JANUS 9 (2020) 534-561

ISSN 2254-7290

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

Stefano Bazzaco
Università di Verona (Italia)
stefano.bazzaco.1@gmail.com

JANUS 9 (2020)
Fecha recepción: 14/10/20, Fecha de publicación: 29/10/20
<URL: https://www.janusdigital.es/articulo.htm?id=160>

Resumen
La presente investigación se centra en los principales aspectos de la digitalización masiva de textos y el reconocimiento automático de las imágenes digitalizadas por medio de software de OCR/HTR. Se presenta pues un experimento de reconocimiento HTR con libros de caballerías del siglo XVI y se propone un modelo apto para transcribir los textos de forma semi-automática y colaborativa.

Palabras clave
Reconocimiento automático de caracteres; OCR; HTR; libros de caballerías; letra gótica; Transkribus; Proyecto READ

Title
Automatic Text Recognition applied to Spanish Golden Age gothic script: creation of an HTR model based on 16th century Spanish Romances of Chivalry on the Transkribus platform

Abstract
The present investigation centres on the main aspects of massive digitalization of texts and the automated recognition of digitalized images thanks to OCR/HTR softwares. Finally, we present an experiment on HTR recognition dealing with XVI Century Spanish Romances of Chivalry and is delivered a model to transcribe in a semi-automated and collaborative way these texts.

Keywords
Automated Character Recognition; OCR; HTR; Spanish Romances of Chivalry; gothic script; Transkribus; READ Project

1. INTRODUCCIÓN[1]

En el panorama de las Humanidades Digitales, el debate acerca de qué sería una edición digital ha subrayado que la migración de una obra literaria al medio informático no puede ser la simple "grabación" de un documento en otro soporte, sino el resultado de un proceso hermenéutico, semántico y pragmático a la vez (Fiormonte y Martiradonna, 2010), llamado *remediation*[2]. Los contenidos de una edición científica digital, siguiendo esta interpretación, componen, por lo tanto, un objeto muy diferente con respecto a la fuente, complejo, como lo era el libro, pero de distinta naturaleza, porque no puede ser impreso sin que se verifique una pérdida de información (Sahle, 2016: 27)[3].

[1] Esta publicación se ha realizado en el marco del Proyecto I+D+i "Biblioteca Digital Siglo de Oro 6", con código: PID2019-105673GB-I00, financiado por el Ministerio de Ciencia e Innovación de España por el programa estatal de Generación del Conocimiento (2020-2022); del *Progetto Mambrino* (Università di Verona); del Progetto di Eccellenza "Le Digital Humanities applicate alle lingue e letterature straniere" 2018-2022 (Università di Verona – <https://dh.dlls.univr.it/>; del Proyecto PRIN "Mapping Chivalry. Spanish Romances of Chivalry from Renaissance to 21th Century: a Digital approach" (2020-2023), prot. 2017JA5XAR (Università di Verona).

[2] Sigue la misma idea Terras, cuando afirma "the act of digitization is one of translation: the resulting digitized representation of an original analogue object is not a replacement for the object. […] In order to record, copy, transmit, or analyse such a complex signal using computational methods, it is necessary to translate this into a form which is more simple, predictable, and processable" (2015: 5). Al respecto, ver también Kichuk (2019: 135).

[3] En la última década, la publicación de tantas ediciones científicas digitales ha llevado a definir el mismo concepto de "texto" a partir de su relación con el ecosistema digital. Los especialistas que se ocuparon del asunto, como Buzzetti y McGann (2006), Caton (2013), Robinson (2019: 117), a la luz de las complejidades que imponía la migración del documento escrito al medio digital, distinguieron dos distintas caras del objeto en cuestión: en la primera, el texto se presentaba como información codificada, secuencia lineal de bits con un valor potencial para el ordenador; en la otra, correspondía a la forma concreta de representación en un soporte de ese contenido informativo para el lector para que pueda comprenderlo de forma inteligible (Priani,

Desde otro punto de vista, sin embargo, una edición científica digital debe necesariamente guardar una relación con el objeto analógico que representa. Tal relación reside en la codificación de los "aspectos" más relevantes de la obra en papel[4], expresados dentro del ecosistema digital en lenguaje máquina, es decir una secuencia discreta de bits. Este proceso se define, en sentido extensivo, digitalización[5] y agrupa entre sí varias prácticas de captación de documentos reales, por lo que atiende a los estudios humanísticos, principalmente libros.

El objeto "libro", en su forma tradicional, presenta una problemática intrínseca para su representación digital, puesto que no distingue entre información codificada y expresión de esta. En otras palabras, podríamos decir que en el libro información y realización concreta se funden, hasta el punto de que la misma tinta que compone las palabras en el papel sería expresión no mediada y consustancial a su propio contenido informativo. La reutilización de un libro real en el ecosistema digital, por consiguiente, pasa por la selección de los aspectos que caracterizan al libro analógico y la migración de sólo algunos de ellos al contexto computacional, cada uno por separado, hasta crear un artefacto unitario pero dependiente de todas sus partes.

Banalmente, los dos principales rasgos del libro analógico que es indispensable captar a la hora de crear una edición científica digital son: el contenido textual (la información) y su forma concreta (el objeto libro). La

---

2015). La distinción entre "texto como codificación" y "texto como realización" llevó a importantes observaciones en el campo de la divulgación en ambiente digital de obras literarias, permitiendo imaginar el texto codificado como una suerte de *metatexto* independiente de sus manifestaciones concretas, pero a la vez capaz de resumirlas y representarlas en un todo. Tal concepción, basada en la falsa suposición de que un contenido textual o verbal pueda ser representado en cualquier medio, fue rebajada ulteriormente por la constatación de los límites intrínsecos de la codificación anidada al estilo SGML/XML, sobre todo cuando esta remite a objetos complejos como son los libros (Priani, 2015: 1227).

[4] Se considere al respecto la siguiente aclaración: "[…] although digital representations of historical artefacts can be seductive, they are not the historical artefact itself, as they are only a digital representation *limited to what has been captured during the sampling process*" (Terras, 2015: 66, cursivas mías).

[5] Aclaramos que, si bien aquí entendemos como digitalización todo proceso de conversión de documentos escritos al contexto digital, en el presente artículo distinguiremos entre digitalización y digitización. La primera corresponde, en nuestra interpretación, a la distribución de imágenes escaneadas de la fuente en papel. La segunda, al contrario, es la transformación de un texto analógico en un texto electrónico en formato *mrf* (*machine readable form*); una interpretación que preferimos por la evidente relación con el verbo anglosajón *to digit*, es decir teclear.

reproducción del contenido corresponde a la transposición del lenguaje escrito en un formato comprensible para la máquina, normalmente codificado en estándar ASCII o Unicode. La forma concreta de ese contenido en un entorno digital, al contrario, suele coincidir con la fotorreproducción de cada una de las partes que componen el objeto libro, es decir su conversión en una imagen virtual basada en una secuencia ordenada de píxeles.

Nótese que, evidentemente, no podríamos hablar de una edición científica digital refiriéndonos a uno de estos dos aspectos de modo independiente, porque caeríamos en el error de considerar solo una de las vertientes que califican el documento original. Cuando nos referimos a contenido textual e imágenes en ámbito digital, hablamos luego de dos caras distintas de la fuente en papel que están en la base de la transmisión y fruición del texto solo si se dan en conjunto[6].

A partir de estas premisas, en el ámbito digital se distinguen el texto digitizado, es decir codificado en formato electrónico, y el texto digitalizado en formato imagen. Estos dos elementos, que representan los dos ejes sobre los que se basa la creación de una edición científica digital, son sustanciales en el presente trabajo: en concreto, en estas páginas se valoran imagen digitalizada y texto digitizado como objetos transmisores de conocimiento y se individualizan las técnicas aptas para transferir de forma automática el contenido de una (la imagen de la página del libro) a otro (la transcripción del texto relacionado)[7].

La investigación que proponemos parte, pues, de la toma en consideración de los aspectos principales de la digitalización masiva de documentos en relación con la extracción de contenidos textuales. Sucesivamente, se detallan los aspectos destacados en el campo del reconocimiento automático y semi-automático de textos, explorando las posibilidades ofrecidas por las herramientas OCR (*Optical Character Recognition*) y HTR (*Handwritten Text Recognition*). Se señalan después los problemas principales que afectan al reconocimiento automático de textos

[6] Al respecto, es interesante que se haya intentado ensalzar la edición científica digital frente a la edición crítica tradicional por ser idealmente más respetuosa con las fuentes, que pueden reproducirse todas y compararse simultáneamente en la pantalla, y más atenta a los destinatarios, pues no necesita representar en un único espacio físico (la página, el libro) los distintos aparatos críticos, ecdóticos, explicativos.

[7] Las imágenes en *Apéndice* son un ejemplo de cómo un libro pasa al contexto digital bajo la forma de imágenes digitalizadas. Es evidente, al respecto, que se trata de escaneos de las fuentes y no de un texto electrónico, que en ese caso sería procesable y buscable mediante la máquina.

antiguos, poniendo la atención en las dificultades específicas encontradas con el reconocimiento de la letra gótica española impresa de los siglos XVI y XVII. Finalmente, se propondrá una herramienta experimental (modelo de HTR) para la transcripción semi-automática de textos en gótica del siglo XVI que creamos en la Universidad de Verona por medio de la plataforma Transkribus <https://transkribus.eu/Transkribus/>.

Este trabajo es fruto de las experimentaciones llevadas a cabo en los últimos dos años en el marco del *Progetto Mambrino*, grupo de investigación que desde 2003, bajo la dirección de Anna Bognolo y Stefano Neri, se ocupa del estudio de las novelas caballerescas del Renacimiento español. Por consiguiente, el corpus de trabajo que se eligió contiene algunos libros de caballerías publicados en la península entre 1520 y 1580[8]: estos textos venían muy al caso por ser bastante homogéneos, extensos y cualitativamente bien digitalizados por parte de bibliotecas y archivos (unas características que, como veremos, están en la base de una exitosa aplicación de s*oftware* de reconocimiento automático de textos). Sin embargo, en estas páginas la aplicación del modelo de reconocimiento no se limita a los textos caballerescos, sino que el módulo de HTR que aquí se propone puede ser enriquecido y el corpus acrecentado por parte de otros especialistas de forma progresiva, con la idea de un proyecto abierto cuyos objetivos últimos son la colaboración y la posibilidad de compartir conocimiento.

2. LA DIGITALIZACIÓN MASIVA Y EL RECONOCIMIENTO AUTOMÁTICO DE TEXTOS

Es evidente que la digitalización masiva de documentos de las últimas cuatro décadas, en concomitancia con el *digital turn* global, ha representado para los estudiosos de literatura un cambio en la relación con su propio objeto de investigación (Terras, 2015). No habría podido ser de otra manera. Sostenida por el imperativo categórico de la innovación y del desarrollo tecnológico, la reproducción del patrimonio global de libros en el contexto digital ha producido en menos de 20 años copias digitales de más de cien millones de libros, que ahora pueden ser consultados cómodamente desde cualquier sitio a través de un ordenador, una tableta o un móvil de última generación.

