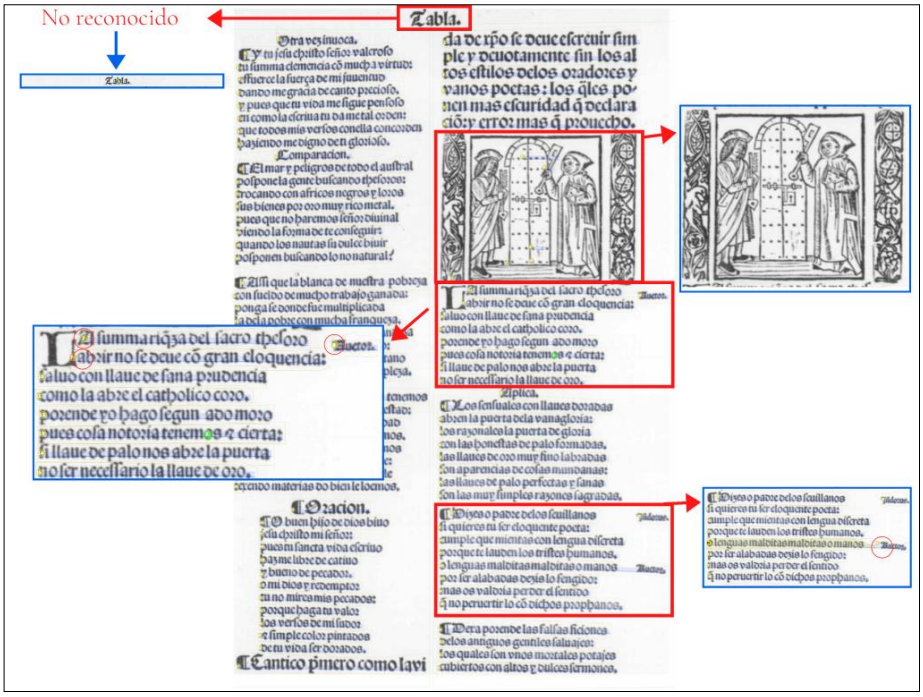


Fig. 1. Muestra del *Layout Analysis* fallido en el ejemplar del *Retablo de la Vida de Cristo*

Respecto a la *Tragicomedia de Calisto y Melibea* de Fernando de Rojas impresa en Roma entre 1512-1515 por Marcello Silber[44], presenta un formato en cuarto con una caja de escritura constituida por una columna de 37 líneas. El proceso de *Layout Analysis* ha reconocido todas las *baselines*, incluida la secuenciación de personajes que aparece al inicio de cada auto a modo de *dramatis personae*. Al presentar los mismos tipos que el texto principal, el programa ha leído el compendio de personajes como una línea de texto al uso; por lo que se ha tenido que indicar de forma manual que constituye otro campo. Al igual que los grabados del *Retablo*, las características figuritas factótum celestinescas se han identificado como campos textuales. Se trata de un error que solamente se ha producido en una ocasión, ya que el programa ha sabido discriminar correctamente los grabados y obviarlos en el resto de páginas.

[44] Se ha trabajado con la digitalización del único ejemplar conocido, y localizado en la Biblioteca Estatal de Ulm (signatura: Schad 4434). Disponible en: <http://openaccess-stadtbibliothek.ulm.de/pdf/Reproduktionen/Schad_4434/?C=M;O=A> (cons. 10/05/2022).

b) Textos de temática caballeresca

Se han empleado once textos de temática caballeresca en el entrenamiento del modelo *Gothic Extended* en el programa *Transkribus*. Por sus rasgos materiales y de *dispositio* textual, podemos dividirlos en dos grupos diferentes, que nos permiten explicarlos en conjunto.

Por un lado, cinco textos que forman parte del género de los libros de caballerías. En primer lugar, dos ediciones diferentes del *Lisuarte de Grecia*: la impresa por Jacobo y Juan Cromberger en 1526 (BNE: R/71) y la edición publicada a cargo de Jácome Cromberger en 1550 (BNE: R/13138(2)). La edición del *Silves de la Selva* de Pedro de Luján impresa en Sevilla por Dominico de Robertis en 1549 (BNE: R/865); la del *Florando de Inglaterra*, impreso en Lisboa por Germán Gallarde en 1545 (British Library: C.62.h.14); y, por último, el ejemplar de la Biblioteca Nacional de España (R/9030) del *Leandro de Bel* de Pedro de Luján en la edición impresa en Toledo por Miguel Ferrer en 1563.

Por otro lado, seis impresos que constituyen parte del conjunto de textos que en los últimos años se han denominado «historias breves de caballerías»[45]. Dentro de esta serie, hemos empleado tres ediciones diferentes de *El libro del conde Partinuplés*: la edición sevillana impresa por Jacobo Cromberger en 1519 (BNL: Res. 401/18) y dos testimonios burgaleses. El primero data de 1558 y fue impreso por los herederos de Juan de Junta (BNE: R/31364/38) y el segundo en 1563 por Felipe de Junta (British Library: C.55.d.4). Además, también hemos utilizado el ejemplar *unicum* de la *Historia de la linda Magalona*, impreso en Sevilla por Jacobo Cromberger en 1519 (British Library: C.7.a.18); la edición de la *Historia de la reina Sebilla* impresa en 1551 por Felipe de Junta (Bibliothèque Nationale de France: Rés. Y2849) y, por último, la edición valenciana de 1527 de la *Historia del rey Canamor*, impresa por Jorge Costilla (Universidad de Oviedo: CEA-227).

[45] Véanse al respecto: Baranda (1991, 183-191); Infantes (1991, 165-182 y 1996, 127-132).

Libros de caballerías

Desde el punto de vista del cuerpo del texto, los cinco textos presentados, al pertenecer al género literario caballeresco, se caracterizan por ser impresos en formato folio, tamaño idóneo para la edición de obras extensas. De hecho, «el formato del género editorial caballeresco que se imprime en talleres peninsulares se limita al folio» (Lucía Megías, 2000, 431). En cuanto a la disposición del interior del libro de caballerías, el texto de este tipo de obras aparece siempre distribuido en dos columnas, así como sucede con las tablas de capítulos, mientras que el de los preliminares legales (privilegio, licencia, aprobación, fe de erratas y tasa) y el de los prólogos aparece siempre a línea tirada (2000, 448). Asimismo, estos textos presentan los títulos o cabeceras centradas en la parte superior de cada una de las planas. Por último, en la esquina superior derecha se coloca la foliación de las páginas mediante números romanos (Figuras 2 y 3).

Fig. 2 y 3. Ejemplos de cabeceras en el *Lisuarte de Grecia* (Sevilla, 1526), fols. 4v-5r.

Como se observa en las imágenes, estas obras presentan el modelo más habitual de cabecera: el título se reparte entre el vuelto y el recto de los folios, de modo tal que al tener abierto el impreso podamos leer el título completo en la parte superior. De acuerdo con la clasificación de Lucía Megías, se trata de las «cabeceras que indican el lugar que ocupan el texto en una serie más amplia de libros» (2000, 451).

