PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Las ediciones de Arnao Guillén de Brocar de BECLaR transcritas con ayuda de Transkribus y OCR4all: creación de un modelo para la red neuronal y posible explotación de los resultados

Manuel Ayuso García

(Universidad Nacional de Educación a Distancia)*

Abstract

Entre el software disponible para el reconocimiento de textos impresos antiguos he decidido emplear dos sistemas, Transkribus y OCR4all, para la transcripción diplomática de las ediciones de Arnao Guillén de Brocar. Se pretende, por una parte, poner de manifiesto las características tipográficas y editoriales de las ediciones de clásicos latinos impresas por Arnao, relevantes para la creación de un modelo de entrenamiento de las redes neuronales empleadas por Transkribus y OCR4all. En segundo lugar, el objetivo es el de presentar algunas herramientas y métodos para mejorar los resultados de la transcripción. Aunque el trabajo aún debe perfeccionarse, ya ofrece resultados que merecen compartirse.
Palabras clave: OCR de impresos antiguos; Modelos de Redes Neuronales; Arnao Guillén de Brocar; Transkribus; OCR4all

Among the software available for the recognition of old printed texts, I have decided to use two different tools, Transkribus and OCR4all, for the diplomatic transcription of the Arnao Guillén de Brocar's editions. Firstly, the following research wants to point out the typographic and editorial characteristics of the editions of Latin classics printed by Arnao that are outstanding for the creation of a training model for the neural networks system on which Transkribus and OCR4all are based. Secondly, it will be intended to present some tools and methods to improve transcription results. Actual outcomes deserve to be shared, though the work is still at an early stage.
Keywords: early printed books OCR; Recurrent Neural Networks; Arnao Guillén de Brocar; Transkribus; OCR4all

§

* Este trabajo se inscribe en el marco de los Proyectos de Investigación PGC 2018-094609-B-I00 (Ministerio de Ciencia e Innovación y Fondo Europeo de Desarrollo Regional, FEDER) y PR[19]_CLA_0084 (Programa Logos, Fundación BBVA de ayudas a la investigación en el área de Estudios Clásicos).

Manuel Ayuso García «Las ediciones de Arnao Guillén de Brocar de BECLaR transcritas con ayuda de Transkribus y OCR4all: creación de un modelo para la red neuronal y posible explotación de los resultados», *Historias Fingidas*, Número Especial 1 (2022) Humanidades Digitales y estudios literarios hispánicos, pp. 151-173.
DOI: https://doi.org/10.13136/2284-2667/1102 - ISSN: 2284-2667.

Introducción

Obtener transcripciones lo más exactas posibles de los textos estudiados es una tarea crucial para proporcionar una herramienta fundamental para la filología. Si el texto o corpus de textos es extenso, esta tarea puede suponer un verdadero cuello de botella en el progreso de una investigación. En el proyecto BECLaR[1] disponer de la transcripción exacta del texto y de los paratextos de las ediciones de nuestro corpus nos permite analizar con más rigor y exactitud las relaciones textuales de las ediciones y otras características importantes para el análisis filológico. Idealmente y para un estudio más profundo de las ediciones del proyecto BECLaR, querríamos contar con la transcripción del corpus completo con una tasa de acierto próxima al 99,9%. La exactitud de los resultados de la transcripción se mide generalmente por la 'Tasa de error de caracteres' (en inglés *Character Error Rate,* o CER), o sea el porcentaje de caracteres reconocidos erróneamente en un conjunto dado. Aunque la revisión por parte del filólogo siempre será necesaria, la parte más costosa en tiempo se puede hacer ya de manera semiautomática.

Este proceso conocido por su acrónimo inglés, OCR (*Optical Character Recognition*) es una cuestión resuelta para los impresos modernos, en los que un CER < 0,5% se consigue con cualquier aplicación comercial o de software libre programada para esta tarea. Sin embargo, este cometido dista aún de estar resuelto para los impresos anteriores a la invención prensa automática en el siglo XIX y presenta generalmente peores resultados con los impresos más antiguos[2]. Los sistemas que se ocupan de esta tarea se basan en las llamadas *RNN, Recurrent neural networks* (Redes neuronales recurrentes) que constituyen una de las aplicaciones más comunes de la llamada AI (Inteligencia Artificial) y el *Machine Learning*.

Las aplicaciones para el reconocimiento del texto de las ediciones antiguas comprenden una parte fundamental que consiste en usar elementos de Inteligencia Artificial para enseñar a la máquina a reconocer el texto en una imagen. Para lograr esto hay que entrenar una red neuronal recurrente a partir de un conjunto de datos verdaderos, es decir, hace falta

[1] <https://www.incunabula.uned.es/>(cons. 08/05/2022).
[2] Para tener un panorama de esta cuestión, véase Springmann *et al.* (2014; 2016).

proporcionar a la máquina ciertas imágenes que contienen texto y su transcripción correspondiente.

Entre los sistemas disponibles para este cometido he usado dos[3]; de manera principal Transkribus[4] y como complemento y contraste OCR4all[5]. Ambos están disponibles libremente, si bien aquel requiere el pago de algunas partes del proceso. Este último sistema es de código abierto y tiene como pieza fundamental Calamari[6], que se basa a su vez en el sistema OCRopus[7].

Cada uno de estos sistemas emplea diversas piezas de software, pero en ambos la parte nuclear la constituye el entrenamiento de una o varias RNN. Sin entrar en detalles que no son el cometido de este trabajo, ambos sistemas emplean redes neuronales de distinta tipología, cuyos parámetros se pueden ajustar para afinar los resultados[8].

Transkribus y OCR4all, mediante los modelos disponibles, pueden proporcionar datos de transcripción, de tal forma que se puede obtener una primera transcripción sin teclear ningún texto, aunque con errores. Estos modelos amplían cada día la cantidad de texto verificado de manera que cada vez ofrecen resultados más exactos con nuevos datos, aunque el resultado puede variar mucho de unos textos a otros[9]. A partir del primer resultado se selecciona una parte del texto para corregir la transcripción proporcionada de manera automática y crear una transcripción sin errores.

Con estos datos el sistema se entrena y la máquina aprende esta tarea. Se creará así un nuevo modelo para conseguir un reconocimiento automático de los textos con menos errores. Este proceso se puede repetir

[3] Sobre mi experiencia anterior con ambos sistemas, véase Ayuso Gracía (2017; 2021)

[4] Sobre los fundamentos de Transkribus, cfr. Kahle *et al.* (2017) y la web del proyecto: <https://readcoop.eu/transkribus/> (cons. 15/05/2022).

[5] Sobre los fundamentos de OCR4all, véase Reul *et al.* (2019) y la web del proyecto: <www.OCR4all.org> (cons. 15/05/2022). Para la descarga e instalación de ambos sistemas de software, se vean los documentos proporcionados en ambos sitios web: <https://readcoop.eu/transkribus/> y <www.OCR4all.org/> (cons. 15/05/2022).