[8] Ver el cuarto apartado y el *Apéndice*.

Se trata de un cambio notable porque implica que las metodologías tradicionales de la filología deben implícitamente aclimatarse a los logros de las nuevas tecnologías y adaptarse al nuevo medio: los avances en campo técnico aseguran abreviar los tiempos de una investigación tradicional, los costes de viajes y alojamiento para acceder a bibliotecas y archivos, los momentos de estudio en situaciones incómodas con condiciones de luz variables y restricciones en la consulta de los documentos. Sin contar que el acceso en remoto y la posibilidad de compartir archivos digitalizados pueden potenciar el estudio de las fuentes, por medio de herramientas informáticas que aumentan la legibilidad de las letras desgastadas y consumidas, a la vez que reducen el manejo de los documentos, lo cual favorece su conservación (Olgivie, 2017: 82).

Terras individua tres distintas etapas de este proceso de digitalización de fuentes primarias (2010). La época de los 80 es la en que los proyectos de digitalización mueven sus primeros pasos: son los años en que se difunde de forma generalizada la recolección en imágenes digitales de volúmenes y documentos conservados en archivos y bibliotecas. Esta primera etapa, lógicamente, corresponde a un periodo de experimentación, en que las instituciones ensayan con las nuevas tecnologías y emprenden proyectos de pequeño alcance, circunscritos a unos ámbitos de interés limitados. Uno de los proyectos más ambiciosos de esta etapa fue el que inició en 1986 el Archivo General de Indias de Sevilla, pionero porque por primera vez interesaba indistintamente a instituciones públicas y privadas (el Ministerio de Cultura Español, IBM España y la Fundación Ramón Araces) la creación de un archivo de imágenes digitales que alcanzó los 7 millones de páginas escaneadas en seis años (Terras, 2010: 4).

En los años 90, con la introducción de la red global de información, los esfuerzos de digitalización de documentos se intensificaron y llevaron al desarrollo de proyectos de largo alcance. Tal etapa, que por su impacto se definió *age of digitalization*, fue estimulada por distintas fuerzas —el cambio de las políticas públicas de digitalización, los avances en el campo de las tecnologías web, la inversión de fondos institucionales y estatales—y correspondió a un periodo de florecimiento de la digitalización, que se hizo masiva. Junto con la difusión de buenas prácticas y de estándares para la conservación, se produjo en este periodo un avance significativo en las infraestructuras de elaboración y gestión de los materiales digitalizados, hasta la creación de entes específicos, como la *Digital Library Federation* (1995), un

consorcio nacido con el fin de dirigir y sostener grandes proyectos de digitalización entre instituciones públicas y privadas americanas.

Alrededor del inicio del nuevo milenio empieza, por fin, la última etapa destacada por Terras, la en que toman forma algunos de los principales planes de digitalización que subsisten hoy en día. En esta época la reproducción y la difusión de fuentes primarias se convierte en un lugar común, tanto que asistimos a la proliferación de tecnologías avanzadas para el tratamiento de las imágenes escaneadas, con un consiguiente crecimiento de iniciativas financiadas por los gobiernos nacionales y de proyectos supranacionales de larguísimo alcance. En esos años se acercaron a la digitalización del libro las empresas comerciales (como *Google*), que conectaron la idea de difusión de una cultura universal con los intereses para controlar el patrimonio textual global.

A partir de este momento, se empezó, pues, a comprender que, a pesar de las enormes ventajas que prometía, el proceso de digitalización implicaba también algunos riesgos, porque la transmisión del conocimiento estaba sometida al imperativo de dominar la información contenida en los libros. La urgencia de controlar el *big data* almacenado en los documentos escritos, tradicionalmente reconocido como un contenido informativo validado y adecuado, favoreció la expansión de una cultura divulgativa de escasa calidad donde los directores de grandes compañías comerciales, tras la quimera del progreso a toda costa, expulsaron en parte a los humanistas del cambio cultural que se estaba verificando.

Entre las empresas comerciales que apostaron por la digitalización con el fin de controlar el acceso y el manejo de los libros en ambiente digital, el caso más relevante es seguramente el de *Google*. La industria de Montain View fue muy rápida en captar el cambio que iba a verificarse en el ámbito del acceso y la consultación de los libros; por consiguiente, haciendo hincapié en la experiencia ya demostrada con el control de los contenidos web, lanzó en 2004 la plataforma *Google Books Search* (que conocemos ahora como *Googlebooks*). Sin embargo, no contó con el hecho de que el patrimonio de los libros representaba una realidad compleja: su transmisión debía necesariamente confrontarse con siglos de tradición impresa, limitando la idea de que los libros eran sencillamente vehículos de información[9]. Dejando al lado cuestiones como la política

[9] Lucía Megías con respecto a este asunto juzga la cuidadosa selección y el control de los documentos (aspecto cualitativo) superiores a la acumulación de los mismos (aspecto cuantitativo). Solo reconociendo e interpretando su patrimonio textual, las bibliotecas virtuales pueden convertirse en motores de la cultura y del conocimiento (2012: 89-102).

descarada de Google para la obtención de digitalizaciones y su tratamiento confuso de los derechos intelectuales[10], lo que me interesa aquí, de acuerdo con las impresiones de Nunberg (2009), es subrayar que Googlebooks no supo convertirse en un punto de referencia fiable para los estudiosos; como mucho podría considerarse un almacén de informaciones útiles, capaz de alimentar una cultura del *good enough*[11] bajo la capa de una democratización del conocimiento.

Desde este prisma de observación, son dos las deficiencias que normalmente se achacan al motor de Google: los metadatos lagunosos y los resultados de búsqueda inciertos. Las razones son múltiples e implican distintas prácticas; lo que es cierto es que en este ámbito desempeñan un papel fundamental el general descuido hacia la extracción de metadatos y la falta de una revisión sistemática[12].

Estos tipos de errores dependen en gran medida del uso de herramientas de reconocimiento automático de textos, empleadas por parte de *Google* con el objetivo último de transformar el contenido gráfico de las imágenes en un flujo textual que el ordenador pudiera fácilmente indexar y hacer buscable por medio de algoritmos y *parsers*. En concreto, lo que hizo *Google* fue implementar un sistema de OCR de acceso abierto llamado *Tesseract* <https://opensource.google/projects/tesseract> apto para la conversión automática de las imágenes escaneadas en un texto en *mrf*. La operación no era nada desdeñable (Roncaglia, 2009: 203-204), porque permitía a la empresa controlar de forma inmediata los contenidos de su biblioteca digital: la transcripción automática por un lado aseguraba una extracción rápida de los metadatos; por otro prometía una interrogación interna de los materiales escaneados, procesables ahora

[10] Se ocupan de estas cuestiones: Roncaglia (2009: 183-203); Borghi, Karapapa (2019: 97-102)

[11] Con este término nos referimos a una cultura en la que "lo que es disponible y fácilmente accesible en red prevalece sobre lo que es cualitativamente superior pero más difícil de encontrarse, como un libro en la biblioteca" (Mancinelli y Pierazzo, 2020: 15).

[12] Al respecto, Nunberg (2009) y Roncaglia (2009: 202-203) consideran que la política poco escrupulosa de *Googlebooks* se deriva del carácter omnívoro de la empresa. Esta idea general, sin embargo, se enlaza con unos problemas de fondo específicos como: la tendencia a falsear la relación con las fuentes primarias, produciendo una suerte de "desmembramiento virtual" de los archivos (Olgivie, 2017: 83); la digitalización inicial de las bibliotecas de grandes universidades americanas, que recogen ediciones en larga medida pertenecientes al siglo XIX, verdaderas joyas para los estudiosos de aquella época, pero inútiles para investigadores de otras épocas (Italia, 2020: 28-29); la desintelectualización de las obras literarias (Borghi y Karapapa, 2019: 111).

directamente en formato imagen gracias a la aplicación de un estrato textual oculto que permitiera organizar la navegación de los contenidos.

Quizás fueron dos las cuestiones que obstaculizaron esta ilusión y causaron lo que Nunberg llamó, refiriéndose a *Googlebooks*, un desastre interpretativo (2009). En primer lugar, la entera operación era bastante compleja porque requería distintas competencias: sólo gracias a una interacción constante entre humanistas e informáticos habría sido posible seleccionar el material, clasificarlo de forma certera y asegurar una conversión correcta de los textos en artefacto digital. En segundo lugar, hay que admitir que las tecnologías de transcripción automática tuvieron un papel fundamental en la multiplicación de errores, puesto que en ese momento no habían alcanzado un nivel suficiente de precisión para permitir una transcripción de las fuentes que no necesitara la revisión por parte de un control humano. Sobre todo, el problema se manifestaba de forma aún más contundente con los textos antiguos, por ejemplo los impresos del XVI, que presentaban unos caracteres difíciles de interpretar por parte de la máquina, variables en su representación y repletos de elementos gráficos que complicaban la inteligibilidad del texto (Bazzaco, 2018: 260-261).

La faceta más alarmante de todo este proceso es que no solo *Google* emplea OCR inadecuados para transcribir el contenido de los libros digitalizados, sino que, a partir de experiencias parecidas, se ha generado un verdadero círculo vicioso, que imbrica instituciones comerciales y públicas a la vez. Archivos digitales como *Internet Archive* y *Hathitrust*, que deberían en sus estatutos promover la difusión de materiales totalmente fiables, proponen a los usuarios versiones de textos derivados del OCR casi ilegibles (Kichuk, 2019: 145-156). Otros proyectos de digitalización, siguiendo el modelo de *Googlebooks*, se basan en texto "OCRizado" para la obtención automática de metadatos y la recuperación de materiales textuales presentes en la red, con una proliferación de inexactitudes de carácter variable (Kichuk, 2019: 157). Existen también repositorios como *Project Gutenberg* que se dirigen decididamente al lector no especialista[13], con el resultado de presentar obras de mediana calidad pero inservibles para el estudioso de disciplinas humanísticas: una política editorial preocupante porque se extiende también a los archivos digitales *partners* del proyecto, que recogen los libros electrónicos de *Project Gutenberg* y contribuyen así a la propagación de las imprecisiones y a la fijación de metodologías erróneas de difusión de los textos (Kichuk, 2019: 145-146).

[13] Lucía Megías define los projectos de este tipo "bibliotecas virtuales generalistas" (2012: 92).

Este proceso de conversión analógica-digital, que tiene sus consecuencias también en el sistema de confección de *e-books* comerciales y en otros sectores editoriales, es finalmente motivo de angustia para el filólogo porque sigue promoviendo unos efectos irreversibles. Dentro de unos años, en un futuro teóricamente próximo, todos los libros que no están bajo derechos comerciales se habrán ya digitalizado por primera vez de forma no supervisada. El coste de digitalización, sin embargo, es de por sí un elemento de disuasión para que se inviertan fondos en segundas o terceras digitalizaciones, puesto que los archivos afrontarán nuevas problemáticas, como la imponente gestión de los repositorios derivados (Kichuk, 2019: 160). Por consiguiente, el impacto negativo generado por una digitalización falaz y la aplicación de OCR inadecuados para la extracción de metadatos y la transcripción descontrolada de los textos no podrá ser eliminado totalmente. Rellenar de modo prematuro los archivos digitales con materiales defectuosos está llevando a una catástrofe bibliográfica. Está claro que este riesgo puede evitarse sólo partiendo de una labor de corrección y regeneración constante de lo que ya se encuentra en la red.