Ambos rasgos propios de los libros de caballerías, tanto la división del texto en dos columnas como la introducción de la cabecera, no generan ningún problema con el programa Transkribus al aplicar el modelo de segmentación *2columns+heading* creado por medio de la función P2PaLA (*Page to Page Layout Analysis*), como se puede ver en la siguiente imagen (Fig. 4):

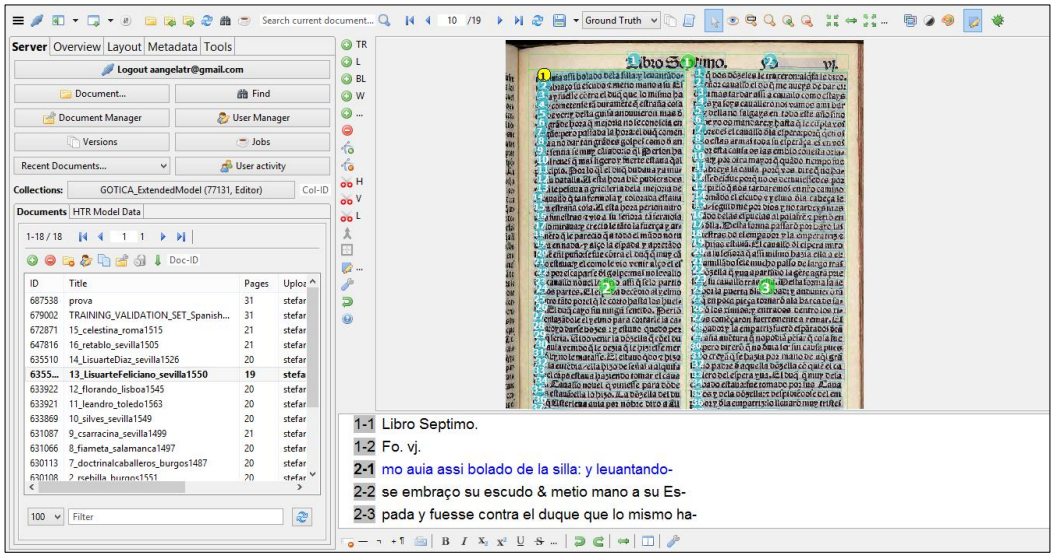


Fig. 4. Segmentación con modelo de P2PaLA del *Lisuarte de Grecia* (Sevilla, 1550)

Asimismo, en todos los casos la primera línea del epígrafe de cada capítulo presenta un tipo diferente y de mayor tamaño. Como ejemplo paradigmático que se puede observar en las figuras, ofrecemos la edición del *Lisuarte de Grecia* de 1526 que emplea para la primera línea del epígrafe una letrería G158 (T:2(C)) y para el resto del texto un tipo G98 (T:8b).

Tras dichos epígrafes que encabezan cada capítulo, el texto se inicia

con una capital xilográfica, que en la edición que hemos tomado como referencia aparece recuadrada en un ribete que ocupa cuatro líneas.

Además, tras varios epígrafes, se inserta un pequeño grabado en blanco y negro que ilustra el contenido de dicho capítulo y ocupa once líneas. Por último, también las portadas aparecen decoradas con un grabado. Por ejemplo, el *Lisuarte de Grecia* ha ilustrado su portada con una xilografía a dos tintas, negra y roja, que ha enmarcado por una orla. La ilustración de la portada contiene seis grabados que representan a diferentes caballeros del ciclo (Fig. 5):

Fig. 5. Portada del *Lisuarte de Grecia* (Sevilla, 1526)

Estos elementos decorativos y explicativos, tanto los que se ubican en la portada como los del interior del relato, han sido pasados por alto por el programa, puesto que la detección del *layout* con imágenes en colores no considera relevantes las zonas con densidad de píxeles menor que la del cuerpo del texto.

Historias breves de caballerías

A diferencia de los extensos libros de caballerías, esta serie de textos se caracteriza por la brevedad, ya que no llegan habitualmente a ocho pliegos (64 páginas) y suelen estar impresos en formato 4º, en lugar de folio. Como resultado de la reducción material, tampoco requieren un excesivo material gráfico ni una especial disposición impresa: el texto se presenta a línea tirada, desaparecen las cabeceras y, por último, la paginación, también en números romanos, se traslada a la esquina inferior derecha de la página. La nueva distribución de la *mise en page* implica que no hay necesidad de aplicar modelos preconcebidos de P2PaLA en Transkribus (Fig. 6).

Por otro lado, de manera similar a las ediciones del *Lisuarte de Grecia*, la primera línea del epígrafe presenta un tipo diferente y de tamaño mayor. En la *Historia del rey Canamor* se emplean tipos de tres fundiciones: para el título Gótica-38: c.190-G; para el epígrafe inicial Gótica-15B: c.132-G y para el texto restante Gótica-22: 100-G (Fig. 7).

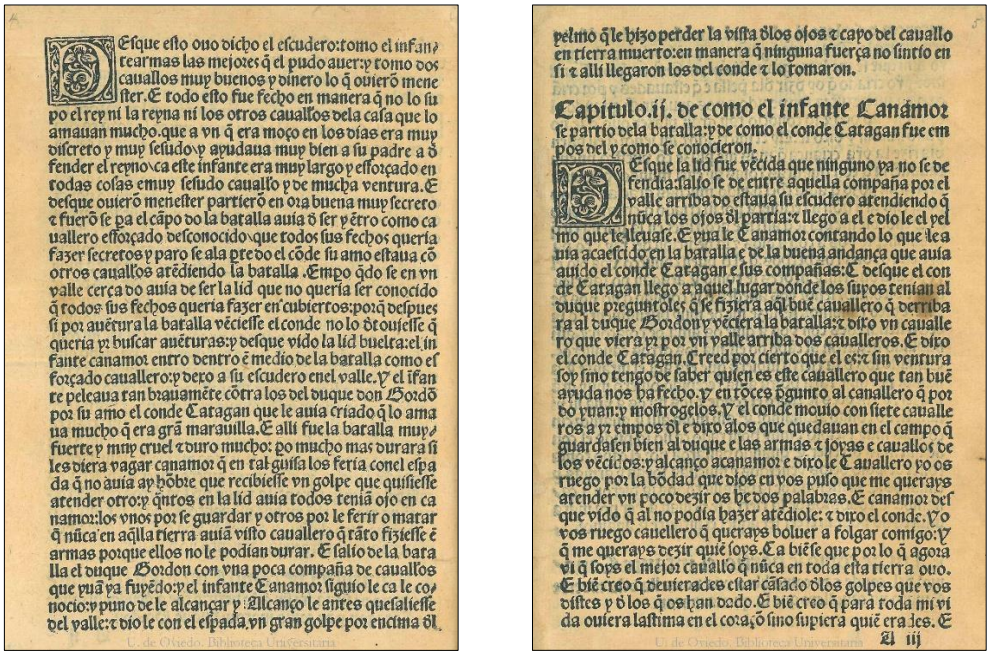


Fig. 6 y 7. Distribución, tamaño de texto y epígrafe en el *Libro del Rey Canamor* (Valencia, 1527)

Por último, todos estos textos incluyen un grabado únicamente en la portada del relato. Es decir, a diferencia de los ejemplares del *Lisuarte de Grecia*, no se ha ilustrado el interior de los testimonios.