[6] <https://github.com/Calamari-OCR/calamari> (cons. 15/05/2022).

[7] <https://github.com/ocropus> (cons. 15/05/2022).

[8] Para conocer más detalles sobre la tipología de las redes, véase la documentación proporcionada por el sitio web de ambos proyectos, que puede consultarse en las referencias bibliográficas.

[9] En el caso de algunos de los modelos disponibles de Transkribus, como Noscemus GM 4.0, el conjunto de entrenamiento consta de 541˙611 palabras o 81˙555 líneas con el que se consigue una CER del 0,79%.

hasta conseguir la menor tasa de error posible. Combinando los dos sistemas Transkribus y OCR4all, con la misma transcripción podemos afinar aún más los resultados.

El presente trabajo muestra la metodología empleada y el proceso para lograrlo. El resultado perseguido de obtener transcripciones exactas aún debe mejorarse y contrastarse con más datos. La finalidad última será conseguir transcripciones lo más exactas posibles con la menor cantidad posible de texto introducido manualmente.

Corpus de trabajo

El punto de partida es el corpus de ediciones estudiadas en el proyecto BECLaR para acometer el trabajo, como se acaba de exponer. Forman parte del mismo ya más de 200 ediciones de los dos primeros siglos de la imprenta de textos mayoritariamente latinos, pero también de sus traducciones castellanas y catalanas. Con la idea de hacer una primera aproximación para automatizar este proceso de transcripción se ha elegido un grupo de ediciones dentro del conjunto de trabajo con la característica común de haber salido de las prensas dirigidas por Arnao Guillén de Brocar. Si el modelo consigue un buen resultado, será fácil extenderlo creando nuevos modelos basados en este en los que se añadan textos de ediciones con distintas tipografías, disposiciones de página y otras características tipográficas hasta lograr el ideal de transcribir el corpus completo[10].

Pese a tratarse de impresos con características tipobibliográficas muy diversas, tienen el nexo común de haberse concebido en el taller dirigido por el mismo maestro impresor con uso de letrerías, disposición de página y convenciones editoriales repetidas. Es cierto que buena parte de los impresos utilizan exclusivamente tipografía gótica o romana, pero en algunos de los trabajos impresos por Arnao de nuestro corpus se combinan ambas tipografías, de manera que será conveniente trabajar con los sistemas automáticos de redes neuronales recurrentes que contenga

[10] Para la selección de este corpus me he guiado por el trabajo de Springmann y Lüdeling (2017).

caracteres de ambas tipografías. Un tema más difícil de resolver es la tipografía griega que está dispersa por buena parte de la producción.

Son en total 19 ediciones, si bien he empleado solo 13 para este trabajo. Este corpus abarca un lapso temporal que va de 1499 a 1521, incluye obras de Cicerón, Juvenal, Ovidio, Persio, Plauto, Pseudo Catón, Salustio y Séneca. En los talleres de Pamplona se imprimió el único incunable, 8 ediciones en Logroño[11], 7 ediciones complutenses[12] y 3 en Valladolid[13] completan este grupo con 3 traducciones y 16 textos en latín.

Con respecto a las características tipobibliográficas lo más significativo para este trabajo es el uso de dos tipografías: gótica y redonda, a los que se añade algunas palabras en tipografía griega dispersas en varios impresos. La tipografía gótica, predominante en la época de Arnao, se utilizó en 13 ediciones, siempre en los textos castellanos, y la segunda en 7. Solo en la última edición de Arnao se combinaron ambas tipografías, si excluimos portadas, títulos y encabezamientos que a menudo se imprimen en la tipografía alternativa al texto principal. Asimismo la disposición de la página, excluyendo las portadas, muestra una variedad importante: algunas ediciones tienen una plana a línea tirada, otras se presentan a doble columna, algunas más contienen un texto principal rodeado por un comentario de cuerpo menor. Además, varias ediciones cuentan con *marginalia.*

Finalmente, el uso de dígrafos, abreviaturas y ligaduras presenta una gran variación entre unas ediciones y otras. Como es habitual en el periodo, el uso de las abreviaturas no es consistente, de manera que, por ejemplo, la grafía *ñ* puede expandirse como *un* o *um* según el contexto, por citar uno de los ejemplos más repetidos.

Solo he podido disponer de imágenes de 13 ediciones, en la que hay representación de todas las tipografías, ciudades, disposiciones de página y se encuentran las tres traducciones castellanas. Los datos detallados se

[11] Se trata de las ediciones siguientes: Persio, *Saturae* 1504-1505 CECLE0138, Cicerón, *Topica* 1506 CECLE0245, Ps. Catón, *Disticha Catonis* 1506 CECLE0206, Ps. Catón, *Disticha Catonis* 1508 CECLE0209, Ps. Catón, *Disticha Catonis* 1510 CECLE0210, Persio, *Saturae* 1510 CECLE0141, Ps. Catón, *Disticha Catonis* 1511 CECLE0211, Ps. Catón, *Disticha Catonis* 1517 CECLE0212.

[12] Cfr. Villarroel (2019, 111-130) para los trabajos complutenses de clásicos latinos de Arnao.

[13] Salustio, *De Bello Iugurthino, De coniuratione Catilinae* 1519 CECLE0, Juvenal, *Sátira VI.*

pueden consultar en el Anexo 1 al final del presente trabajo[14].

Por lo demás, el corpus presenta las dificultades propias de los impresos de la época para que el OCR sea satisfactorio: impresión de los tipos con diversos resultados por el uso de la prensa manual, espaciado irregular entre palabras y los ya citados usos de ligaduras, abreviaturas, dígrafos y la presencia de caracteres no usados en las tipografías actuales como *s longa*, *r rotunda*, entre otros.

Preparación de la transcripción

Una vez seleccionado el corpus, el siguiente paso es cargar las imágenes de los textos en los sistemas. Antes de dar este paso se deben considerar algunos aspectos fundamentales para obtener unos resultados satisfactorios. Las imágenes deben tener alta resolución, buena nitidez, ángulo adecuado y ausencia de sombras y manchas. Muchas de las grandes bibliotecas del mundo proporcionan en sus sitios web imágenes de las ediciones del corpus de trabajo, pero aún son muchas las que faltan. Los archivos proporcionados por las bibliotecas cuentan generalmente con imágenes con una calidad suficiente para este cometido. Las imágenes tomadas por uno mismo, a veces, no consiguen una fotografía lo suficientemente buena para su empleo. Por este motivo es aconsejable en determinados casos aplicar una corrección a las imágenes. En este campo el software disponible también es inmenso. Me permito ceñirme a la aplicación recomendada para este cometido por el grupo CIS[15] de la Universidad Ludwig Maximilian de Múnich, Scantailor[16]. Esta aplicación es capaz de corregir desviaciones del ángulo, eliminar manchas, cambiar la resolución, dividir páginas, etc. No obstante, no he podido completar un experimento completo para presentar los resultados concretos y rigurosos de la transcripción antes y después del proceso de mejora de las imágenes, pero el resultado final, sin duda, es mejor. Existe una cantidad ingente de

[14] Aún no hemos podido experimentar con las 6 ediciones de las que carecemos de imágenes.
[15] CIS - Center for Information and Language Processing, <https://www.cis.uni-muenchen.de/ueber_uns/index.html> (cons. 15/05/2022).
[16] <https://scantailor.org/> (cons. 15/05/2022).

software de tratamiento de imagen, pero este, en concreto, está diseñado para la mejora de imágenes de texto fotografiadas y automatiza la tarea en buena medida, de manera que se puede mejorar el resultado de todas las imágenes de una edición en un proceso que dura tan solo unos minutos.