3. Los sistemas de reconocimiento automático de textos impresos: OCR y HTR

La resolución de los problemas existentes en el campo de la transmisión del patrimonio textual al medio digital que acabamos de exponer pasa por la introducción de buenas prácticas que aseguran una correcta difusión de los libros en la red. Al respecto, son varias las soluciones que han sido propuestas, como la creación de ediciones digitales certificadas por la comunidad científica, la difusión de archivos y metadatos totalmente fiables, la corrección en *crowdsourcing* de textos digitizados, etc. Para aligerar estos procedimientos, el punto de partida puede ser el de establecer un uso adecuado de los sistemas de reconocimiento automático de textos, un campo en que la contribución metodológica y científica de los filólogos puede ser determinante (Mancinelli y Pierazzo, 2020: 16).

Evidentemente, las herramientas de reconocimiento automático de textos, con frecuencia reunidas bajo el término OCR, son en la actualidad muy distantes del campo de investigación filológico. La razón de este alejamiento consiste en una inicial constatación por parte de los humanistas de la insuficiencia de los primeros sistemas de OCR, que por distintas razones –el escaneo defectuoso de las fuentes, los problemas gráficos particulares de las

obras impresas– no prometían resultados fiables y útiles para el avance de la disciplina. Se produjo, por lo tanto, una suerte de prejuicio (*bias*) hacia los instrumentos de transcripción automática, enfatizado por la distinción, comúnmente muy difundida en los ambientes filológicos, entre "*clean transcription*" y "*dirty OCR*" (Smith y Cordell, 2018: 10-11).

Sin embargo, y afortunadamente, la tecnología ha seguido adelante, prometiendo resultados cada vez más fiables en el reconocimiento automático de textos. Actualmente, la transcripción automática de fuentes impresas contemporáneas se considera un problema solucionado[14], puesto que se aseguran unos resultados de transcripción muy cercanos a la perfección (Sprigmann y Lüdeling, 2017). Los defectos en el reconocimiento de textos impresos antiguos, al contrario, siguen representando un obstáculo para la difusión en la red de contenidos textuales fiables.

Desde un punto de vista conservador, los mejores sistemas de foto-reproduccion no deberían alterar el estado de las páginas, porque el documento debe reproducirse de forma exacta manteniendo intacta su materialidad; por otro lado, en esta misma exactitud de representación del objeto real se complica la reproducción del contenido en formato digitizado, puesto que la transcripción automática está obstaculizada por elementos gráficos de difícil interpretación, principalmente con respecto a los libros antiguos.

Los principales componentes no textuales de la imagen que limitan la aplicación fructífera de sistemas de reconocimiento automático de textos son las manchas en las páginas, las deformaciones procedentes del escaneo manual y las transferencias de la tinta al vuelto de la página (Sprigmann y Lüdeling, 2017). Si bien contamos en la actualidad con óptimos proyectos de digitalización lanzados por bibliotecas y archivos, como es el caso del portal *Biblioteca Digital Hispánica* de la BNE, donde el escaneo de las fuentes desde el punto de vista técnico está sometido a una adecuada declaración de calidad[15], el problema de la aplicación errónea de sistemas de transcripción automática puede presentarse con frecuencia y la única solución para gestionarlo reside en el empleo de instrumentos para la corrección automatizada de imágenes y su elaboración posterior para hacerlas aptas a la lectura por parte del OCR.

[14] Participan de este conjunto todas las grafías impresas del siglo XX, con excepción del Fraktur, una grafía que se difundió en las primeras tres décadas de 1900 en área germanófona (cfr. Reul *et al.*, 2018).

[15] Ver al respecto la siguiente documentación: <http://www.bne.es/webdocs/ Catalogos/ ProcesoDigitalizacionBNE.pdf [08/10/2020].

Con respecto al segundo asunto, el reconocimiento automático de letras impresas antiguas representa un problema aún más evidente. En los impresos antiguos aparecen varios componentes gráficos que dificultan la interpretación automática por parte del ordenador. Con respecto a los textos que nos interesan aquí, es decir los que se publicaron durante el Siglo de Oro español, contamos con varios de ellos: variabilidad gráfica de las letras, abreviaturas, signos tironianos, ligaduras entre caracteres, a los que se añaden elementos accidentales como el frecuente desgaste de los tipos y los defectos del entintado. Todos estos elementos complican el reconocimiento con programas de OCR porque se oponen a la norma general de estas tecnologías, que impone la correspondencia directa y constante entre un signo gráfico en la fuente y una letra en el texto transcrito[16]. En la imprenta manual antigua es natural que las realizaciones concretas de cada tipo impreso sean oscilantes dentro de una única fuente y hasta pueden coincidir con más de una letra transcrita (por ejemplo en el caso de las abreviaturas), los resultados que se obtienen no pueden alcanzar un nivel de precisión satisfactorio.

Junto con el desarrollo de los softwares de OCR, han surgido más recientemente plataformas de HTR (*Handwritten Text Recognition*), principalmente diseñadas para la interpretación de textos manuscritos, capaces de asegurar buenos resultados también con los impresos antiguos. El origen de estas herramientas está íntimamente relacionado con el desarrollo de los sistemas de OCR; pronto, sin embargo, el HTR se convirtió en un área de estudio independiente por las problemáticas específicas que debía solucionar, la más contundente de las cuales era la variabilidad en las grafías de los textos manuscritos, un asunto complejo y que requería un considerable desarrollo tecnológico. Hace dos décadas se dieron avances determinantes en la estadística computacional gracias a la integración de la inteligencia artificial a los sistemas de *pattern recognition* y al creciente empleo de algoritmos de *machine learning* capaces de aprender cómo acrecentar sus *performances* a partir de unos ejemplos dados. Todo ello, sumado al incremento de las capacidades de procesamiento de los ordenadores, permitió que los *softwares* de HTR se

[16] Sugiere lo mismo Orlandi cuando afirma: "il procedimento di OCR avveniva in modo relativamente rigido. Il pacchetto funzionava esclusivamente in base al riconoscimento dei caratteri di un certo numero di polizze (font) ritenute di maggior uso; ovvero poteva essere istruito da un operatore in modo da riconoscere una polizza non prevista. […] Ove occorressero imperfezioni nella stampa dell'originale, ovvero macchie o altro sulla carta, il sistema non era in grado di ricostruire la lettera in questione" (2010: 30-31).

convirtieran en productos cumbre dentro del campo del reconocimiento automático de textos.

En el marco del *Progetto Mambrino* se ha experimentado una herramienta de la plataforma *Transkribus* que permite transcribir con precisión los impresos antiguos: un *software* de HTR *user friendly* que facilita el proceso de entrenamiento del ordenador en el reconocimiento de letras impresas del Siglo de Oro.

Transkribus es un instrumento que nació dentro del Proyecto READ *(Recognition and Enrichment of Archival Documents*) financiado por el Programa Europeo Horizonte2020, en principio sostenido por colaboradores que provenían de trece naciones europeas distintas (comprendida España, con la Universitat Politècnica de València). El *software* cuenta ahora con un impresionante número de usuarios, que va multiplicándose constantemente, y se ha convertido en un punto de referencia para muchos estudiosos, como demuestran los numerosos participantes en el último encuentro anual del grupo en Innsbruck <https://readcoop.eu/transkribus-user-conference-2020> y la reciente fundación de la Cooperativa Europea READ COOP SCE, un proyecto que agrupa más de sesenta instituciones de todo el globo <https://readcoop.eu/members>[17].

El proyecto de *Transkribus* se basa en una doble articulación: por un lado, las instituciones culturales, los académicos y, más en general, los usuarios proveen los materiales digitalizados y unas transcripciones fiables de los mismos para el entrenamiento del s*oftware*; por otro, los informáticos confeccionan las infraestructuras necesarias y elaboran los datos. El resultado de este proceso es la consolidación de una red de usuarios (*growing users' network*), central en el desarrollo de la plataforma porque el sistema de *machine learning* empleado se alimenta y fortalece por medio de los nuevos materiales procesados, los cuales, a pesar de permanecer como objetos privados de los usuarios, son fundamentales para las redes neurales, que en el fondo aprenden a través de ellos y se potencian (Carbonell *et al.*, 2013).

Los últimos avances del *software* (que ofrece una cómoda interfaz gráfica *java-based*) residen en la inclusión de *deep neural networks* que permiten la creación de modelos de reconocimiento automático que no dependen de la lengua y la fecha de producción del documento –lo cual coincide

[17] Entre las instituciones que adhirieron a la cooperativa, en su mayoría austriacas y alemanas, figuran también la Universitat Politècnica de València (España) y la Fondazione Banco di Napoli (Italia, cfr. Zappulli e Iorio, 2018).

potencialmente con la posibilidad de reconocer cualquier grafía de cualquier época (Leifert *et al.*, 2016). A la luz de los muchos logros alcanzados por las instituciones *partner* y viendo los excelentes resultados obtenidos en el ámbito del *Progetto Mambrino* en las transcripciones de la letra impresa cursiva del siglo XVI (Mancinelli, 2017; Bazzaco, 2018), nos hemos propuesto experimentar un modelo de HTR extendido para la transcripción semi-automática de la letra gótica española del siglo XVI.

4. La creación de un modelo HTR para la gótica impresa por medio de la plataforma Transkribus

El proyecto de crear un modelo HTR para transcribir de forma semi-automática la gótica impresa pasó por tres distintas fases: (1) en principio, se valoraron los límites de la plataforma *Transkribus* y las posibilidades de alcanzar nuestros objetivos; (2) después se seleccionaron las obras digitalizadas que iban a constituir nuestro corpus de trabajo; (3) finalmente, preparamos las transcripciones manuales necesarias para entrenar el programa y generamos, a partir de esos materiales, un modelo de reconocimiento.

(1) Con respecto a la evaluación de *Transkribus,* la experiencia previa con otros textos renacentistas fue determinante. *Transkribus*, de hecho, es una plataforma que asegura buenos resultados con los materiales impresos, principalmente por su uniformidad en comparación con los documentos manuscritos; no obstante, era evidentemente imposible obtener transcripciones totalmente fiables. En nuestro caso, se habría podido, quizás, alcanzar un nivel de precisión muy cercano al 99% de texto correcto, como ya se hizo en el caso de la letra cursiva, pero nunca se habría obtenido una transcripción exenta de errores (Bazzaco, 2018). Al respecto, téngase en cuenta que una transcripción con el 1% de error en nuestro caso no sería considerada un producto digno de ser destinado a la lectura, porque cada página en formato folio contiene de media 450 palabras y 1.500 caracteres, lo cual correspondería con tener en cada hoja alrededor de 10-15 errores. Por supuesto, muchos de estos errores no serían significativos (faltas de acentuación, ausencia de guiones a finales de líneas), pero de todos modos no sería una transcripción fiable si prescindiera de un control humano de los resultados[18].