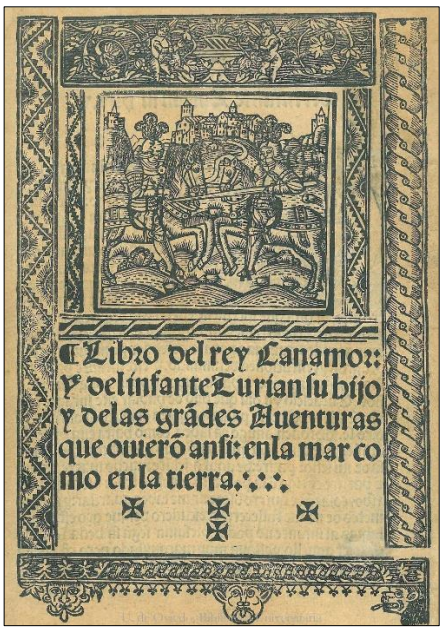


Fig. 8. Portada del *Libro del Rey Canamor* (Valencia, 1527)

El grabado, que representa un combate de dos caballeros con una ciudad de fondo con una pieza, se encuentra dentro de un marco formado por cuatro piezas xilográficas y encima del título.

A pesar de la distinción en cuanto a la *mise en page* entre los libros de caballerías y las ediciones que se corresponden con las historias breves de caballerías, en todos los casos Transkribus ha delimitado correctamente el *layout*. De hecho, el reconocimiento de los grabados de portada e interiores, en el caso de los libros de caballerías, ha sido omitido de manera automática; mientras que las capitales xilográficas –debido a su composición artística–, no han sido reconocidas por el programa y se han excluido manualmente de la delimitación de líneas.

Recapitulación letra gótica

En conclusión, realizado el recorrido por estos textos de acuerdo con los principios descriptivos de la bibliografía, observamos lo siguiente: las características editoriales y bibliográficas de los dieciséis impresos en letra gótica que constituyen el *dataset* del modelo son muy variadas, pero no afectan de forma significativa al reconocimiento del texto. Podríamos dividir estos textos impresos en folio (*Lisuarte de Grecia, Doctrinal de los Caballeros, La Fiameta, Crónica del Rey Don Rodrigo, Retablo de la Vida de Cristo, Florando de Inglaterra, Silves de la Selva* y *Leandro el Bel*) o en cuarto (*Partinuplés, Magalona, Reina Sebilla, Tragicomedia de Calisto y Melibea*); en ediciones con el texto presentado a doble columna, el cual plantea más problemas por la segmentación de la imagen (*Lisuarte de Grecia*, *La Fiameta, Crónica del Rey Don Rodrigo, Retablo de la Vida de Cristo, Florando de Inglaterra, Silves de la Selva* y *Leandro el Bel*) o a línea tirada (*Partinuplés, Magalona, Reina Sebilla, Tragicomedia de Calisto y Melibea, Doctrinal de los Caballeros*); o por contener grabados internos (*Lisuarte, Tragicomedia de Calisto y Melibea*) o prescindir de ellos. Si bien, dicha casuística no supone ningún problema insalvable en el manejo de Transkribus, solamente aparecen dificultades cuando las digitalizaciones son defectuosas o de baja calidad, o el ejemplar está dañado.

Apéndice 2. Modelo de HTR SpanishRedonda_extended_sXVI-XVII

Descripción Dataset:
Tipo de documentos: impresos
Nr. de palabras: 61 938
Nr. de líneas: 7 675
CER Training Set: 0.21%
CER Validation Set: 1.07%

Autores:
Stefano Bazzaco (coord.), Gaetano Lalomia, Daniela Santonocito, Manuel Garrobo Peral, Mónica Martín Molares, Carlota Cristina Fernández Travieso

Versión actualmente disponible: versión 1.0.0 (julio 2021)

Cómo citar: Stefano Bazzaco (coord.), Gaetano Lalomia, Daniela Santonocito, Manuel Garrobo Peral, Mónica Martín Molares, & Carlota Cristina Fernández Travieso (2021). HTR model SpanishRedonda_XVI-XVII_extended DATASET (1.0.0) [Data set]. Zenodo.
DOI: <https://doi.org/10.5281/zenodo.4889218>

Enlaces:
https://github.com/stefanobazzaco/HTR-model-SpanishRedonda_XVI-XVII_extended
https://zenodo.org/record/4889218#.YmUUF9pBw2x
https://readcoop.eu/model/spanish-redonda-round-script-16th-17th-century/

Al igual que los dieciséis impresos seleccionados para el corpus en gótica, para la creación del *dataset* que constituye la base del entrenamiento del modelo *SpanishRedonda* en el programa Transkribus se empleó una quincena de impresos, que podemos dividir en dos grupos.

Por un lado, se identifican cinco obras de carácter histórico-caballeresco, publicadas entre 1578 y 1607. Todas ellas escritas en verso con una disposición textual variada (ya sea en una o dos columnas). Hacemos referencia a dos obras de Cristóbal de Mesa: el libro primero de la *Restauración de España* (impreso en Madrid por Juan de la Cuesta en 1607) y el argumento del canto primero de *Las navas de Tolosa* (Madrid, viuda de P. Madrigal, 1594). A estas dos sumamos la *Historia de las hazañas y hechos del invencible caballero Bernardo del Carpio*, de Agustín Alonso (Toledo, Pero López de Haro, 1585); el canto primero de *Las lágrimas de Angélica*, de Luis Barahona de Soto (Granada, Hugo de Mena, 1586) y el canto primero de la segunda parte del *Libro del Orlando determinado que prosigue la materia de Orlando el Enamorado*, compuesto por Don Martín de Bolea y Castro (Lérida, Miguel Prats, 1578).

Por otro lado, constituye el grueso del corpus una decena de relaciones de sucesos, sobre las que nos detendremos en el apartado siguiente para comentar sus características bibliográficas. Y es que, con motivo de poder ofrecer la transcripción de estos textos en acceso abierto en el portal del proyecto *Biblioteca Digital Siglo de Oro (BIDISO)*[46], se consideró fundamental la creación de un modelo de transcripción en redonda. Para ello, se fue elaborando un corpus documental sobre distintas colecciones vinculadas con las fuentes primarias de las bibliotecas digitales del proyecto BIDISO.