En el corpus de trabajo de este artículo contamos con imágenes procedentes de las bibliotecas que conservan los ejemplares para 10 ediciones y disponemos de imágenes tomadas por el grupo de investigación para 3 ediciones. Para las restantes aún no contamos con las digitalizaciones[17].

Tras cargar las imágenes en los sistemas, se procederá a analizar la disposición de la página y segmentarla en zonas y líneas de texto. El proceso es automático en ambos sistemas, pero los resultados pueden editarse y mejorarse en ambos sistemas.

Antes de proseguir es necesario comprobar los resultados y corregirlos si es necesario. En OCR4all se incluye para esta operación la herramienta LAREX acrónimo de *Layout Analysis and Region EXtraction*[18], que se abre en una ventana diferente a la de la aplicación principal, mientras que la interfaz de Transkribus incluye las herramientas para el análisis de la página en la ventana principal de la aplicación.

La siguiente etapa consiste en realizar un reconocimiento del texto empleando alguno de los modelos que los sistemas nos ofrecen.

En el caso de Transkribus se dispone de un elevado número de modelos que se adaptan a nuestro corpus[19]. También OCR4all proporciona diversos modelos adecuados para el corpus de trabajo[20]. La elección del modelo se ha basado en los siguientes parámetros: tipología de textos similares y mayor tamaño de los datos de creación del modelo y coincidencia en el idioma de los textos.

---

[17] Se pueden obtener los datos concretos en el sitio web del proyecto BECLaR.

[18] Si bien forma parte de OCR4all, se puede descargar y usar como pieza independiente. Manual de uso al siguiente enlace: <https://www.uni-wuerzburg.de/fileadmin/10030600/Mitarbeiter/Reul_Christian/Projects/Layout_Analysis/LAREX_Quick_Guide.pdf> (cons. 15/05/2022).

[19] En Transkribus hay un elevado número de modelos basados en dos tecnologías distintas PyLaia, más reciente, y HTR+. Para usar esta herramienta, una vez consumido un crédito inicial, se debe pagar una pequeña cantidad.

[20] Los modelos para el reconocimiento de textos de OCR4all se pueden descargar e instalar desde la URL: <https://github.com/Calamari-OCR/calamari_models> (cons. 15/05/2022). Hay una breve descripción de cada modelo.

Este proceso se puede reiterar con distintos modelos base. No obstante, una ojeada proporciona, en general, impresiones sobre cuál de los modelos ha arrojado una transcripción más adecuada.

Estos resultados provisionales se exportarán con las herramientas ofrecidas por ambos sistemas en formato de salida TXT sobre el cual vamos a realizar algunas operaciones.

Ejecución de la transcripción

El siguiente paso es la transcripción manual de algunas páginas para la creación de los modelos que más tarde se emplearán en el reconocimiento del corpus de estudio.

Se trata del punto crucial del trabajo, de cuya exactitud dependen los resultados finales. Si el rendimiento final no es satisfactorio, se podrán corregir, en primer lugar, las páginas transcritas y añadir después más páginas hasta conseguir un mejor resultado.

Varias son las consideraciones antes de acometer el trabajo de mecanografiar la transcripción. En primer lugar, la selección de las páginas. Estas deben ser representativas del conjunto. Los caracteres que no formen parte de la transcripción no se reconocerán correctamente, pues el sistema no habrá sido entrenado. Por esta razón, es importante que haya en la transcripción una representación suficiente de todos los caracteres del corpus.

Para asegurar que este paso crucial sea correcto he escrito en lenguaje Python un pequeño guion o *script* que devuelve ordenados los caracteres y el número de cada uno de ellos presentes en el conjunto de entrenamiento[21]. Si después de ejecutar este *script* se observa la falta de uno o varios caracteres, se deberá ampliar el conjunto para que los incluya todos. La ausencia de cada carácter crearía un «punto ciego», de modo que el sistema no podría aprender a reconocer dicho carácter.

[21] El *script* de Python que he llamado process_texts.py tiene como argumento de entrada el archivo TXT con la transcripción que se pretende usar para el entrenamiento y devuelve como salida un archivo con los caracteres presentes en la entrada ordenados y enumerados.

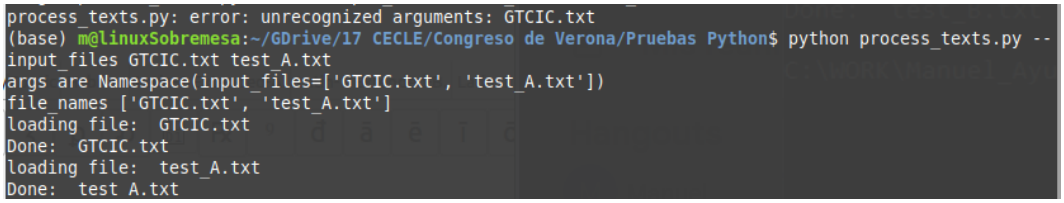


Fig. 1. Terminal en la ejecución del guión de Python

	135	E	9	S	8	i	607	x	16	ɋ	4
	991	F	1	T	7	l	223	ā	16	⁹	2
&	33	G	4	V	6	m	311	ē	10		1
,	4	H	3			n	324	ō	7		
-	34	I	9	a	549	o	310	ũ	1		
.	94	L	3	b	62	p	150	ū	29		
:	76	M	21	c	212	q	79	ſ	286		
;	2	N	11	d	161	r	350	ʒ	12		
?	1	O	9	e	608	s	156	¯	1		
A	11	P	5	f	36	t	484	⍰	2		
C	33	Q	9	g	54	u	510	ꝑ	1		
D	3	R	8	h	9	v	1	ꝙ	11		

Tabla 1. Resultado del guión ejecutado sobre una transcripción

En el ejemplo de la tabla anterior se aprecia la falta de 'B', 'y', por ejemplo, de modo que habrá que añadir a la transcripción líneas que contengan los caracteres ausentes.