[18] Al respecto, hay estudios que aseguran que tales resultados se pueden corregir con menor esfuerzo que transcribiendo manualmente el texto *from scratch*; al contrario, cuando los

Otra cuestión que tuvimos en cuenta es que *Transkribus* funcionaba mucho mejor si se entrenaba sobre un único texto, lo cual, sin embargo, no habría producido un modelo extensible. Para nuestro objetivo, habría resultado más útil facilitarles a los estudiosos un modelo experimental de HTR que reuniera diferentes textos parecidos (*extended model*) que pudiera ser empleado como módulo base para la transcripción de nuevos textos que progresivamente, una vez corregidos, pudieran sumarse alimentando el modelo. De tal manera, nos propusimos crear un modelo HTR que, fortalecido por las adiciones de los usuarios, permitiría en el futuro transcribir todo tipo de texto en letra gótica con un margen de error muy bajo, disminuyendo el aporte de la indispensable ulterior supervisión humana.

(2) Partiendo de estas premisas, pasamos a seleccionar las obras que iban a constituir la base del modelo HTR. Estos textos debían de ser bastante homogéneos en su conformación tipográfica; ser extensos, porque solo así aumenta la probabilidad de encontrar elementos de difícil interpretación por parte del *software*; y, además, debían cumplir con unos elevados estándares de calidad en su digitalización en formato imagen[19].

A partir de estos criterios de selección, con el fin de componer nuestro corpus de trabajo, entre los textos que conocíamos y que teníamos fácilmente a disposición, elegimos los siguientes ejemplares reproducidos:

título	**autor**	**impresor(es)**	**localización**
Lisuarte de Grecia	Juan Díaz	Jacobo y Juan Cromberger, Sevilla, 1526	BNE R/71
Florando de Inglaterra	Anónimo	German Gallarde, Lisboa, 1545	BL C.62.h.14.
Silves de la Selva	Pedro de Luján	Dominico de Robertis, Sevilla, 1549	BNE R/865
Leandro el Bel	Pedro de Luján	Miguel Ferrer, Toledo, 1563	BNE R/9030

Tabla 1. Ejemplares seleccionados para la creación del modelo HTR

Como se ve en la tabla, los textos escogidos para la creación del modelo de HTR proceden de diferentes lugares e impresores y abarcan un periodo de

resultados superan el 10% de error las transcripciones son de por sí inservibles (Alvermann y Blüggel, 2017).

[19] Roling (2020) sugiere que es lógico emplear materiales íntegros y bien preservados para aviar la creación de un modelo de reconocimiento textual.

producción bastante extenso, que va de 1526 a 1563. Las imágenes digitalizadas de las fuentes provienen en tres casos del portal Biblioteca Digital Hispánica (reproducciones que alcanzan una resolución superior a los 300 ppp); las imágenes del *Florando de Inglaterra* (1545) proceden del proyecto de digitalización emprendido por la *British Library* en colaboración con *Google*[20], de manera que desconocemos el grado de resolución de las imágenes, que, a pesar de ello, parecen adecuadas para nuestros propósitos.

Todos los textos están impresos en letra gótica (unas páginas están reproducidas como se muestra en *Apéndice*)[21].

(3) Una vez que decidimos qué textos iban a ser la base de nuestro modelo de HTR, procedimos a trabajar directamente dentro de *Transkribus*. El *software*, como explicamos en otras ocasiones (Bazzaco, 2018; Bognolo y Bazzaco 2019), funciona de la siguiente manera: inicialmente se suben las imágenes digitalizadas de las páginas a la plataforma; después se segmentan las imágenes en regiones y líneas de texto (*Layout Analysis*); se pasa pues a la transcripción manual de unas páginas de la fuente (el texto transcrito se llama *Ground Truth* o *Golden Transcription*); finalmente, gracias al entrenamiento del *software* y a la creación de un modelo de HTR, se pueden transcribir de forma automática las restantes páginas.

Como se puede observar (Fig. 1), Transkribus es una herramienta que se apoya a un servicio *client-server.* Los instrumentos que se encuentran en la nube son dispositivos de acceso abierto; por otra parte, la adquisición de las transcripciones derivadas del reconocimiento automático es un servicio con limitaciones[22]. Restringimos por lo tanto nuestra experimentación a la sección *open access* del *software*.

[20] Recogimos la información del dispositivo de visualización de fuentes digitalizadas promovido por la British Library: <http://access.bl.uk/item/viewer/ark:/81055/_vdc_100035857199.0x000001#?c=0&m=0&s=0&cv=0&xywh=2055%2C1%2C6407%2C3490>.

[21] No es de momento posible indexar con seguridad los tipos que constituyen la base de nuestro modelo de reconocimiento. Basándonos en las reproducciones en facsímil contenidas en los catálogos, podemos suponer que el modelo de HTR en este estadio inicial podrá reconocer con bastante precisión los tipos catalogados por Haebler con los números 26, 53, 54, 67, 68, 72, 73, 77, 83, 86, 87, 91, 92, 99, 100, 109, 125, 126, 163, 167 (1902: XIV y ss.); los ejemplos de letras inventariados por Norton como II, IV, V (1966: 186-192); y los tipos góticos registrados por Griffin como 1, 2, 3, 7, 8 (1991: 271-282). Se ha consultado también el inventario en línea *Typenrepertorium der Wiegendrucke* patrocinado por la Biblioteca Estatal de Berlín <https://tw.staatsbibliothek-berlin.de>).

[22] La distribución de los textos transcritos se convertirá pronto en un servicio de pago; sin embargo, los estudiosos podrán descargar 500 páginas transcritas de forma gratuita y se

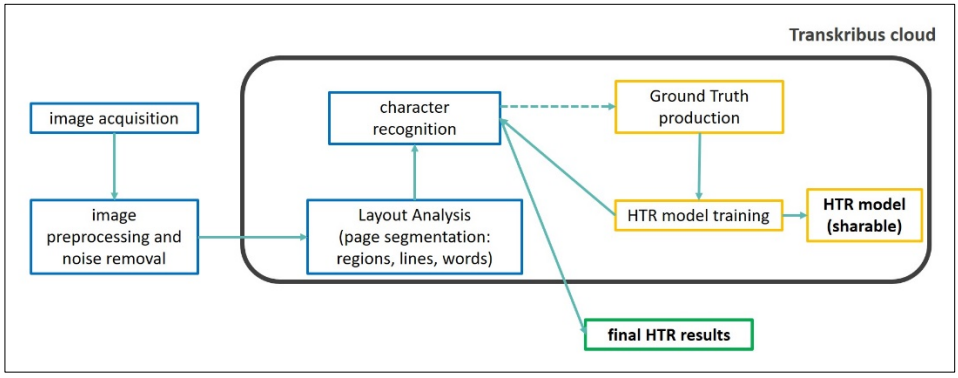


Fig. 1. Síntesis del flujo de trabajo de la plataforma Transkribus

En principio, procesamos las imágenes digitalizadas antes de subirlas a la plataforma con el fin de perfeccionar los resultados del reconocimiento. Resolvimos de forma automática los problemas de legibilidad del texto y la presencia de elementos gráficos que estorbaban el entrenamiento del software (generalmente llamados ruido) por medio de Scantailor <https://scantailor.org/>, una herramienta de acceso abierto que consiente la modificación de imágenes por lote.

Pasamos pues a la segmentación de las imágenes digitalizadas. El texto de los libros de caballerías que empleamos está normalmente dispuesto en doble columna, como habitualmente se dispone este género en la imprenta: tal distribución dificulta el correcto orden de lectura de cada línea de texto por parte de la máquina. Para resolver el problema, instruimos el programa gracias a la función interna a Transkribus llamada Page to Page Layot Analysis (P2PaLA) y creamos un módulo capaz de segmentar de forma automática las dos columnas de texto y asignarle a cada línea un correcto orden de lectura. El modelo de segmentación automática de la página que creamos se basa en 200 páginas y ha demostrado ser totalmente fiable con las obras de nuestro interés. El modelo de P2PaLA se pondrá a disposición de los usuarios y será incluido en la carpeta *public models* de Transkribus con el nombre "SpanishGothic_2columns"[23].

Para la transcripción de las fuentes se mantuvieron criterios estrictamente conservadores, puesto que los proveedores del s*oftware* aconsejan

prevén, en futuro, diversos planes económicos para las universidades y otras instituciones culturales.

[23] Para ulteriores informaciones acerca de la función P2PaLA ver el siguiente enlace: <https://transkribus.eu/wiki/index.php/P2PaLA> [07/10/2020].

respetar en lo posible la correspondencia entre un signo gráfico de la fuente y un carácter de la transcripción. Sin embargo, la intención es la de crear en un futuro dos distintos modelos de reconocimiento: el primero, basado en transcripciones más conservadoras, podría constituir la base de una edición digital documental[24]; el segundo, basado en criterios de transcripción más modernizantes –normalización de la acentuación, desarrollo de las abreviaturas, regularización de las variantes gráficas– permitiría obtener unas transcripciones de tipo interpretativo. Este segundo modelo de HTR, a pesar de no cumplir con la norma de correspondencia en el caso de las abreviaturas, promete igualmente un alto grado de precisión, como ha demostrado Thöle (2017).

Finalmente, a partir de la transcripción de 20 páginas de cada obra del corpus (correspondientes a 8.500 líneas y alrededor de 67.000 palabras transcritas) logramos crear un modelo de HTR experimental apto para el reconocimiento automático de la letra gótica impresa. Con los textos de nuestro corpus, es decir los que se utilizaron para su creación, el modelo alcanza un margen de precisión muy alto, que se acerca al 1% de error para las páginas no transcritas (Fig. 2).

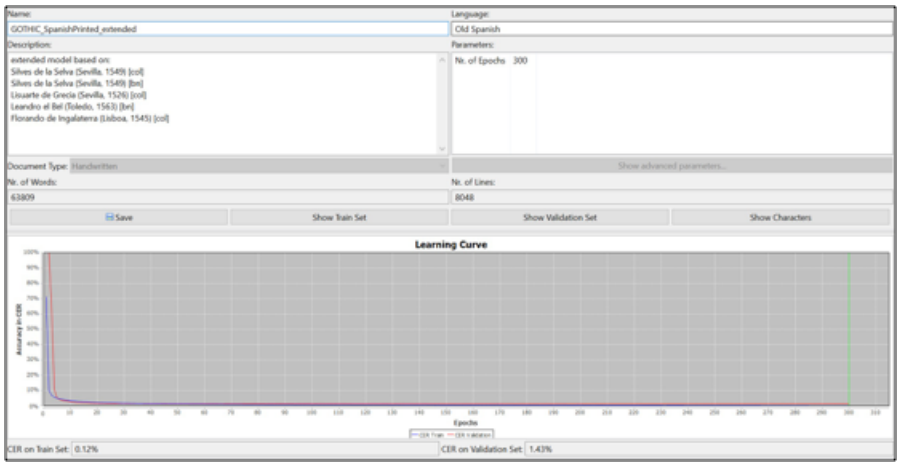


Fig. 2. Detalle de una fase del proceso de *training* y creación del modelo HTR

[24] En opinión de Mancinelli y Pierazzo, las *softwares* de reconocimiento automático de textos están contribuyendo a la creación de un nuevo modelo editorial, la *Digital Documentary Edition*. De esta manera, según las estudiosas, la transcripción en ámbito digital deja de ser una operación preliminar "de servicio" y se convierte en un producto editorial a todos los efectos (2020: 57).