Relaciones de sucesos

Entendemos por *relaciones de sucesos* «aquellos documentos noticieros que solían versar sobre asuntos muy diversos y cuya forma era variada. Podían ser manuscritas o impresas, estar en verso o prosa, y constar de un

[46] Se pretende, además, que estos textos transcritos sean enriquecidos con marcación formal y semántica, a través de una codificación en XML-TEI, para la creación de ediciones académicas digitales. Véase, en este mismo número, Fernández Travieso y Garrobo Peral (2022).

solo pliego o, incluso, llegar a tener las dimensiones de un libro voluminoso»[47]. Por tanto, se evidencia ya en su propia definición el carácter tan heterogéneo, y a su vez complejo, que presenta este género editorial[48]. Como vemos, se hace mención a distintos aspectos: la temática variada, la modalidad del discurso o la extensión e, incluso, la forma de difusión (que, si bien puede ser manuscrita o impresa, por razones obvias al estudiar una tipografía concreta, nos centramos solo en esta última).

Habida cuenta de las particularidades del género, se buscó una selección lo más representativa posible de estos textos (Tabla 6). De ahí que se hayan empleado relaciones de sucesos con diferente temática, impresos por distintos tipógrafos en talleres ubicados en emplazamientos variados –peninsulares, como Madrid, Valencia, Sevilla o Cuenca, y extranjeros como Roma, Bruselas o Lima– y que abarcasen un abanico de años lo más amplio posible.

Tipología

Buena parte de los acontecimientos festivos, políticos y sociales de la Edad Moderna fueron plasmados en los impresos noticieros para dejar memoria de lo ocurrido. En este corpus de relaciones que hemos empleado (Tabla 6), vemos cómo destaca una tipología: las relaciones de ceremonias o festejos. De estas conservamos relaciones más extensas, que daban lugar a los llamados *libros de fiestas*. Aunque parezca una obviedad, tal particularidad influye en la configuración misma de los volúmenes, tanto en el contenido como en la forma. Debido a esto, y ante las necesidades iniciales de proveer a Transkribus del mayor número de páginas posible para poder crear el modelo, optamos por servirnos de este subgénero caracterizado por una extensión mayor principalmente por ser «recomendable transcribir [al menos] una veintena de páginas manualmente» (Bazzaco, 2020, 548).

[47] Información disponible en el portal BIDISO <https://www.bidiso.es/estaticas/ver.htm?id=6> (cons. 15/05/2022). Para una aproximación a las Relaciones de Sucesos (RdS), véanse Pena Sueiro (2001) y Pena Sueiro y Ruiz Astiz (2019).
[48] Infantes (1996, 208).

Título	Año	Lugar de impresión	Tipología	Form.	Columnas	CBDRS
Relación de la solemne entrada hecha en Ferrara a los 13 días de noviembre MDXCVIII por la serenísima Margarita de Austria	1598	Roma. Nicolás Mucio	Ceremonias y festejos. Entrada	4º 12 h.	1 (prosa)	0001160 B
Relacion del aparato que se hizo en la ciudad de Valencia para el recebimiento de la serenissima reina doña Margarita de Austria	1599	Valencia. Pedro Patricio Mey	Ceremonias y festejos. Entrada	8º 16 h.	1 (prosa)	0002620 A
Relación del nacimiento del nuevo infante y de la muerte de la reina nuestra señora	1612	Cuenca. Salvador Viader	Ceremonias y festejos. Nacimiento y exequias	4º 2 h.	2 (verso)	0002764 A
Relación verdadera del acompañamiento y bautismo de la serenísima princesa Margarita María Catalina	1623	Madrid. Diego Flamenco	Ceremonias y festejos. Bautismo	Folio 2 h.	1 (prosa)	0004060 B
Fiesta que se hizo en Aranjuez a los años del Rey Nuestro Señor	1623	Madrid. Juan de la Cuesta	Ceremonias y festejos.	4º 26 h.	1 (prosa)	0007066 A
Relación verdadera en que se da cuenta de todo el daño que causó las crecientes del rio Guadalquivir en la ciudad de Sevilla y Triana	1626	Lima. Gerónimo de Contreras	Suceso extraordinario de la naturaleza	Folio 2 h.	1 (prosa)	0007022 A
Relacion de las fiestas que se han hecho en la fidelissima Ciudad de Napoles por el nazimiento del Principe N.S. […], hasta cinco de Março de este año de 1658	[1658]	[s.l.] [s.n.]	Ceremonias y festejos. Nacimiento	4º 12 h.	1 (prosa)	0007308 A
Primera parte de la relación de las reales disposiciones […] jornada a la provincia de Guipuzcoa a entregar a la serenísima	1660	Sevilla. Juan Gómez de Blas	Acontecimiento político. Relaciones de viajes	4º 4 h.	1 (prosa)	0003549 A
Segunda parte de la relación diaria del itinerario que su Majestad ha seguido desde que salió de Madrid hasta llegar a Fuenterrabía	[1660]	Sevilla. Juan Gómez de Blas	Acontecimiento político. Relaciones de viajes	4º 4 h.	1 (prosa)	0002944 A
Relación de un nuevo milagro obrado por intercesión del glorioso apóstol de las Indias, san Francisco Xavier	1663	Bruselas	Suceso extraordinario. Milagro	4º 4 h.	1 (prosa)	0003722 A

Tabla 6. Listado de relaciones de sucesos empleadas para el modelo *SpanishRedonda*

En la selección aparecen cuatro libros de fiestas, es decir, cuatro relaciones que exceden el volumen de un pliego de cordel: la *Relación de la solemne entrada hecha en Ferrara [...]*, de Ioan Paolo Mocante (Roma, Nicolás Mucio, 1598)[49]; la *Relación del aparato que se hizo en la ciudad de Valencia […]*, de Juan Bautista Confalionero (Valencia, Patricio Mey, 1599)[50]; la *Fiesta que se hizo en Aranjuez a los años del rey nuestro señor don Felipe IIII [...],* de don Antonio de Mendoza (Madrid, Juan de la Cuesta, 1623)[51] y la *Relación de las fiestas que se han hecho en la fidelísima ciudad de Nápoles por el nacimiento del príncipe […]* ([1658])[52]. En total, estaríamos hablando de 101 páginas, lo que supone alimentar a Transkribus con 3249 líneas con solo estas cuatro ediciones.

Portada y grabados

Además, son justo «estos libros [de fiestas], más que ninguna otra clase de relaciones, [los que] suelen adornarse con ilustraciones de grabados xilográficos o calcográficos» (López Poza, 1999, 220). Constatan esta afirmación las portadas de las ediciones que transcribimos (Fig. 9).

Generalmente, en ellas se incluían elementos heráldicos (reales, papales o nobiliarios), dependiendo del editor o promotor en algunos casos, o bien los motivos xilográficos que poseía un impresor en su taller. Así, podemos ver en la portada de la izquierda el escudo papal y el real, en medio el del Reino de Valencia y, a la derecha, el escudo de Felipe IV (Figs. 9a y 9b). La edición sobre las fiestas de Nápoles no incorpora ningún grabado, pero, por la disposición de la portada, no parece descabellado sospechar que podría haberse reservado el espacio inferior de la misma para incluir uno (Fig. 9c).