Más importante aún será conseguir una transcripción exacta del texto que recoja fielmente el texto transmitido en las ediciones. Esto implica transcribir también las erratas evidentes del texto original. Una transcripción defectuosa dará como resultado un reconocimiento erróneo. La evaluación de estos errores no será correcta, de manera que se distorsionarán los resultados.

La tercera consideración antes de emprender el mecanografiado manual del texto es decidir qué clase de transcripción se quiere, diplomática o normalizada. Si se decide hacer la transcripción diplomática, Transkribus no cuenta, hasta donde hemos podido averiguar, con modelos que permitan una transcripción de esta clase. Por el contrario, OCR4all

ofrece algunos modelos que hacen una primera predicción automática de esta clase.

Transcripción diplomática	Transcripción normalizada
Ciceronis Topica M. TVLIVS CICERO. S.D. C. TREBATIO Ide quāti apd' me ſis: & ſi iure id qui dē. Nō enī te amore uīco uerūtamē qd' p̄ſenti tibi ꝑprie ſubnegarē nō tribuerē: certe id abſenti debere nō potui. Itaqꝫ ut primū Velia nauiga- re coepi: īſtitui Topica Ariſtotelica cōſcribere: ab ipa urbe cōmonit⁹ amātiſſima tui. Et libri tibi miſi Rhegio ſcriptū: q̄ planiſſime res illa ſcribi potuit. Sintibi q̄dā uidebūt' obſcuriora: cogitare debebis: nullā artē fine lenis: fine īterp̄te: & fine ali qua exercitatiōe pcipi poſſe. Nō lōge abieris. nū ius ciuile uſm ex libriscognoſci pōt. Qui q̄q̄ plu rimi ſūt: doctorē in deſiderāt. q̄q̄ ſi tu attēte leges ſaepi⁹: ꝑ te oīa cōſeq̄re: ut certe ītelligas. Vt uero etiā tibi ipi loci ꝓpoſita q̄one occurrit: exercita- tiōe conſeq̄re. In qua quidem nos te continebi- mus: ſi & ſalui redierimus: & ſalua iſta offenderi mus. Vale V. Cal' Sextil'. Rhegio. MARCI TVLLII CICERONIS TOPICO rum liber ad Caium trebatium. M	Ciceronis Topica M. TVLIVS CICERO. S.D. C. TREBATIO Ide quanti apud me sis: & si iure id qui dem. Non enim te amore uinco uerum tamen quod praesenti tibi proprie subnegarem non tribuerem: certe id absenti debere nō potui. Itaque ut primū Velia nauiga- re coepi: īſtitui Topica Aristotelica conscribere: ab ipsa urbe commonitus amantissima tui. Et libri tibi misi Rhegio scriptum: qui planissime res illa scribi potuit. Sintibi quadam uidebuntur obscuriora: cogitare debebis: nullam artem fine lenis: fine interprete: & fine ali qua exercitatione percipi posse. Non longe abieris. num ius ciuile uestrum ex libris cognosci possunt. Qui quamquam plu rimi ſunt: doctorem in desiderant. quamquam si tu attente leges saepius: per te omnia consequere: ut certe intelligas. Vt uero etiam tibi ipsi loci proposita quaestione occurrit: exercita- tione consequere. In qua quidem nos te continebi- mus: si & salui redierimus: & salua ista offenderi mus. Vale V. Calendas Sextilias. Rhegio. MARCI TVLLII CICERONIS TOPICO rum liber ad Caium trebatium.

Tabla 2. Ejemplo de transcripción de la edición de los *Topica* de Cicerón (Logroño 1506)

En este trabajo he recurrido a ambas clases de transcripción combinando los resultados de los dos sistemas. El número de páginas transcritas manualmente ha procurado seguir las recomendaciones de los creadores de cada sistema. Así, para Transkribus he transcrito 8 páginas por cada edición[22], mientras que para OCR4all he empleado entre 60 y 150 líneas[23].

Para la transcripción manual diplomática ambos sistemas cuentan con extensiones de los teclados que permiten la inserción de glifos de manera razonablemente cómoda no presentes en el teclado físico. En ambos sistemas la herramienta para este propósito se llama *Virtual Keyboard*. En Transkribus se abre en una ventana flotante, mientras que en OCR4all ocupa la parte derecha de la ventana de LAREX. Este teclado virtual se puede editar y añadir cualquier glifo con código Unicode, de manera que se puede representar virtualmente cualquier glifo presente en los textos. El *Virtual Keyboard* se puede exportar e instalar en cualquier equipo. En OCR4all es un archivo TXT y en Transkribus XML. La fuente informática deberá ser capaz de reproducir glifos especiales.

Para la tarea de la transcripción manual, Transkribus cuenta con la ventaja de trabajar sobre el servidor remoto de forma que se pueden hacer transcripciones colaborativas compartiendo los documentos del servidor con otros usuarios. Sin embargo, en OCR4all todo el trabajo se hace en local, por consiguiente la colaboración no es fácil para los usuarios no expertos.

Asimismo, Trankribus también cuenta con la herramienta *Text2Image* para acoplar automáticamente transcripciones de texto a las imágenes de un documento, de modo que es posible usar las transcripciones creadas en OCR4all. Esta operación en sentido inverso, Transkribus a OCR4all, no está automatizada.

Como esta parte del proceso es crítica, recomiendo revisar las transcripciones manuales, a ser posible, por varias personas.

---

[22] La documentación recomienda entre 25 y 35 páginas para los textos manuscritos y la tercera parte para los impresos, <https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/> (cons. 15/05/2022).

[23] La recomendación la he hecho siguiendo el trabajo de Reul *et. al.* (2017, 426).

Entrenamiento de los sistemas de reconocimiento y creación de los modelos

En el momento en que la transcripción se considere suficientemente correcta se procede al entrenamiento del sistema con vistas a la creación de un modelo que sirva para el reconocimiento de todo el corpus. Cada modelo obtendrá una CER que servirá para predecir el modelo que arrojará mejores resultados con documentos similares. Este paso es también sustancialmente diferente entre ambos sistemas, pues en Transkribus esta operación la hace el servidor, mientras que con OCR4all es el ordenador local el que ejecuta esta tarea, que puede ser costosa en términos de tiempo.

Las posibilidades para el entrenamiento de modelos son muy numerosas, pues admiten en ambos sistemas muchas combinaciones, con o sin un modelo base, parámetros y —en el caso de Transkribus— elegir tipología de red neuronal. Con los modelos OCR4all he seguido las indicaciones de Reul *et al.* (2017a; 2017b, 38-51).

En las siguientes tablas (3a y 3b) se puede ver un resumen de los modelos creados y la CER lograda en cada uno de ellos[24]. Se puede observar que se ha creado un modelo mixto en el cual se usan ediciones con texto en redonda y en cursiva. A continuación, se han creado modelos solo para tipografía romana o gótica.