Para comprobar la precisión del modelo, se subieron a la plataforma otras imágenes de páginas de textos en letra gótica, distintos en su representación gráfica y disposición tipográfica, procedentes de diferentes impresores. Los resultados con estos textos fueron buenos, pero confirmaron la hipótesis de que será necesaria una progresiva implementación del modelo de HTR para generar otras transcripciones suficientemente correctas. En el futuro, gracias a la colaboración de una amplia comunidad de usuarios, y aumentando la cantidad de textos procesados, confiamos en que será posible perfeccionar el modelo de reconocimiento automático y obtener resultados de transcripción satisfactorios con los textos en letra gótica.

5. CONSIDERACIONES FINALES

Las prácticas de *remediation* de los textos en soporte digital están sometidas a los grandes cambios impuestos por el avance tecnológico. La digitalización masiva de documentos, vehiculada por los gigantes de la web, supuso una proliferación de versiones en texto electrónico de obras literarias repletas de inexactitudes, principalmente debidas a varios factores: la falta de precisión y transparencia en las declaraciones sobre la fuente digitalizada; las condiciones materiales de la fuente; los errores derivados del proceso de reconocimiento automático. Estos problemas, además, se presentan de forma reiterada y aún más despistante y peligrosa en la trasmisión de obras impresas antiguas, porque, junto a los errores propios del reconocimiento automático, el riesgo es que con herramientas de transcripción automática defectuosas se imponga una sistemática uniformización y modernización de formas que no corresponden a lo que se encuentra en diccionarios históricos en uso y ausentes en el mismo texto fuente.

Publicados con metadatos lagunosos, imprecisos en sus formas y contenidos, los textos electrónicos pertenecen por lo tanto a un campo desprovisto de un contexto de legitimación como lo es la publicación en papel, que al contrario cuenta con siglos de tradición metodológica, debate científico y recepción por parte de los especialistas. Para limitar la creación y difusión de ediciones digitales erróneas y engañosas, se propone aquí empezar por una controlada utilización de los *softwares* de reconocimiento automático de textos, de modo que los humanistas vuelvan a tomar su papel desde el principio de la cadena de transmisión de las obras literarias, también dentro del ecosistema digital.

Con esta finalidad, se ha presentado aquí un flujo de trabajo que desde los facsímiles digitales lleva, de forma semi-automática, a la obtención de transcripciones fiables en formato digital. Por lo tanto, para probar la eficiencia de la plataforma *Transkribus* en el reconocimiento de la letra gótica española del Renacimiento, ante todo seleccionamos un corpus de cuatro textos pertenecientes al género de los libros de caballerías.

El proceso de transcripción automática en la plataforma *Transkribus* se articuló en distintos pasajes: la manipulación de las imágenes digitalizadas para facilitar el reconocimiento; la subida de las imágenes a la plataforma; la segmentación automática de cada página escaneada; la transcripción de una porción mínima de cada obra para el entrenamiento del *software*[25] y la obtención de un modelo de HTR.

Toda la operación demostró solidez y llevó a la creación de un modelo de HTR extendido para la letra gótica que puede representar un punto de partida firme para la creación de materiales textuales fiables.

Finalmente, para aprovechar nuestros resultados, el estudioso tiene dos posibles soluciones.

En primer lugar, de forma inmediata, puede utilizar el modelo que creamos y que se pone a disposición de los usuarios de Transkribus en la sección "*public models*". El nombre del modelo es SpanishGothic_XVIc[26]. De tal manera, es posible conseguir de forma automática la transcripción de la obra elegida, pero este simple pasaje no asegura un alto índice de precisión en los resultados. Sin embargo, el usuario puede beneficiarse de los numerosos instrumentos disponibles dentro de la plataforma *Transkribus*. Por ejemplo, puede codificar el texto en formato XML Tei, maquetando el contenido a través de etiquetas específicas que indiquen las intervenciones editoriales o la presencia de nombres de personajes y lugares. Además, puede emplear la función de búsqueda de palabras clave (*Keyword Spotting*) para detectar las palabras que se asemejan en su representación gráfica y visualizarlas en un listado organizado jerárquicamente[27].

---

[25] Les agradezco a Stefano Neri y a Giada Blasut la ayuda que me dieron con las transcripciones.

[26] Para más informaciones acerca de Transkribus y la utilización de modelos HTR públicos ver la página wiki del software: <https://transkribus.eu/wiki/index.php/Main_Page> [07/10/2020].

[27] Ver al respecto: <https://transkribus.eu/wiki/images/1/1b/HowToTranscribe_Keyword_Search.pdf>. En el ámbito de la literatura del Siglo de Oro, la función de Keyword Spotting fue presentada por primera vez por parte del PRHLT (Universitat Politècnica de València), con el

En alternativa se ofrece la perspectiva de una colaboración directa con nuestro trabajo: en el caso de que el estudioso quiera implementar el modelo HTR que creamos, tendrá que subir a la plataforma 20 imágenes digitalizadas de la fuente en gótica de su interés, lanzar el proceso de *Layout Analysis* y transcribir línea por línea el texto. Una vez ultimado el proceso de transcripción podrá contactar con el autor del presente artículo y compartir su propia colección[28]. De tal manera su trabajo se convertirá en material útil para otros especialistas interesados en la misma letra, porque el modelo de HTR implementado consentirá ampliar convenientemente las posibilidades de reconocimiento automático.

Concluimos señalando que el presente trabajo comparte la que Fusi[29] define como una filosofía de *crowd-sourcing* típica del *framework Transkribus*, donde los contenidos, y en particular los modelos de HTR, son enriquecidos y determinados por parte de los usuarios en el ámbito de investigaciones particulares o de contribuciones voluntarias. La sostenibilidad del proyecto completo para el reconocimiento automático de la letra gótica, por lo tanto, pasa por el crecimiento y el soporte de la comunidad de usuarios, lo cual, a pesar de los costes en términos de tiempo de trabajo, promete evitar que tales operaciones sean tierra de conquista de iniciativas de tipo comercial, consintiendo la adopción de una perspectiva transparente y filológicamente adecuada para la transmisión digital de nuestro patrimonio literario.

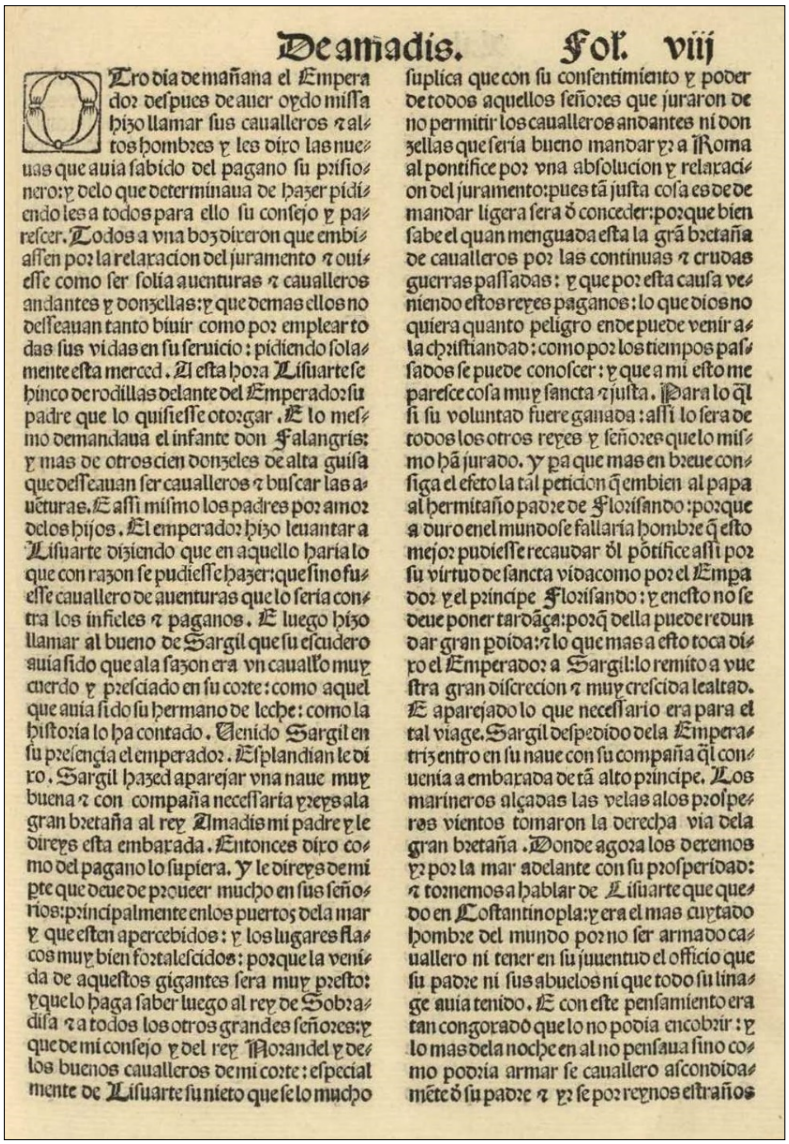
---

apoyo del grupo PROLOPE (Universitat Autònoma de Barcelona) y de la Biblioteca Nacional de España, en ocasión de la exhibición "Lope y el teatro del Siglo de Oro" (BNE, noviembre 2018 – marzo 2019) Enlace: <http://www.bne.es/es/Actividades/Exposiciones/Exposiciones/ Exposiciones 2018/lope-teatro-siglo-de-oro.html> [08/10/2020].

[28] El contacto puede darse por medio de correo electrónico (stefano.bazzaco.1@gmail.com) o directamente dentro de la plataforma Transkribus con la función "*manage users*", que permite añadir usuarios a la colección seleccionada.

[29] Entrevista a la página <https://www.gramsciforthehumanities.org/le-potenzialita-di-transkribus-per-la-ricerca-testuale-lopinione-di-daniele-fusi/> [08/10/2020].