---

[49] Ejemplar conservado en la Biblioteca Valenciana, sign. XVI/F-33. Disponible en: <https://bivaldi.gva.es/es/catalogo_imagenes/grupo.cmd?presentacion=pagina&posicion=1&path=1004644&registrardownload=0&texto_busqueda=&interno=S> (cons. 25/04/2022).

[50] Ejemplar conservado en la British Library, sign. 9930.aa.9. Disponible en: <https://www.bl.uk/treasures/festivalbooks/pageview.aspx?strFest=0142&strPage=001> (cons. 25/04/2022).

[51] Ejemplar conservado en la Biblioteca Nacional de España, sign. R/15515. Disponible en: <http://bdh-rd.bne.es/viewer.vm?id=0000052047&page=1> (cons. 25/04/2022).

[52] Ejemplar conservado en la Biblioteca Nacional de España, sign. VE/1558/11. Disponible en: <https://archive.org/details/relaciondelasfie00napl/mode/2up> (cons. 25/04/2022).


RELACION
DE LA SOLENE ENTRADA HECHA EN FERRARA
a los 13. dias de Noviembre M. D. XCVIII.
Por la Sereniſſima S. Doña Margarita de Auſtria Reyna de Eſpaña, y del Conſiſtorio publico con todos los aparatos que ſu S.S.N. Clemente Papa VIII. mando hazer y hizo para tal affecto.
Con particular auiſo de la Miſſa Pontifical que canto ſu Beatitud, y de las Ceremonias de los deſpoſorios que ſe hizieron en la Iglefia Cathedral de dicha Ciudad Domingo alos 15. del meſmo con la ceremonia de la Roſa que ſu S. acabada la Miſſa dio alla Reyna.
Compueſta por Ioan Paolo Mecante Vno de los Maeſtros de ceremonias de la Capilla de ſu S. en lengua Italiana, y traducida, e interpretada, en lengua Eſpañola por el Doctor Andres Catalan natural de la Ciudad de Toro.
CON LICENCIA DE LOS SVPERIORES.
EN ROMA.
Impreſſa por Nicolas Mucio. Año M. D. XCVIII.
En plaça Capranica.

RELACION DEL
APARATO QVE SE
HIZO EN LA CIVDAD DE
Valencia para el recebimiento de la Sereniſſima Reyna Doña Margarita de Auſtria deſpoſada con el Catholico y potentiſſimo Rey de Eſpaña Don Phelipe Tercero deſte nombre.
Vendenſe en caſa de Franciſco Miguel a la calle de Caualleros.

RELACION
DELAS FIESTAS
Que ſe han hecho en la Fideliſsima Ciudad de Napoles.
POR EL NAZIMIENTO DEL
PRINCIPE N.S.
Que DIOS guarde,
Haſta cinco de Março de eſte año de 1658.

Fig. 9. Portadas de algunas RdS incluidas en el corpus

Las portadas de las demás relaciones constan de un encabezado a modo de título. Este podía ocupar desde tres líneas, como en el caso de la *Relación verdadera del acompañamiento y bautismo de la serenísima princesa Margarita María Catalina* (Madrid, Diego Flamenco, 1623)[53] hasta casi media plana, como en la *Relación de un nuevo milagro obrado por intercesión del glorioso apóstol de las Indias, san Francisco Xavier, en 2 de septiembre de 1662* (Palermo, 1663)[54]. En estos encabezados suele destacarse tipográficamente –generalmente con el empleo de mayúsculas– algunos elementos textuales del título y se puede incluir otra información sobre la edición como el pie de imprenta (lugar de impresión, nombre del impresor y/o del costeador, fecha), los datos legales (mención del privilegio, de la licencia, de la tasa, etc.) u otras menciones o alusiones (mención de edición, autor, traductor). Así, en el ejemplo de la relación sobre el milagro de san Francisco Javier, se añade: *en Palermo de Sicilia, aprobado por el ilustrísimo arzobispo de dicha ciudad. Según la copia italiana, impresa en Palermo el*

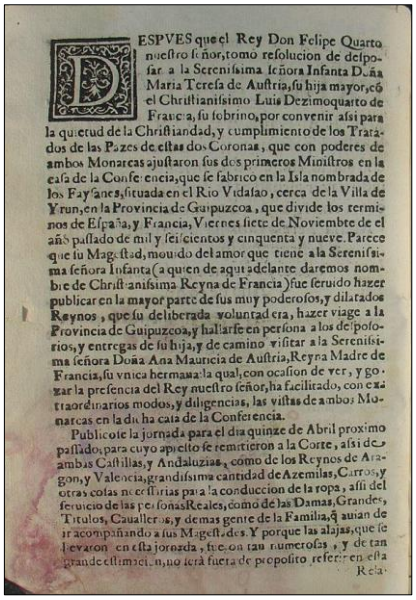
[53] Ejemplar conservado en el fondo antiguo de la Biblioteca de la Universidad de Sevilla, sign. A 109/085(033). Disponible en: <https://archive.org/details/A109085109> (cons. 25/04/2022). Puede consultarse la edición crítica anotada y un estudio sobre las tres ediciones de esta relación en Martín Molares (2021).

[54] Ejemplar conservado en el fondo antiguo de la Biblioteca de la Universidad de Sevilla, sign. A 111/025(17). Disponible en: <https://archive.org/details/A11102517> (cons. 25/04/2022).

mes de agosto de 1663 y sacada de la copia francesa, impresa en Bruselas a 5 de septiembre de dicho año.

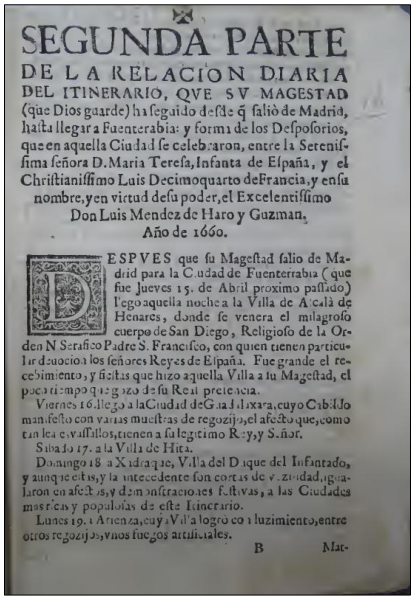
No tan frecuente es el caso de la *Primera parte de la relación de las reales disposiciones y majestuosos aparatos con que su Majestad […] se ha servido hacer jornada a la provincia de Guipúzcoa, a entregar a la serenísima señora doña María Teresa Bibiana de Austria, su hija, al cristianísimo Luis decimocuarto de Francia, su esposo* (Sevilla, Juan Gómez de Blas, 1660). En esta breve relación, de cuatro hojas en formato 4.º, se reserva el primer recto a la portada. En ella se incluye el título ya citado, el escudo real xilográfico, la mención a la licencia y, tras un filete, el pie de imprenta. Se entiende que pueda deberse al reclamo editorial por tratarse de una relación festiva sobre un acontecimiento monárquico tan relevante como la unión de Luis XIV con María Teresa de Austria. Sin embargo, la *Segunda parte de la relación diaria* […], en línea con lo anteriormente mencionado, ocupa aproximadamente un tercio del primer recto.