Los resultados son todavía modestos y requieren una importante revisión aún para conseguir transcripciones útiles en filología. No obstante, se pueden sacar algunas conclusiones. Los modelos de Transkribus ofrecen un resultado mejor en todos los casos excepto en uno. La excepción es el modelo creado a partir de imágenes propias. Estas imágenes tienen resolución suficiente, pero presentan las líneas con ángulo (*skew*), que resuelve mejor OCR4all. Las líneas transcritas en Transkribus son muchas más, de modo que con respecto al rendimiento el resultado

[24] El modelo base *NOSCEMUS 4.0* usado en Transkribus ha sido compartido por Stefan Zathammer y se basa en los datos de entrenamiento tomados de the Digital Sourcebook of the NOSCEMUS Project <https://www.uibk.ac.at/projects/noscemus/> (cons. 15/05/2022). Por su parte, *Spanish_Gothic_XV-XVI_extended* ha sido compartido por Stefano Bazzaco a partir de los datos del Progetto Mambrino <http://www.mambrino.it/> (cons. 15/05/2022). Para los modelos base de OCR4all no he podido determinar exactamente la autoría.

es mejor con OCR4all, pero tendríamos que comparar cuántas líneas de transcripción son necesarias para obtener un CER aceptable.

Transkribus				
Ediciones base	Nombre del modelo	Modelo Base	CER	# Líneas
Ovidio 1519, Juvenal 1519, Cicerón 1506	Arnao 3 (Pylaia)	--	3,71	1766
Ovidio 1519, Juvenal 1519, Cicerón 1506	Arnao 2 (Pylaia)	--	2,1	1699
Ovidio 1519, Juvenal 1519, Cicerón 1506	Arnao Pylaia (Pylaia)	--	11,3	570
Cicerón 1506	Arnao Latin Roman (Pylaia)	--	5,2	851
Cicerón 1515	Arnao roman2 (Pylaia)	--	8,14	134
Cicerón 1517	Arnao 2 (HTR+)	Noscemus GM 4.0	3,5	270
Ovidio 1519	Arnao_Spanish_Gothic (Pylaia)	--	7,2	114
Ovidio 1519	Arnao_Spanish_Gothic (HTR+)	SpanishGothic_ XV-XVI_extended	0,75	566

Tabla 3a. Reconocimiento con Transkribus (READ Coop)

OCR4all				
Ediciones base	Nombre del modelo	Modelo Base	CER	# Líneas
Ovidio 1519	Ovidio 1519	(+ Fraktur)	9,3	65
Juvenal 1519	Ovidio 1519	(+ Fraktur)	5,84	132
Ovidio 1519, Juvenal 1519, Cicerón 1506	Arnao	Antiqua lig	18,7	176
Cicerón 1515	Cicerón 1515	Antiqua lig	8,6	182

Tabla 3b. Reconocimiento con OCR4all

Por otro lado, los modelos de Transkribus que han logrado mejores resultados son los que han sido entrenados con transcripción normalizada, que es la empleada en los modelos base. Partiendo de cero y usando la

transcripción diplomática, OCR4all logra mejores resultados.

Estas conclusiones solo pueden ser provisionales, pues la experimentación no ha sido completa. Queda acreditado, no obstante, que el resultado mejora con mayor número de líneas transcritas en los conjuntos de entrenamiento.

Reconocimiento de texto

Tras el entrenamiento hemos procedido al reconocimiento empleando los modelos que presentan mejor CER[25].

Mostramos algunos ejemplos de los datos obtenidos en las imágenes y transcripciones que pueden verse en los anexos.

Los resultados se pueden exportar en diversos formatos como XML y TXT en ambos sistemas, pero Transkribus ofrece unas posibilidades más amplias, proporcionando, por ejemplo, PDF con capa de texto o TEI.

El objetivo es lograr tener el texto íntegro y fiable de las ediciones antiguas, de manera que se puedan hacer búsquedas, edición del texto, colaciones, entre otras operaciones, como hacemos con cualquier archivo de texto. Idealmente el texto resultante debería tener una transcripción diplomática y una distribución en páginas y líneas lo más semejante al original. La transcripción diplomática se puede transformar fácilmente en su forma normalizada, mientras que el proceso inverso no es posible. Este proceso de normalización incluye la supresión de las divisiones de palabras por salto de línea, unificación de los caracteres, expandiendo las posibles abreviaturas, etcétera, para diversos propósitos.

Estos resultados deben mejorarse usando las herramientas que ambos sistemas proporcionan para poder ser útiles en filología.

[25] En el caso de Transkribus, los modelos se han entrenado basándose en los de *NOSCEMUS*, creados por el proyecto homónimo, y los modelos *Spanish Gothic* y *Spanish Redonda* creados por el equipo de S. Bazzaco. Con OCR4all he usado los modelos de CALAMARI antiqua ligature.

Anexo 1: Tabla de ediciones

Se presentan las ediciones del corpus de trabajo con sus datos más relevantes y el número de identificación de las mismas en el proyecto BECLaR, donde se pueden obtener información detallada de las mismas.

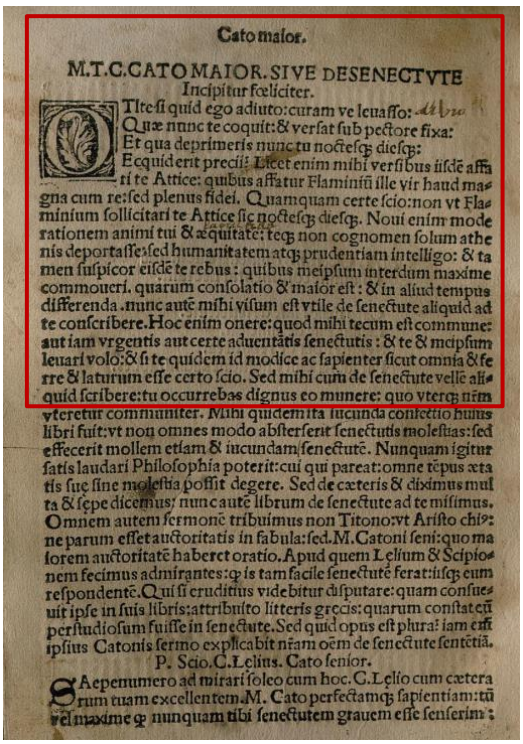
Aclaro el contenido de algunas columnas de esta tabla: «Título» aparece el título normalizado de la obra latina. «D.(isposición) de la pág.(ina)» se informa de la disposición en línea tirada (LT), dos columnas (2 col.) o dos textos, uno rodeando al otro (2 text.), que además puede tener *marginalia* (m.) seguido por el número de líneas de la página. Por último, la columna Img. B/P informa de la procedencia de las imágenes de la edición: de una biblioteca (B) o de mis propias fotografías (P). Si no se dispone aún de ellas figura (NO).