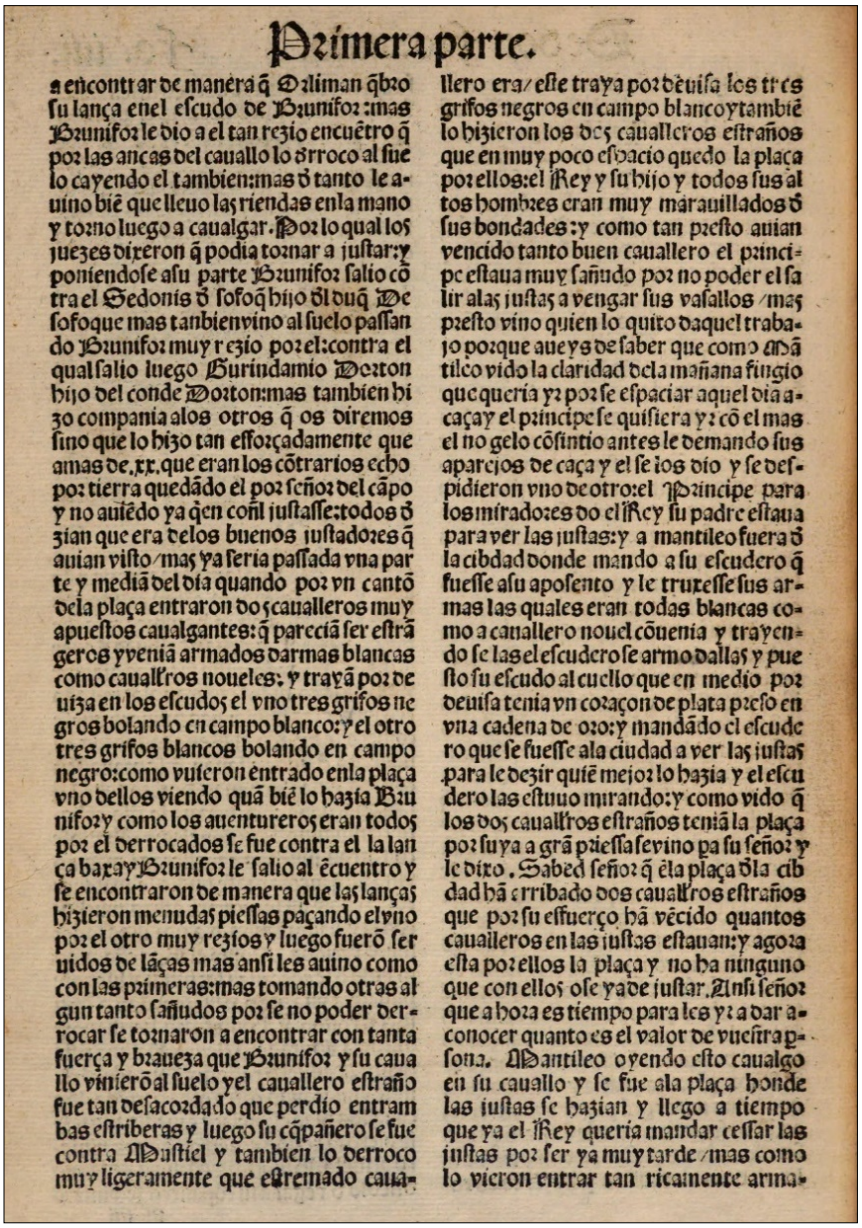
APÉNDICE



De amadis. Fol. viij

Otro dia de mañana el Empera
dor despues de auer oydo missa
hizo llamar sus caualleros ⁊ al
tos hombres y les dixo las nue
uas que auia sabido del pagano su prisio
nero: y delo que determinaua de hazer pidi
endo les a todos para ello su consejo y pa
rescer. Todos a vna boz dixeron que embi
assen por la relaxacion del juramento ⁊ oui
esse como ser solia auenturas ⁊ caualleros
andantes y donzellas: y que demas ellos no
desseauan tanto biuir como por emplear to
das sus vidas en su seruicio: pidiendo sola
mente esta merced. Al esta hora Lisuarte se
hinco de rodillas delante del Emperador: su
padre que lo quisiesse otorgar. E lo mes
mo demandaua el infante don Falangris:
y mas de otros cien donzeles de alta guisa
que desseauan ser caualleros ⁊ buscar las a
uẽturas. E assi mismo los padres por amor
delos hijos. El emperador hizo leuantar a
Lisuarte diziendo que en aquello haria lo
que con razon se pudiesse hazer: que si no fu
esse caualleros de auenturas que lo seria con
tra los infieles ⁊ paganos. E luego hizo
llamar al bueno de Sargil que su escudero
auia sido que ala sazon era vn caualleo muy
cuerdo y preciado en su corte: como aquel
que auia sido su hermano de leche: como la
historia lo ha contado. Venido Sargil en
su presencia el emperador. Esplandian le di
xo. Sargil hazed aparejar vna naue muy
buena ⁊ con compañia necessaria y reys ala
gran bretaña al rey Amadis mi padre y le
direys esta embarada. Entonces dixo co
mo del pagano lo supiera. Y le direys de mi
pte que deue de proueer mucho en sus seño
rios: principalmente en los puertos dela mar
y que esten apercebidos: y los lugares fla
cos muy bien fortalescidos: porque la veni
da de aquestos gigantes sera muy presto:
y que lo haga saber luego al rey de Sobra
disa ⁊ a todos los otros grandes señores: y
que de mi consejo y del rey Norandel y de
los buenos caualleros de mi corte: especial
mente de Lisuarte su nieto que se lo mucho
suplica que con su consentimiento y poder
de todos aquellos señores que juraron de
no permitir los caualleros andantes ni don
zellas que seria bueno mandar y a Roma
al pontifice por vna absolucion y relaxaci
on del juramento: pues tã justa cosa es de de
mandar ligera sera ỗ conceder: porque bien
sabe el quan menguada esta la grã bretaña
de caualleros por las continuas ⁊ crudas
guerras passadas: y que por esta causa ve
niendo estos reyes paganos: lo que dios no
quiera quanto peligro ende puede venir a
la christiandad: como por los tiempos pas
sados se puede conoscer: y que a mi esto me
paresce cosa muy sancta ⁊ justa. Para lo q̃l
si su voluntad fuere ganada: assi lo sera de
todos los otros reyes y señores que lo mis
mo hã jurado. Y pa que mas en breue con
siga el efeto la tal peticion q̃ embien al papa
al hermitaño padre de Florisando: porque
a duro enel mundo se fallaria hombre q̃ esto
mejor pudiesse recaudar ỗl põtifice assi por
su virtud de sancta vida como por el Empa
dor y el principe Florisando: y en esto no se
deue poner tardãça: por q̃ della puede redun
dar gran poida: ⁊ lo que mas a esto toca di
xo el Emperador a Sargil: lo remito a vue
stra gran discrecion ⁊ muy crescida lealtad.
E aparejado lo que necessario era para el
tal viage. Sargil despedido dela Empera
triz entro en su naue con su compañia q̃l con
uenia a embarada de tã alto principe. Los
marineros alçadas las velas alos prospe
res vientos tomaron la derecha via dela
gran bretaña. Donde agora los dexemos
y: por la mar adelante con su prosperidad:
⁊ tornemos a hablar de Lisuarte que que
do en Costantinopla: y era el mas cuytado
hombre del mundo por no ser armado ca
uallero ni tener en su juuentud el officio que
su padre ni sus abuelos ni que todo su lina
ge auia tenido. E con este pensamiento era
tan congoxadỗ que lo no podia encobrir: y
lo mas dela noche en al no pensaua sino co
mo podria armar se cauallero ascondida
mẽte ỗ su padre ⁊ y: se por reynos estraños

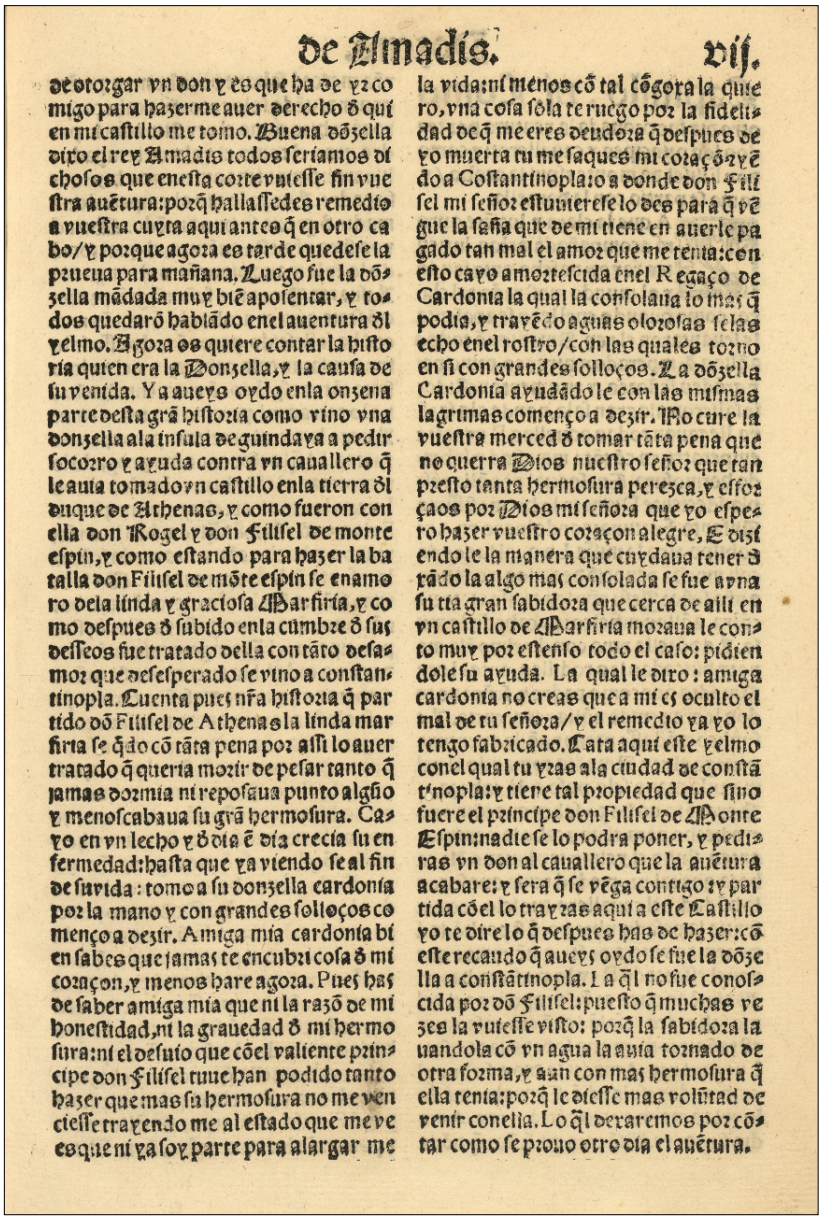
Lisuarte de Grecia, Juan Díaz, imp. Jacobo y Juan Cromberger, 1526, Sevilla. Fol. 8r.