En el interior de las relaciones de sucesos con las que se ha trabajado no encontramos otros grabados. Solo adornaban algunos textos las letras capitulares xilográficas al inicio del impreso. Cinco de las diez relaciones, todas ellas político-festivas, añaden este adorno tipográfico. Por ejemplo, las dos relaciones sevillanas –la primera y la segunda parte del viaje para la entrega de la novia– comienzan por la misma letra (*D*), por lo que el impresor Juan Gómez de Blas utilizó el mismo grabado xilográfico (Fig. 10).



DESPVES que el Rey Don Felipe Quarto nuestro señor, tomo resolucion de desposar a la Serenissima señora Infanta Doña Maria Teresa de Austria, su hija mayor, cõ el Christianissimo Luis Dezimoquarto de Francia, su sobrino, por convenir assi para la quietud de la Christiandad, y cumplimiento de los Tratados de las Pazes de estas dos Coronas, que con poderes de ambos Monarcas ajustaron sus dos primeros Ministros en la casa de la Conferencia, que se fabrico en la Isla nombrada de los Faysanes, situada en el Rio Vidasao, cerca de la Villa de Yrun, en la Provincia de Guipuzcoa, que divide los terminos de España, y Francia, Viernes siete de Noviembre de el año passado de mil y seiscientos y cinquenta y nueve. Parece que su Magestad, mouido del amor que tiene a la Serenissima señora Infanta (a quien de aqui adelante daremos nombre de Christianissima Reyna de Francia) fue seruido hazer publicar en la mayor parte de sus muy poderosos, y dilatados Reynos, que su deliberada voluntad era, hazer viage a la Provincia de Guipuzcoa, y hallarse en persona a los desposorios, y entregas de su hija, y de camino visitar a la Serenissima señora Doña Ana Mauricia de Austria, Reyna Madre de Francia, su vnica hermana: la qual, con ocasion de ver, y gozar la presencia del Rey nuestro señor, ha facilitado, con extraordinarios modos, y diligencias, las vistas de ambos Monarcas en la dicha casa de la Conferencia.

Publicose la jornada para el dia quinze de Abril proximo passado, para cuyo apresto se remitieron a la Corte, assi de ambas Castillas, y Andaluzias, como de los Reynos de Aragon, y Valencia, grandissima cantidad de Azemilas, Carros, y otras cosas necessarias para la conduccion de la ropa, assi del seruicio de las personas Reales, como de las Damas, Grandes, Titulos, Caualleros, y demas gente de la Familia, q̃ auian de ir acompañando a sus Magestades. Y porque las alajas, que se llevaron en esta jornada, fueron tan numerosas, y de tan grande estimacion, no será fuera de proposito referir en esta Rela-



SEGUNDA PARTE
DE LA RELACION DIARIA
DEL ITINERARIO, QVE SV MAGESTAD (que Dios guarde) ha seguido desde q̃ salió de Madrid, hasta llegar a Fuenterabia: y forma de los Desposorios, que en aquella Ciudad se celebraron, entre la Serenissima señora D. Maria Teresa, Infanta de España, y el Christianissimo Luis Decimoquarto de Francia, y en su nombre, y en virtud de su poder, el Excelentissimo Don Luis Mendez de Haro y Guzman.
Año de 1660.

DESPVES que su Magestad salio de Madrid para la Ciudad de Fuenterrabia (que fue Jueves 15. de Abril proximo passado) llego aquella noche a la Villa de Alcalà de Henares, donde se venera el milagroso cuerpo de San Diego, Religioso de la Orden N. Serafico Padre S. Francisco, con quien tienen particular deuocion los señores Reyes de España. Fue grande el recebimiento, y fiestas que hizo aquella Villa a su Magestad, el poco tiempo que gozo de su Real presencia.

Viernes 16. llego a la Ciudad de Guadalaxara, cuyo Cabildo manifesto con varias muestras de regozijo, el afecto que, como tan leal vassallos, tienen a su legitimo Rey, y Señor.

Sabado 17. a la Villa de Hita.

Domingo 18. a Xidraque, Villa del Dique del Infantado, y aunque estas, y la antecedente son cortas de vezindad, signalaron en afectos, y demonstraciones festivas, a las Ciudades mas ricas, y populosas de este Itinerario.

Lunes 19. a Arienza, cuya Villa logro con luzimiento, entre otros regozijos, vnos fuegos artificiales.

B Mar-

Fig. 10. Inicial grabada en dos relaciones seriadas

Formato y extensión

La mayor parte de los textos transcritos –siete de las diez relaciones– presentan un formato en 4.º, muy frecuente en las hojas o pliegos sueltos. No sorprende, por tanto, que la extensión de estos impresos sea breve: una relación de dos hojas, tres de cuatro hojas, dos de doce hojas y solo una tiene una extensión superior al ocupar 26 hojas.

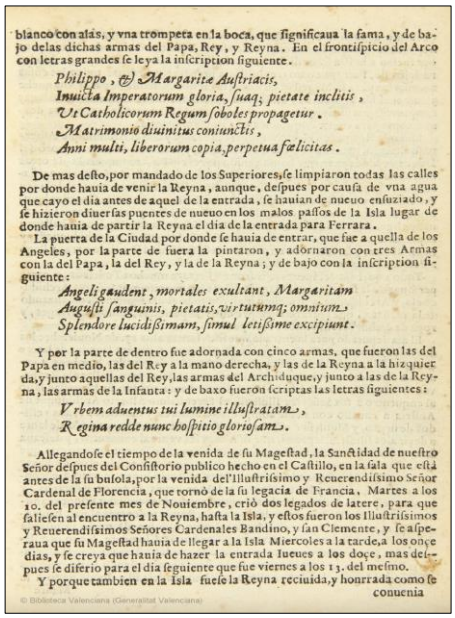
Las otras tres relaciones restantes, que ocupan dos hojas, están en formato folio (*Relación verdadera del […] bautismo de la serenísima princesa Margarita María Catalina* y la *Relación verdadera en que se da cuenta de todo el daño que causó las crecientes del rio Guadalquivir en la ciudad de Sevilla y Triana*)[55], y solo una en formato 8.º (*Relación del aparato que se hizo en la ciudad de Valencia para el recibimiento de la serenísima reina doña Margarita de Austria*)[56].

[55] Ejemplar conservado en la Biblioteca Nacional de España, sign. VE/59/64. Disponible en <http://bdh-rd.bne.es/viewer.vm?id=0000075249&page=1> (cons. 25/04/2022).