Fecha	Ciudad	Título	Tipografía	Idioma	D. pag.	ID_CECLE	# text lines	Img. B/P
1499	Pamplona	Disticha Catonis	G.	Lat.	LT 26	82	240	B
1505	Logroño	Saturae (Pers.)	G.	Lat.	2 text., 62	138	2760	B
1506	Logroño	Disticha, Fabulae	G	Lat.	LT 26	206	240	NO
1506	Logroño	Topica (Cic.)	R.	Lat.	2 col.,	245		B
1508	Logroño	Disticha, Fabulae	G	Lat.	LT, 26	209	240	NO
1510	Logroño	Disticha, Fabulae	G.	Lat.	LT 36	210	240	B
1510	Logroño	Saturae (Pers.)	G.	Lat.	LT 32	141	680	B
1511	Logroño	Disticha, Fabulae	G	Lat.	LT 36	211		NO
1514	Alcalá de H.	Saturae (Pers.)	G	Lat.	2 text, 43, m.	142	2744	B
1515	Alcalá de H.	Orationes (Cic.)	R.	Lat.	LT, 32	265	640	P
1517	Alcalá de H.	Comoediae 1 (Plaut.)	R.	Lat.	LT, 38, m.	204	9480	NO

1517	Alcalá de H.	Amphytruo (Plaut.)	G.	Es.	LT, 32, m.	203	2500	NO
1517	Alcalá de H.	Senec. Amic. Re Pub. Parad. (Cic.)	R.	Lat.	LT, 37	266	2550	B
1517	Alcalá de H.	Tragoediae (Sen.)	G.	Lat.	LT, 34	234	12240	P
1517	Logroño	Disticha, Fabulae	G.	Lat.	LT, 34	212	240	NO
1518	Alcalá de H.	Comoediae 2 (Plaut.)	R.	Lat.	LT, 38, m.	207	10600	P
1519	Valladolid	Bell. Iug., Cat. (Sall.)	G	Es.	LT, 34, m.	185	5984	B
1519	Valladolid	Metamorphoseon (Ov.)	G.	Es.	LT, 32	124	620	B
1519	Valladolid	Saturae (Iuv.)	G.	Es.	LT. 32	155	1728	B
1521	Alcalá de H.	Saturae (Pers.)	R., G.	Lat.	2 text., 43, m.	144	2744	B
		TOTAL					56470	

Anexo 2: Imágenes de las ediciones y sus transcripciones

A continuación se presentan algunos ejemplos de páginas y su transcripción diplomática correspondiente a la derecha.



Cato maior.

M.T.C.CATO MAIOR. SIVE DESENECTVTE
Incipitur fœliciter.

Tte ſi quid ego adiuto:curam ve leuaſſo:
Quæ nunc te coquit:& verſat ſub pectore fixa:
Et qua deprimeris nunc tu noctesq; dieſq:
Ecquid erit preci? Licet enim mihi verſibus iiſdē affa
ti te Attice: quibus affatur Flaminiū ille vir haud ma-
gna cum re:ſed plenus fidei. Quamquam certe ſcio:non vt Fla-
minium ſollicitari te Attice ſic noctesq; dieſq;. Noui enim mode
rationem animi tui & æquitate: teq; non cognomen ſolum athe
nis deportaſſe,ſed humanitatem atq; prudentiam intelligo: & ta
men ſuſpicor eiſdē te rebus: quibus meipſum interdum maxime
commoueri. quarum conſolatio & maior eſt: & in aliud tempus
differenda. nunc autē mihi viſum eſt vtile de ſenectute aliquid ad
te conſcribere. Hoc enim onere: quod mihi tecum eſt commune:
aut iam vrgentis aut certe aduentātis ſenectutis: & te & meipſum
leuari volo:& ſi te quidem id modice ac ſapienter ſicut omnia & fe
re & laturum eſſe certo ſcio. Sed mihi cum de ſenectute velle ali-
quid ſcribere:tu occurrebas dignus eo munere: quo vterq; nūm
vteretur communiter. Mihi quidem ita iucunda confectio huius
libri fuit:vt non omnes modo abſterſerit ſenectutis moleſtias:ſed
effecerit mollem etiam & iucundam ſenectutē. Nunquam igitur
ſatis laudari Philoſophia poterit:cui qui pareat:omne tēpus æta
tis ſuę ſine moleſtia poſſit degere. Sed de cæteris & diximus mul
ta & ſepe dicemus; nunc autē librum de ſenectute ad te miſimus.
Omnem autem ſermonē tribuimus non Titono:vt Ariſto chi?:
ne parum eſſet auctoritatis in fabula:ſed.M.Catoni ſeni:quo ma
iorem auctoritatē haberet oratio. Apud quem Lelium & Scipio-
nem fecimus admirantes:q; is tam facile ſenectutē ferat:iiſq; eum
reſpondentē. Qui ſi eruditius videbitur diſputare: quam conſue-
uit ipſe in ſuis libris:attribuito litteris grecis: quarum conſtat eū
perſtudioſum fuiſſe in ſenectute. Sed quid opus eſt plura? iam eiſ
ipſius Catonis ſermo explicabit nram oēm de ſenectute ſentētiā.

P. Scio.C.Lelius. Cato ſenior.

Saepenumero ad mirari ſoleo cum hoc.C.Lelio cum cætera
rum tuam excellentem.M. Cato perfectamq; ſapientiam:tū
vel maxime q; nunquam tibi ſenectutem grauem eſſe ſenſerim:

Cato malor.

M.T. CATO MAIOR SIVE DE SENECTVTE

Incipitur fOeliciter.

Tite ſi quid ego adiuto: curam ve leuaſſo:

Quae nunc te coquit: & verſat ſub pectore fixa:

Et qua deprimeris nunc tu noceſq: die ſq3s

Ecquid erit preci. eēet enim mihi rerſibus iſcē aff

ti te Atice: qubus affatur Faminiū ile it haud ma-

gn eum re: ſed plenus fide. Cuamquam certe ſcio: non t a-

minium ſolicitari te Attice ſiq gocteſq3 dieſq3s. Noui enim

mode

rationem animi tui & aqualitate: teq3s non cognomen ſolum

athe

nis deportaſſeſed humanitatem atq3s prudentilam intelligo: &

ta

men ſuſpicor eiſē te rebus qubus meipſum interdum mexime

eommoueri. quarum conſolatio & meior eſt: & in aliud tempus

differenda tunc autē mſhi ſum eſt rtile de ſenectute aliquid ad

te conſcribere. Hoc eriim onere: quod mihi tecum eſt

commume:

aut iam rgentis aut certe aduentaātis ſenectutis: & te &

mcipſum

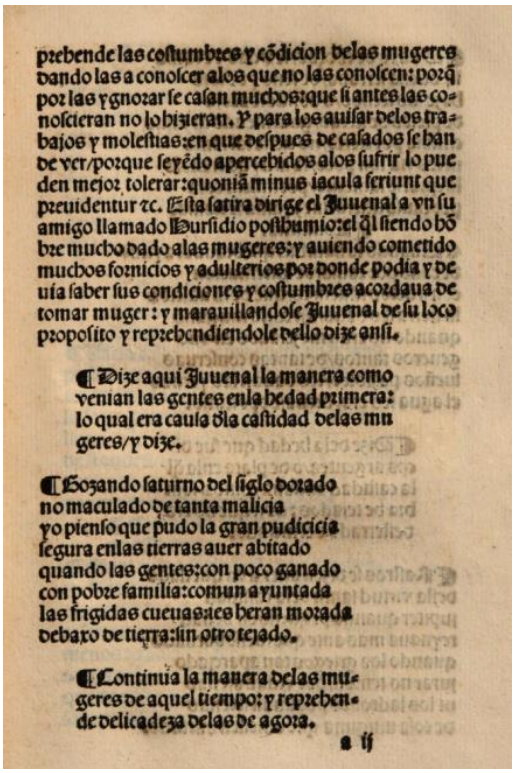
lerar' olo: & ſi tae quidem id modee ac fapiente ficut ommia &

fe

re & laturum eſſe certo ſcio. Sed mibi cum de ſenectute ee al-

quid ſcrībere: tu occurrebas dignus eo munere: quo terq3 nrtm

Ejemplo 1. Folio a1v del ejemplar BH FLL 18902(2), Universidad Complutense de Madrid, Biblioteca Histórica Marqués de Valdecilla, CECLE0266



prehende las coſtumbres y cõdicion delas mugeres
dando las a conoſcer alos que no las conoſcen: porq̃
por las ygnorarſe caſan muchos: que ſi antes las co-
noſcieran no lo hizieran. Y para los auiſar delos tra-
bajos y moleſtias: en que deſpues de caſados ſe han
de ver / porque ſeyendo apercebidos a los ſufrir lo pue-
den mejor tolerar: quoniã minus iacula feriunt que
preuidentur &c. Eſta ſatira dirige el Iuuenal a vn ſu
amigo llamado Hurſidio poſthumio: el q̃l ſiendo hõ
bre mucho dado a las mugeres: y auiendo cometido
muchos fornicios y adulterios por donde podia y de
uia ſaber ſus condiciones y coſtumbres acordaua de
tomar muger: y marauillandoſe Iuuenal de ſu loco
propoſito y reprehendiendole dello dize anſi.

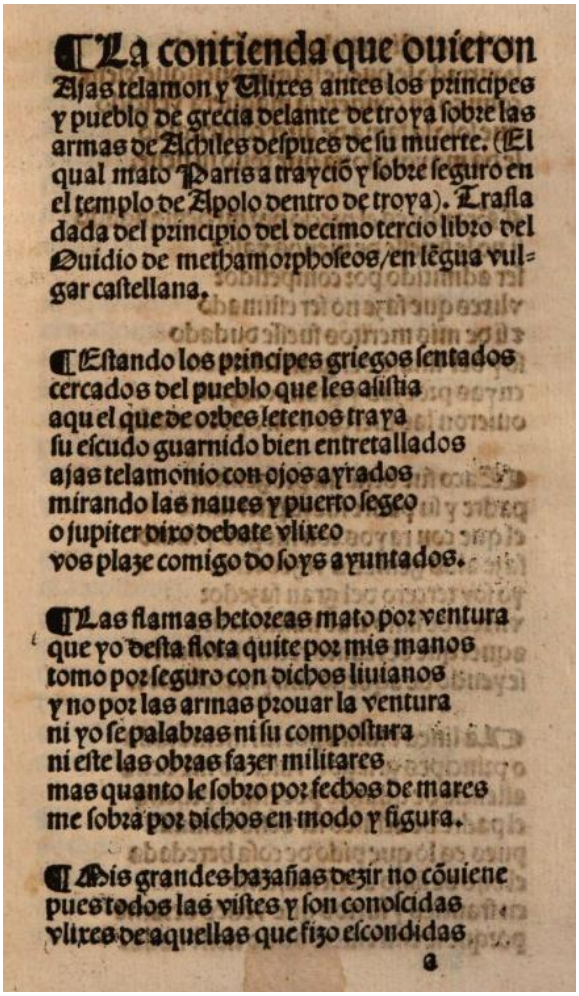
¶ Dize aqui Iuuenal la manera como
venian las gentes en la hedad primera:
lo qual era cauſa de la caſtidad de las mn
geres / y dize.

¶ Gozando ſaturno del ſiglo dorado
no maculado de tanta malicia
yo pienſo que pudo la gran pudicicia
ſegura en las tierras auer abitado
quando las gentes: con poco ganado
con pobre familia: comun ayuntada
las frigidas cueuas: les heran morada
debaxo de tierra: ſin otro tejado.

¶ Continua la manera delas mu-
geres de aquel tiempo: y reprehen-
de delicadeza de las de agora.

a ij

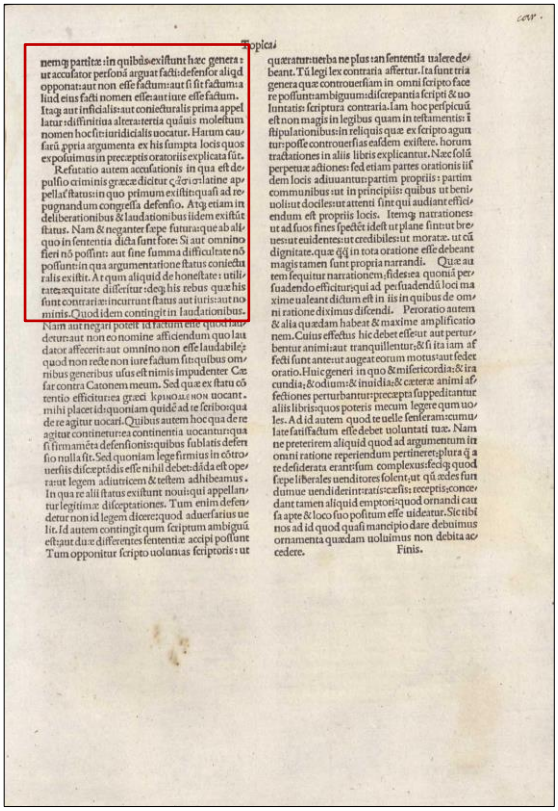
Ejemplo 2. Folio a2r del ejemplar BE.8.S.76 PS, Österreichische Nationalbibliothek, CECLE0155