Primera parte.

a encontrar de manera q̃ Oliman q̃bro su lança enel escudo de Brunifor: mas Brunifor le dio a el tan rezio encuẽtro q̃ por las ancas del cauallo lo derroco al suelo cayendo el tambien: mas ỹ tanto le auino biẽ que lleuo las riendas enla mano y torno luego a caualgar. Por lo qual los juezes dixeron q̃ podia tornar a justar: y poniendose asu parte Brunifor salio cõtra el Sedonis ỹ sofoq̃ hijo dl duq̃ De sofoque mas tanbien vino al suelo passando Brunifor muy rezio por el: contra el qual salio luego Burindamio Dorton hijo del conde Dorton: mas tambien hizo compania alos otros q̃ os diremos sino que lo hizo tan esforçadamente que amas de. xx. que eran los cõtrarios echo por tierra quedãdo el por señor del cãpo y no auiẽdo ya q̃en conl justasse: todos ỹ zian que era delos buenos justadores q̃ auian visto / mas ya seria passada vna parte y mediã del dia quando por vn cantõ dela plaça entraron dos caualleros muy apuestos caualgantes: q̃ parecia ser estrãgeros y venia armados darmas blancas como caualleros noueles: y trayã por deuiza en los escudos el vno tres grifos negros bolando en campo blanco: y el otro tres grifos blancos bolando en campo negro: como vuieron entrado enla plaça vno dellos viendo quã biẽ lo hazia Brunifor y como los auentureros eran todos por el derrocados se fue contra el la lança baxay Brunifor le salio al ẽcuentro y se encontraron de manera que las lanças hizieron menudas piessas paçando el vno por el otro muy rezios y luego fuerõ seruidos de lãças mas ansi les auino como con las primeras: mas tomando otras algun tanto sañudos por se no poder derrocar se tornaron a encontrar con tanta fuerça y braueza que Brunifor y su cauallo vinierõ al suelo y el cauallero estraño fue tan desacordado que perdio entrambas estriberas y luego su cõpañero se fue contra Mantiel y tambien lo derroco muy ligeramente que estremado caualllero era / este traya por deuisa los tres grifos negros en campo blanco y tambiẽ lo hizieron los dos caualleros estraños que en muy poco espacio quedo la plaça por ellos: el Rey y su hijo y todos sus altos hombres eran muy marauillados ỹ sus bondades: y como tan presto auian vencido tanto buen cauallero el principe estaua muy sañudo por no poder el salir alas justas a vengar sus vasallos / mas presto vino quien lo quito daquel trabajo porque aueys de saber que como Mãtileo vido la claridad dela mañana fingio que queria yr por se espaciar aquel dia a caça y el principe se quisiera yr cõ el mas el no gelo cõsintio antes le demando sus aparejos de caça y el se los dio y se despidieron vno de otro: el Principe para los miradores do el Rey su padre estaua para ver las justas: y a mantileo fuera ỹ la cibdad donde mando a su escudero q̃ fuesse asu aposento y le truxesse sus armas las quales eran todas blancas como a cauallero nouel cõuenia y trayendo se las el escudero se armo dallas y puesto su escudo al cuello que en medio por deuisa tenia vn coraçon de plata preso en vna cadena de oro: y mandãdo el escudero que se fuesse ala ciudad a ver las justas para le dezir quiẽ mejor lo hazia y el escudero las estuuo mirando: y como vido q̃ los dos caualleros estraños teniã la plaça por suya a grã priessa se vino pa su señor: y le dixo. Sabed señor q̃ ẽla plaça ỹ la cibdad hã derribado dos caualleros estraños que por su esfuerço hã vẽcido quantos caualleros en las justas estauan: y agora esta por ellos la plaça y no ha ninguno que con ellos ose ya de justar. Ansi señor que a hora es tiempo para les yr a dar a conocer quanto es el valor de vuestra persona. Mantileo oyendo esto caualgo en su cauallo y se fue ala plaça honde las justas se hazian y llego a tiempo que ya el Rey queria mandar cessar las justas por ser ya muy tarde / mas como lo vieron entrar tan ricamente arma-

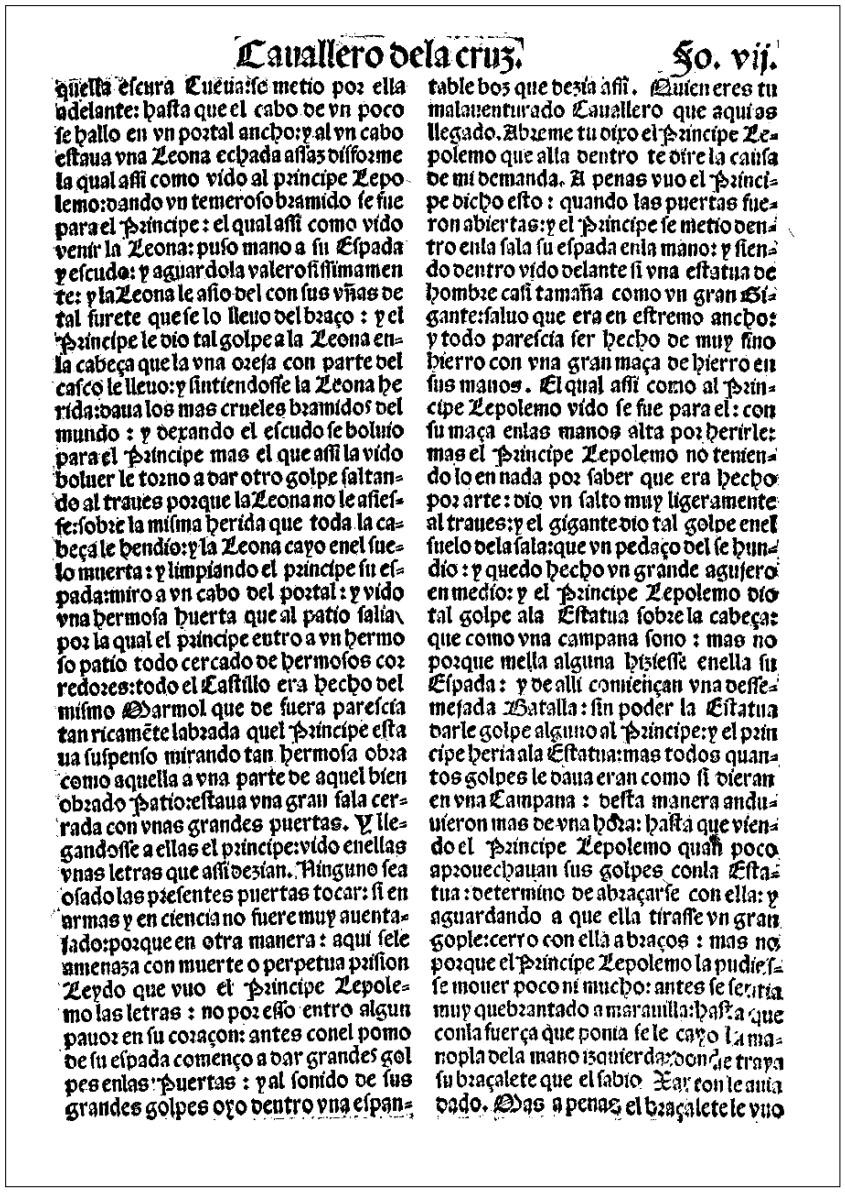
Florando de Inglaterra, imp. G. Gallarde, 1545, Lisboa. Fol. 4v



de Amadís. vij.

de otorgar vn don y eſo que ha de yr co
migo para hazerme auer derecho ỏ qui
en mi caſtillo me tomo. Buena dõzella
dixo el rey Amadis todos ſeriamos di
choſos que eneſta corte vnieſſe fin vue
ſtra auẽtura: porq̃ hallaſſedes remedio
a vueſtra cuyta aqui antes q̃ en otro ca
bo/y porque agora es tarde quedeſe la
pruena para mañana. Luego fue la dõ
zella mãdada muy biẽ apoſentar, y to
dos quedarõ hablãdo enel auentura ỏl
yelmo. Agora os quiere contar la hiſto
ria quien era la Donzella, y la cauſa de
ſu venida. Ya aueys oydo enla onzena
parte deſta grã hiſtoria como vino vna
donzella ala inſula de guindaya a pedir
ſocorro y ayuda contra vn cauallero q̃
le auia tomado vn caſtillo enla tierra ỏl
duque de Athenas, y como fueron con
ella don Rogel y don Filiſel de monte
eſpin, y como eſtando para hazer la ba
talla don Filiſel de mõte eſpin ſe enamo
ro dela linda y graciosa Barfiria, y co
mo deſpues ỏ ſubido enla cũmbre ỏ ſus
deſſeos fue tratado della con tãto deſa
mor que deſeſperado ſe vino a conſtan
tinopla. Cuenta pues nra hiſtoria q̃ par
tido dõ Filiſel de Athenas la linda mar
firia ſe q̃do cõ rãta pena por aſſi lo auer
tratado q̃ queria morir de peſar tanto q̃
jamas dormia ni reposaua punto alguo
y menoſcabaua ſu grã hermoſura. Ca
yo en vn lecho y ỏ dia ẽ dia crecia ſu en
fermedad: haſta que ya viendo ſe al fin
de ſu vida: tomo a ſu donzella cardonia
por la mano y con grandes ſollocos co
mençoa dezir. Amiga mia cardonia bi
en ſabes que jamas te encubri coſa ỏ mi
corazon, y menos hare agora. Pues has
de ſaber amiga mia que ni la razõ de mi
honeſtidad, ni la grauedad ỏ mi hermo
ſura: ni el deſuio que cõel valiente prin
cipe don Filiſel tuue han podido tanto
hazer que mas ſu hermoſura no me ven
cieſſe trayendo me al eſtado que me ve
es que ni ya ſoy parte para alargar me
la vida: ni menos cõ tal cõgoja la que
ro, vna coſa ſola te ruego por la fidelt
dad deq̃ me eres deudora q̃ deſpues de
yo muerta tu me ſaques mi corazõ y ẽ
do a Conſtantinopla: o a donde don Fili
ſel mi ſeñor eſtuuiereſe lo des para q̃ vẽ
gue la ſaña que de mi tiene en auerle pa
gado tan mal el amor que me tenia: con
eſto cayo amortecida enel Regaço de
Cardonia la qual la conſolaua lo mas q̃
podia, y trayẽdo aguas olorosas ſe las
echo enel roſtro/con las quales torno
en ſi con grandes ſollocos. La dõzella
Cardonia ayudãdo le con las miſmas
lagrimas començoa dezir. No cure la
vueſtra merced ỏ tomar tãta pena que
no querra Dios nueſtro ſeñor que tan
preſto tanta hermoſura perezca, y eſfor
çaos por Dios mi ſeñora que yo eſpe
ro hazer vueſtro corazon alegre, E diz
iendo le la manera que cuydaua tener ỏ
tãdo la algo mas conſolada ſe fue a vna
ſu tia gran ſabidora que cerca de alli en
vn caſtillo de Barfiria morava le con
to muy por eſtenſo todo el caſo: pidien
dole ſu ayuda. La qual le dixo: amiga
cardonia no creas que a mi es oculto el
mal de tu ſeñora/y el remedio ya yo lo
tengo fabricado. Cata aqui eſte yelmo
conel qual tu yras ala ciudad de conſtã
tinopla: y tiene tal propiedad que ſino
fuere el principe don Filiſel de Monte
Eſpin: nadie ſe lo podra poner, y pedi
ras vn don al cauallero que la auẽtura
acabare: y ſera q̃ ſe vẽga contigo: y par
tida cõel lo trayras aqui a eſte Caſtillo
yo te dire lo q̃ deſpues has de hazer: cõ
eſte recaudo q̃ aueys oydo ſe fue la dõze
lla a conſtãtinopla. La q̃l no fue conoſ
cida por dõ Filiſel: pueſto q̃ muchas ve
zes la vuieſſe viſto: porq̃ la ſabidora la
uandola cõ vn agua la auia tornado de
otra forma, y aun con mas hermoſura q̃
ella tenia: porq̃ le dieſſe mas voluntad de
venir conella. Lo q̃l dexaremos por cõ
tar como ſe propuo otro dia el auẽtura.