[56] Esta distribución podríamos entender que es también representativa de los formatos más empleados para este género editorial. Así, el *Catálogo y Biblioteca Digital de Relaciones de Sucesos (CBDRS)* devuelve resultados similares en cuanto a la preferencia de los formatos: en 4.º se registran 2 475 ediciones; en formato folio, 1 420 ediciones; en 8.º, 191 y en 12.º, 13 ediciones.

Disposición del texto

En cuanto a la disposición del texto, la mayor parte de las relaciones escogidas son en prosa y solo una es en verso: la *Relación del nacimiento del nuevo infante y de la muerte de la reina nuestra señora* (Cuenca, Salvador Viader, 1612)[57]. No obstante, en las relaciones, una u otra forma no eran excluyentes puesto que pueden aparecer combinadas. De este modo, en alguno de los volúmenes impresos –por ejemplo, en los libros de fiesta– se introducen en la narración los distintos versos que se recitaban en las justas o certámenes, así como las composiciones que adornaban los elementos efímeros.



blanco con alas, y vna trompeta en la boca, que significaua la fama, y de baxo delas dichas armas del Papa, Rey, y Reyna. En el frontispicio del Arco con letras grandes se leya la inscription siguiente.

Philippo, & Margaritæ Austriacis,
Inuicta Imperatorum gloria, suaq; pietate inclitis,
Vt Catholicorum Regum soboles propagetur.
Matrimonio diuinitus coniunctis,
Anni multi, liberorum copia, perpetua fœlicitas.

De mas desto, por mandado de los Superiores, se limpiaron todas las calles por donde hauia de venir la Reyna, aunque, despues por causa de vna agua que cayo el dia antes de aquel de la entrada, se hauian de nueuo enxuziado, y se hizieron diuersas puentes de nueuo en los malos passos de la Isla lugar de donde hauia de partir la Reyna el dia de la entrada para Ferrara.

La puerta de la Ciudad por donde se hauia de entrar, que sue a quella de los Angeles, por la parte de fuera la pintaron, y adornaron con tres Armas con la del Papa, la del Rey, y la de la Reyna; y de bajo con la inscription siguiente:

Angeli gaudent, mortales exultant, Margaritam
Augusti sanguinis, pietatis, virtutumq; omnium,
Splendore lucidissimam, simul lætissime excipiunt.

Y por la parte de dentro sue adornada con cinco armas, que fueron las del Papa en medio, las del Rey a la mano derecha, y las de la Reyna a la hizquierda, y junto aquellas del Rey, las armas del Archiduque, y junto a las de la Reyna, las armas de la Infanta: y de baxo fueron scriptas las letras siguientes:

Vrbem aduentus tui lumine illustratam,
Regina redde nunc hospitio gloriosam.

Allegandose el tiempo de la venida de su Magestad, la Sanctidad de nuestro Señor despues del Consistorio publico hecho en el Castillo, en la sala que está antes de la su bulola, por la venida del'Illustrissimo y Reuerendissimo Señor Cardenal de Florencia, que tornó de la su legacia de Francia, Martes a los 10. del presente mes de Nouiembre, crió dos legados de latere, para que saliesen al encuentro a la Reyna, hasta la Isla, y estos fueron los Illustrissimos y Reuerendissimos Señores Cardenales Bandino, y san Clemente, y se aspe-raua que su Magestad hauia de llegar a la Isla Miercoles a la tarde, a los onze dias, y se creya que hauia de hazer la entrada Iueues a los doze, mas despues se diferio para el dia siguiente que sue viernes a los 13. del mesmo.

Y porque tambien en la Isla sue sue la Reyna reciuida, y honrrada como se conuenia

Fig. 11. Ejemplo de RdS con alternancia de prosa y verso

Lo que sí evidenciamos en este ejemplo es que se emplean también distintas letras dentro de la misma relación (redonda y cursiva), así como lenguas diferentes. Por ejemplo, los fragmentos en latín de las ceremonias

[57] Ejemplar conservado en la Biblioteca Nacional de España, sign. R/12676. Disponible en: <http://bdh-rd.bne.es/viewer.vm?id=0000061653&page=1> (cons. 25/04/2022).

religiosas (Fig. 11).

Por último, y a diferencia de los libros de caballerías en gótica, las relaciones no suelen tener titulillos en los que se incluyan los nombres de los epígrafes del capítulo. Solo encontramos un caso, el de las *Fiestas de Aranjuez*, en donde en los rectos encabeza el titulillo «de Aranjuez» y en los versos aparece «Fiestas». Además, es el único ejemplo del texto que presenta apostillas marginales.

En definitiva, la heterogeneidad de un género como las relaciones de sucesos ofrece múltiples posibilidades de estudio bibliográfico, por lo que parece ser el contexto ideal para la explotación de herramientas de reconocimiento de textos.

§

Bibliografía citada

Allés Torrent, Susanna, «Tiempos hay de acometer y tiempos de retirar: literatura áurea y edición digital», *Studia Aurea*, 11 (2017), pp. 13-30.

Baranda, Nieves, «Compendio bibliográfico sobre narrativa caballeresca breve», en *Evolución narrativa e ideológica de la literatura caballeresca*, ed. M.ª Eugenia Lacarra, Bilbao, Servicio Editorial de la Universidad del País Vasco, 1991, pp. 183-191.

Bazzaco, Stefano, «El Progetto Mambrino y las tecnologías OCR: estado de la cuestión», *Historias Fingidas*, 6 (2018), pp. 257-272. DOI: <https://doi.org/10.13136/2284-2667/89> (cons. 13/05/2022).

—, «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. Estudios sobre el Siglo de Oro*, 9 (2020), pp. 534-561.

—, «Experimentos de estilometría en el ámbito de los libros de caballerías. El caso de atribución de un original italiano: *Il terzo libro di Palmerino d'Inghilterra* (Portonari, 1559)», *Actas de la Asociación Hispánica de Literatura Medieval*, 2022, en prensa.

Bognolo, Anna y Stefano Bazzaco, «Tra Spagna e Italia: per un'edizione digitale del Progetto Mambrino», *eHumanista/IVITRA*, 16 (2019), pp. 20-36.

Burnard, Lou; O'Brian O'Keeffe, Katherine; Unsworth, John (eds.), *Electronic Textual Editing*, New York, MLA, 2006.

Calvo-Tello, José, *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*, Transcript, Verlag, 2021.

Causer, Tim; Terras, Melissa, «"Many hands make light work. Many hands together make merry work": Transcribe Bentham and crowdsourcing manuscript collections», en *Crowdsourcing our Cultural Heritage*, Ashgate, Farnham, 2014, pp. 57-88.

Floridi, Luciano, *The 4th Revolution. How the Infosphere is Reshaping Human Reality*, Oxford, Oxford University Press, 2014.

Franzini, Greta; Kestemont, Mike; Rotari, Gabriela; Jander, Melina; Ochab, Jeremi K.; Franzini, Emily; Byszuk, Joanna; Rybicki, Jan, «Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm», *Frontiers in Digital Humanities*, 5 (2018), s.p.