¶ La contienda que ouieron
Ajas telamon y Vlixes antes los pꝛincipes
y pueblo de grecia delante de troya ſobre las

armas de Achiles deſpues de ſu muerte. (El
qual mato Paris a trayciõ y ſobre ſeguro en
el templo de Apolo dentro de troya). Traſla
dada del principio del decimo tercio libro del
Ouidio de methamoꝛphoſeos / en lẽgua vul=
gar caſtellana.
¶ Eſtando los principes griegos ſentados
cercados del pueblo que les aſiſtia
aquel que de oꝛbes ſetenos traya
su eſcudo guarnido bien entretallados
ajas telamonio con ojos ayrados
mirando las naues y puerto ſegeo
o jupiter dixo debate vlixeo
vos plaze comigo do ſoys ayuntados.
¶ Las flamas hetoꝛeas mato poꝛ ventura
que yo deſta flota quite poꝛ mis manos
tomo poꝛ ſeguro con dichos liuianos
y no poꝛ las armas pꝛouar la ventura
ni yo ſe palabꝛas ni su compoſtura
ni eſte las obꝛas fazer militares
mas quanto le ſobꝛo poꝛ fechos de mares
me ſobꝛa poꝛ dichos en modo y figura.
¶ Mis grandes hazañas dezir no cõuiene
pues todos las viſtes y ſon conoſcidas
vlixes de aquellas que fizo eſcondidas
a

Ejemplo 3. Folio a1r del ejemplar BE.8.S.76 Alt-Punk.,
Österreichische Nationalbibliothek, CECLE0124



Topicas

nemqȝ partitae: in quibūs. exiſtunt haec genera: ut accuſator perſonā arguat facti: defenſor aliqd opponat: aut non eſſe factum: aut ſi ſit factum: a liud eius facti nomen eſſe: aut iure eſſe factum Itaqȝ aut inficialis: aut coniecturalis prima appel latur: diffinitiua altera: tertia quāuis moleſtum nomen hoc ſit: iuridicialis uocatur. Harum cau ſarū ꝓpria argumenta ex his ſumpta locis quos expoſuimus in precaeptis oratoriis explicata ſūt. Refutatio autem accuſationis in qua et de- pulſio criminis graecae dicitur cēie: latine ap- pellat' ſtatus: in quo primum exiſtit: quaſi ad re- pugnandum congreſſa defenſio. Artqȝ etiam in deliberationibus & laudationibus iidem exiſtūt ſtatus. Nam & neganter ſaepe futura: que ab ali- quo in ſententia dicta ſunt fore: Si aut omnino fieri nō poſſint: aut ſine ſumma difficultate nō poſſunt: in qua argumentatione ſtatus coniectu ralis exiſtit. At qum aliquid de honeſtate: utili- tate: aequitate diſſeritur: deqȝ his rebus quae his ſunt contrariae: incurrunt ſtatus aut iuris: aut no minis. Quod idem contingit in laudationibus.

Ejemplo 4. Folio a5v del ejemplar del Seminario de Santa Catalina de Mondoñedo, e78-135(2) CECLE0245

Bibliografía citada

Ayuso García, Manuel, «OCR of a mixed corpus: early printings and manuscripts of Martianus Capella's work», *DATeCH2017* (Göttingen, Germany, 2017), ACM, 2017, pp. 77-82.

—, «Las primeras ediciones hispanas de Persio. Aproximación a su estudio empleando OCR y otras herramientas de reconocimiento automático», en *La edición de los clásicos latinos en el Renacimiento: textos, contextos y herencia cultural. Los textos clásicos en los inicios de la tradición impresa*, ed. A. Moreno Hernández, Madrid, Ediciones Complutense de Madrid, 2021, pp. 163-181.

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.

Kahle, Philip, Sebastian Colutto, Gunter Häckl, Gunter Mühlberger, «Transkribus - a Service Platform for Transcription,Recognition and Retrieval of Historical Documents», en *14th IAPR International Conference on Document Analysis and Recognition*, 2017, pp. 19-24.

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).

Reul, Christian, Christoph Wick, Uwe Springmann, Frank Puppe, «Transfer Learning for OCRopus Model Training on Early Printed Books», en: *Zeitschrift für Bibliothekskultur / Journal for Library Culture* 5.1 (2017), pp. 38–51.

Reul, Christian, Marco Dittrich, Martin Gruner, «Case Study of a Highly Automated Layout Analysis and OCR of an Incunabulum: *Der Heiligen Leben* (1488)», en *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage* (DATeCH 2017, Göttingen), ACM, 2017, pp. 155–160.

Reul, Christian, Uwe Springmann, Christoph Wick, Frank Puppe, «Improving OCR Accuracy on Early Printed Books by Utilizing Cross Fold Training and Voting», en *13th IAPR International Workshop on Document Analysis Systems* (DAS 2018, Vienna, Austria, April 24-27), 2018, pp. 423–428. DOI: < https://doi.org/10.1109/DAS.2018.30>.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, Frank Puppe, «OCR4all - An open-source tool providing a (semi-) automatic OCR workflow for historical printings», *Applied Sciences*, 9/22 (2019).

Springmann, Uwe, Dietmar Najock, Hermann Morgenroth, Helmut Schmid, Annette Gotscharek, Florian Fink, «OCR of historical printings of Latin texts: problems, prospects, progress», en *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (DATeCH 2014, Madrid, Spain), ACM, 2014, pp. 57–61.

Springmann, Uwe y Florian Fink, *CIS OCR Workshop v1.0: OCR and postcorrection of early printings for digital humanities*, 2016. DOI: <https://doi.org/10.5281/zenodo.46571>.

Springmann, Uwe, Florian Fink, Klaus U. Schulz, «Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings», en *ArXiv e-prints*, 2016, s.p. URL: <http://arxiv.org/abs/1606.05157> (cons. 18/05/2022).

Springmann, Uwe y Anke Lüdeling, «OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus», en *Digital Humanities Quarterly*, 11/2 (2017).

Springmann, Uwe, Christian Reul, Stefanie Dipper, Johannes Baiter, *GT4HistOCR: Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin*, 2018. DOI: <https://doi.org/10.5281/zenodo.1344131> (cons. 18/05/2022).

Reul, Christian, Uwe Springmann, Christoph Wick, Frank Puppe, «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. URL: <https://arxiv.org/ftp/arxiv/papers/1810/1810.03436.pdf> (cons.18/05/2022).

Springmann, Uwe, Florian Fink, Klaus U. Schulz, «Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical», *ArXiv e-prints*, 2016. URL: <https://arxiv.org/abs/1606.05157> (cons. 18/05/2022).

Enlaces a softwares citados (cons. 10/05/2022)

Calamari	<https://github.com/Calamari-OCR/calamari>
OCRopus	<https://github.com/ocropus>
OCR4all	<https://www.OCR4all.org/>
LAREX	<https://github.com/OCR4all/LAREX>
Scantailor	<https://scantailor.org/>
Transkribus	<https://readcoop.eu/transkribus/>