Silves de la Selva, Pedro de Luján, imp. D. De Robertis, 1549, Sevilla. Fol. 7r.



Caualleró dela cruz. Fo. vij.

quella escura Cueua:se metio por ella adelante: hasta que el cabo de vn poco se hallo en vn portal ancho: y al vn cabo estaua vna Leona echada assaz disforme la qual assi como vido al príncipe Lepolemo: dando vn temeroso bramido se fue para el Príncipe: el qual assi como vido venir la Leona: puso mano a su Espada y escudo: y aguardola valerosissimamente: y la Leona le asio del con sus vñas de tal furete que se lo lleuo del braço: y el Príncipe le dio tal golpe a la Leona en la cabeça que la vna oreja con parte del casco le lleuo: y sintiendosse la Leona herida: daua los mas crueles bramidos del mundo: y dexando el escudo se boluio para el Príncipe mas el que assi la vido boluer le torno a dar otro golpe saltando al traues porque la Leona no le affiesse: sobre la misma herida que toda la cabeça le hendio: y la Leona cayo enel suelo muerta: y limpiando el príncipe su espada: miro a vn cabo del portal: y vido vna hermosa huerta que al patio salia por la qual el príncipe entro a vn hermoso patio todo cercado de hermosos corredores: todo el Castillo era hecho del mismo Marmol que de fuera parescia tan ricamẽte labrada quel Príncipe estaua suspenso mirando tan hermosa obra como aquella a vna parte de aquel bien obrado Patio: estaua vna gran sala cerrada con vnas grandes puertas. Y llegandosse a ellas el príncipe: vido enellas vnas letras que assi dezian. Ninguno sea osado las presentes puertas tocar: si en armas y en ciencia no fuere muy auentajado: porque en otra manera: aqui sele amenaza con muerte o perpetua passion Leydo que vuo el Príncipe Lepolemo las letras: no por esso entro algun pauor en su coraçon: antes conel pomo de su espada començo a dar grandes golpes enlas Puertas: y al sonido de sus grandes golpes oyo dentro vna espantable boz que dezia assi. Quien eres tu malauenturado Caualiero que aqui as llegado. Abreme tu dixo el Príncipe Lepolemo que alla dentro te dire la causa de mi demanda. A penas vuo el Príncipe dicho esto: quando las puertas fueron abiertas: y el Príncipe se metio dentro enla sala su espada enla mano: y siendo dentro vido delante si vna estatua de hombre casi tamaña como vn gran Gigante: saluo que era en estremo ancho: y todo parescia ser hecho de muy fino hierro con vna gran maça de hierro en sus manos. El qual assi como al Príncipe Lepolemo vido se fue para el: con su maça enlas manos alta por herirle: mas el Príncipe Lepolemo no temiendo lo en nada por saber que era hecho por arte: dio vn salto muy ligeramente al traues: y el gigante dio tal golpe enel suelo dela sala: que vn pedaço del se hundio: y quedo hecho vn grande aguiero en medio: y el Príncipe Lepolemo dio tal golpe ala Estatua sobre la cabeça: que como vna campana sono: mas no porque mella alguna hiziesse enella su Espada: y de alli comiençan vna desemesada Batalla: sin poder la Estatua darle golpe alguno al Príncipe: y el Príncipe heria ala Estatua: mas todos quantos golpes le daua eran como si dieran en vna Campana: desta manera anduuieron mas de vna hora: hasta que viendo el Príncipe Lepolemo quan poco aprouechauan sus golpes conla Estatua: determino de abraçarse con ella: y aguardando a que ella tirasse vn gran gople: cerro con ella a braços: mas no porque el Príncipe Lepolemo la pudiesse mouer poco ni mucho: antes se sentia muy quebrantado a maranilla: hasta que conla fuerça que ponia se le cayo la manopla dela mano izquierda: donde traya su braçalete que el sabio Xartonte auia dado. Mas a penas el braçalete le vuo

Leandro el Bel, Pedro de Luján, imp. M. Ferrer, 1563, Toledo. Fol. 7r.

Bibliografía

Alvermann, Dirk y Bruno Blüggel, "Transkribus at Greifswald. Idea, practice, results, perspective", ponencia dictada en *Transkribus User Conference*, 2–3 November 2017, Technical University of Vienna, Vienna <https://readcoop.eu/wp-content/uploads/2017/07/Alvermann_Bluegel_Greifswald.pdf > [07/10/2020].

Bazzaco, Stefano, "El Progetto Mambrino y las tecnologías OCR: estado de la cuestión", *Historias Fingidas*, 6 (2018), pp. 257-272.

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

Borghi, Maurizio y Stavroula Karapapa, "Dal cartaceo al 'digitale di massa': biblioteche virtuali, diritto d'autore e il caso Google Books", en *Teoria e forme del testo digitale*, introducción, edición y notas de Michelangelo Zaccarello, Roma, Carocci Editore, 2019, pp. 95-113.

Buzzetti, Dino y Jerome McGann, "Critical editing in a digital horizon", en *Electronic Textual Editing*, eds. Lou Burnard y Katherine O'Brien O'Keeffe, Nueva York, Modern Language Association of America, 2006, 53-73.

Carbonell, Jamie G; Michalski, Ryszard S.; Mitchell, Tom M., "An overview of machine learning", en *Machine Learning: An Artificial Intelligence Approach*, eds. Jamie G. Carbonell, Ryszard S. Michalski y Tom M. Mitchell, Berlin-Heidelberg, Springer-Verlag, 2013, pp. 3-23.

Caton, Paul, "On the term 'text' in digital humanities", *Literary and Linguistic Computing*, n. 28 (2013), pp. 209-220.

Fiormonte, Domenico y Valentina Martiradonna, "La representación digital de la génesis del texto: un caso de estudio", en *En el taller del escritor: génesis textual y edición de textos*, Bilbao, Universidad del País Vasco, 2010, pp. 147-176.

Griffin, Clive, *Los Cromberger. La historia de una imprenta del siglo XVI en Sevilla y Méjico*, Madrid, Eds. Cultura Hispánica, 1991.

Haebler, Conrado, *Tipografía Ibérica del siglo XV. Reproducción en facsímile de todos los caracteres tipográficos empleados en España y Portugal hasta el año de 1500*, La Haya-Leipsig, a costa de Martinus Nijhoff y Karl W. Hiersemann, 1902.

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*, introducción, edición y notas de Michelangelo Zaccarello, Roma, Carocci Editore, 2019, pp. 135-166.

Leifert, Gundram *et al*., "CITlab ARGUS for historical handwritten documents", 2016, arXiv:1605.08412 <https://www.researchgate.net/publication/269577757_CITlab_ARGUS_for_historical_handwritten_documents> [07/10/2020].

Lucía Megías, José Manuel, *Elogio del texto digital. Claves para interpretar el nuevo paradigma*, Madrid, Fórcola Ediciones, 2012.

Mancinelli, Tiziana y Elena Pierazzo, *Che cos'è un'edizione scientifica digitale*, Roma, Carocci Editore, 2020.

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.

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

Norton, Frederick John, *Printing in Spain, 1501-1520*, Cambridge, Cambridge University Press, 1966.

Nunberg, Geoffrey, "Google's Book Search: A disaster for scholars", en *The Chronicle of Higher Education*, 31 agosto 2009 <https://www.chronicle.com/article/googles-book-search-a-disaster-for-scholars/> [08/10/2020].

Ogilvie, Brian, "Scientific Archives in the Age of Digitization", *Isis*, vol. 107, n. 1 (2016), pp. 77-85.

Orlandi, Tito, *Informatica testuale. Teoria e prassi*, Roma-Bari, Editori Laterza, 2010.

Priani, Ernesto, "El texto digital y la disyuntiva de las humanidades digitales", *Palabra Clave*, n. 18 (2015), pp. 1215-1234.

Reul *et al.*, "State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines", en *Proceedings of the DHd 2019 Digital Humanities: Multimedial & Multimodal*, Mainz, 2019 <https://arxiv.org/ftp/arxiv/papers/1810/1810.03436.pdf> [08/10/2020].

Robinson, Peter, "Il contesto 'collaborativo' degli studi letterari e la dimensione 'sociale' delle edizioni scientifiche", en *Teoria e forme del testo digitale*, introducción, edición y notas de Michelangelo Zaccarello, Roma, Carocci Editore, 2019, pp. 115-133.

Roling, Marco, "Does Handwriting Text Recognition work for damaged archives?", 2020 <https://www.cortsfoundation.org/pdf/RolingMDP_HTR_on_damaged_archives_V20200317-cmp.pdf> [08/10/2020].

Roncaglia, Gino, *La quarta rivoluzione. Sei lezioni sul futuro del libro*, Roma-Bari, Laterza, 2009.

Sahle, Patrick, "What is a scholarly digital edition (SDE)?", *Digital Scholarly Editing. Theory, Practice and Future Perspectives*, eds. Matthew Driscoll y Elena Pierazzo, Cambridge, Open Book Publishers, 2016, pp. 19-39.

Smith, David A. y Ryan Cordell, *A Research Agenda for Historical and Multilingual Optical Character Recognition*, NULab – Northeastern University, 2018 <https://repository.library.northeastern.edu/downloads/neu:m043p093w?datastream_id=content> [08/10/2020].

Springmann, Uwe y Anke Lüdeling, "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus", *DH Quartely*, vol. 11, n. 2 (2017), sin paginación <http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html> [08/10/2020].

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

Terras, Melissa, "Cultural Heritage Information: Artefacts and Digitization Technologies", en *Cultural Heritage information: Access and Management*, ed. Ian Ruthven y Gobinda Chowdhury, 2015, pp. 63-88.

Thöle, Karen, "Transcribing a highly abbreviated incunable (and some more manuscript sources)", ponencia dictada en *Transkribus User Conference*, 2–3 November 2017, Technical University of Vienna, Vienna <https://read.transkribus.eu/wp-ontent/uploads/2017/07/Thoele_Incunable.pdf> [08/10/2020].

Zappulli, Andrea y Sabrina Iorio, "La digitalizzazione dell'Archivio Storico del Banco di Napoli", *DigItalia. Rivista del digitale nei beni culturali*, año XIII, n. 2 (2018), pp. 46-51.