García-Reidy, Alejandro, «Deconstructing the Authorship of *Siempre ayuda la verdad*: a play by Lope de Vega?», *Neophilologus*, 103 (2019), 493-510.

Gifford Fenton, Eileen; Duggan, Hoyt N., «Effective methods of producing machine-readable text from manuscript and print sources», en *Electronic Textual Editing*, eds. Lou Burnard, Katherine O'Brian O'Keeffe y John Unsworth, New York, MLA, 2006, pp. 241-261.

González-Sarasa Hernáez, Silvia, *Tipología editorial del impreso antiguo español*, Tesis doctoral, dir. Fermín de los Reyes Gómez, Universidad Complutense de Madrid, 2013.

Hernández-Lorenzo, Laura, «Poesía áurea, estilometría y fiabilidad: métodos supervisados de atribución de autoría atendiendo al tamaño de las muestras», *Caracteres: estudios culturales y críticos de la esfera digital*, n. 8/1 (2019), pp. 189-229.

Infantes, Víctor, «La narración caballeresca breve», en *Evolución narrativa e ideológica de la literatura caballeresca,* ed. María Eugenia Lacarra, Bilbao, Servicio Editorial Universidad del País Vasco, 1991, pp. 165-182.

—, «La prosa de ficción renacentista: entre los géneros literarios y el 'género editorial'», en *Actas del X Congreso de la Asociación Internacional de Hispanistas*, ed. Antonio Vilanova, Barcelona, PPU, 1992, vol. 1, pp. 467-474.

—, «¿Qué es una relación?: divagaciones varias sobre una sola divagación», in *Las «Relaciones de sucesos» en España (1500-1750). Actas del primer Coloquio Internacional (Alcalá de Henares, 8, 9 y 10 de junio de 1995)*, eds. María Cruz García de Enterría, Henry Ettinghausen, Víctor Infantes de Miguel y Agustín Redondo, Alcalá de Henares, Editorial Universidad de Alcalá - Publications de la Sorbonne, 1996, pp. 203-216.

—, «El *género editorial* de la narrativa caballeresca breve», *Voz y letra. Revista de literatura,* 7/2 (1996), pp. 127-132.

—, «La tipología de las formas editoriales», en *Historia de la edición y de la lectura en España 1472-1914*, dir. Víctor Infantes, François López y Jean-François Botrel, Madrid, Fundación Germán Sánchez Ruipérez, 2003, pp. 39-49.

Italia, Paola, *Editing 2000. Per una filologia dei testi digitali*, Roma, Salerno Editrice, 2020.

Kichuk, Diana, «Quantità e qualità dei testi online: il problema della digitalizzazione di massa», en *Teoria e forme del testo digitale*, ed. Michelangelo Zaccarello, Roma, Carocci Editore, 2019, pp. 135-166.

López Poza, Sagrario, «Las peculiaridades de las relaciones festivas en forma de libro», en *La fiesta. Actas del II Seminario de Relaciones de Sucesos (A Coruña, 1998)*, ed. Sagrario López Poza y Nieves Pena Sueiro, A Coruña, Sociedad de Cultura Valle Inclán, 1999, pp. 213-222.

Lucía Megías, José Manuel, *Imprenta y libros de caballerías*, Madrid, Ollero & Ramos, 2000.

Mancinelli, Tiziana, «Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work», *Historias Fingidas*, 4 (2016), pp. 255-260. DOI: <https://doi.org/10.13136/2284-2667/65> (cons. 13/05/2022).

Martín Molares, Mónica, «El bautismo de la princesa Margarita María Catalina de Austria (1623): tres ediciones de una relación», en *Buenas noticias. Relaciones de sucesos en los siglos XVI-XVIII: estudios y textos*, eds. Gabriel Andrés y Sandra M.ª Peñasco González, Pesaro, Metauro Edizioni, coll. Ispanica urbinate n. 4, 2021, pp. 65-89.

Mordenti, Raul, *Informatica e critica dei testi*, Roma, Bulzoni, 2001.

Moretti, Franco, *Graphs, Maps, Trees: Abstract Models for a Literary History*, Londres - Nueva York, Verso, 2005.

—, *Falso movimento. La svolta quantitativa nello studio della letteratura*, Milano, Nottetempo, 2022.

Mühlberger, Günter *et al.*, «Transforming scholarship in the archives through Handwritten Text Recognition. Transkribus as a case study», *Journal of Documentation - Emerald Publishing*, 75/5 (2019), pp. 954-976.

Narang, Sonika Rani; Jindal, M. K.; Kumar, Munish, «Ancient text recognition: a review», *Artificial Intelligence Review*, 53 (2020), pp. 5517-5558.

Orlandi, Tito (ed.), *Il problema della formalizzazione*, Roma, Accademia Nazionale dei Lincei, 1994.

Pena Sueiro, Nieves, «Estado de la cuestión sobre el estudio de las Relaciones de sucesos», *Pliegos de Bibliofilia*, 13/1 (2001), pp. 43-66.

Pena Sueiro, Nieves; Ruiz Astiz, Javier, «Las relaciones de sucesos: producto y género editorial en la Monarquía Hispánica», *Memoria y Civilización. Anuario de Historia*, 22 (2019), pp. 371-380.

Pierazzo, Elena, *Digital scholarly editing: Theories, models and methods*, Aldershot, Ashgate, 2015.

Reul, Christian; Springmann, Uwe; Wick, Christoph; Puppe, Frank, «State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open-Source Engines», *ArXiv e-prints*, 2018. URL: <https://arxiv.org/abs/1810.03436> (cons. 24/03/2022).

Rockwell, Geoffrey; Passarotti, Marco, «The Index Thomisticus as a Digital Humanities Big Data Project», *Umanistica Digitale*, 5 (2019), pp. 13-34. DOI: <http://doi.org/10.6092/issn.2532-8816/8575> (cons. 13/05/2022).

Roncaglia, Gino, «Google Book Search e le politiche di digitalizzazione libraria», *DigItalia web. Rivista del Digitale nei Beni Culturali*, 2 (2009), pp. 17-35.

—, *La quarta rivoluzione. Sei lezioni sul futuro del libro*, Roma-Bari, Laterza, 2010.

Rosselli del Turco, Roberto; di Pietro, Chiara; Martignago, Chiara, «Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo», *Umanistica Digitale*, 7 (2019), pp. 5-21. DOI: <http://doi.org/10.6092/issn.2532-8816/9322> (cons. 13/05/2022).

Shillingsburg, Peter L., *From Gutenberg to Google. Electronic Representations of Literary Texts*, Cambridge, Cambridge University Press, 2006.

Smith, David A., Ryan Cordell, *A Research Agenda for Historical and Multilingual Optical Character Recognition*, NULab, Northeastern University, 2018.

Terras, Melissa, «The Rise of Digitization: An Overview», en *Digital Libraries*, ed. Rico Rukowski, Olanda, Sense Publishers, 2010, pp. 3-20.