Revista de Humanidades Digitales

Inteligencia artificial para la transcripción de letra itálica española del siglo XVIII: Transkribus como herramienta para las Humanidades Digitales

Artificial Intelligence for the Transcription of XVIIIth Century Spanish Italics: Transkribus as a Tool for Digital Humanities

Jaime BERMÚDEZ CARREÑO
Universidad Industrial de Santander
jaime2211010@correo.uis.edu.co
https://orcid.org/0009-0001-9139-0221

RESUMEN

El propósito de este artículo es desarrollar un modelo de inteligencia artificial para la lectura y transcripción de la escritura itálica manuscrita producida en el siglo XVIII español utilizando la plataforma Transkribus. Esta investigación hace un recorrido historiográfico de los esfuerzos de investigadores para crear modelos de lectura de textos (HTR) con el objetivo de determinar los retos y problemas encontrados, así como la metodología y las buenas prácticas a emplear en aras de obtener un modelo de lectura preciso. Finalmente se realiza una presentación metodológica del desarrollo del modelo y se presentan las conclusiones obtenidas.

PALABRAS CLAVE

Inteligencia artificial, Transkribus, itálica, modelos HTR, transcripción de textos antiguos.

ABSTRACT

The purpose of this article is to develop an artificial intelligence model to read and transcribe italics handwriting produced in the Spanish XVIIIth century using Transkribus as platform. This research makes an historiographic journey of the efforts of other researchers to create handwritten text models (HTR) with the objective of determining the challenges and problems encountered, as well as the methodology and good practices to be used in order to obtain an accurate reading model. Finally, a methodological presentation of the development of the model is made and the conclusions obtained are presented .

KEYWORDS

Artificial Intelligence, Transkribus, Italics, HTR Models, Ancient Texts Transcription.

ISSN
2531-1786

Revista de Humanidades Digitales http://revistas.uned.es/index.php/RHD/index
Recibido 07/08/2023 – Aceptado 23/10/2023
Editado por G. del Rio Riande

1.INTRODUCCIÓN

La labor del escribano es un ejercicio que requiere paciencia y disciplina. Se escribe para registrar momentos, sean económicos, políticos, sociales o de índole artística. Sea el diario de viaje de un navío o la crónica de acontecimientos en guerra, el escribano ha puesto pluma sobre papel para registrar momentos desde mucho antes de la imprenta, e incluso después de ella. Cada escritura es diferente, atada a quien escribe, cada trazo es único y cada papel antiguo que leamos hoy puede estar afectado por los rigores del tiempo, degradando la tinta y dificultando su lectura.

Estas dificultades de lectura pueden ser experimentadas por cualquier investigador y es por esta razón, entre muchas otras, que se han buscado alternativas que faciliten la lectura de los textos, entendiendo esta facilidad no como el dominio de la paleografía como disciplina, sino en el uso de herramientas y nuevos medios que faciliten y, en muchos casos, agilicen la labor de lectura y transcripción de textos para su uso en investigación.

Uno de los grandes avances en el desarrollo de herramientas para la lectura de textos tiene que ver con la microfilmación[1], técnica que permite fotografiar y luego traspasar una fotografía de un texto a un pequeño rollo de celulosa que después puede consultarse en un dispositivo específico de lectura. El microfilm permitió la lectura de documentos enteros sin manipular físicamente el papel, disminuyendo la posibilidad de deterioro del original; además, otorgó la posibilidad de hacer copias de un documento, permitiendo la exploración colaborativa del mismo, acelerando así su transcripción y uso. Pero el microfilm tiene sus limitantes, entre ellos la rápida degradación de la celulosa y la necesidad de contar con dispositivos especiales para su consulta.

Entre los años sesenta y ochenta del siglo veinte surgieron los computadores personales como son conocidos en la actualidad[2], dotados de un monitor (con interfaz gráfica), teclado y *mouse*. Se trataba de equipos pequeños pero potentes con capacidades de procesamiento de datos nunca antes vistas. Con el nacimiento del PC nació también la era digital, dando paso a la visualización multimedia (audio, foto y video) a través de los equipos de cómputo. Las instituciones de la memoria harían entonces grandes esfuerzos por capturar imágenes digitales de sus documentos (Bazzaco et al., 2022, p. 73) para que pudiesen ser consultados por medio de computadores de escritorio; así mismo, la masificación del internet como medio para compartir información introdujo la consulta virtual de archivo (Baucom, 2019, p. 4). En un espacio relativamente corto de tiempo el investigador tuvo herramientas poderosas y de fácil acceso que permitieron la consulta rápida de diversos archivos y documentos, en la mayoría de los casos, ubicados físicamente a miles de kilómetros de distancia, pero presentes gracias a estos nuevos medios digitales.

---

[1] Archivo General de la Nación. Guía de digitalización a partir de soportes en microfilm. Accesible desde: https://www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Consulte/Recursos/Publicaciones/GuiaDigitalizacionMicrofilm.pdf.

[2] Sobre la historia de la computadora, véase: https://humanidades.com/historia-de-la-computadora/.

Podría decirse que el problema de acceso a la información estaba casi resuelto, la consulta digital era ágil y económica, ahora quedaba el problema de la lectura. Versiones antiguas en diferentes idiomas, tipos de escritura a mano alzada variados como la cortesana, la procesal y la procesal encadenada o el deterioro propio del documento son retos que se presentan a quien desee hacer uso de esta información. La única manera de interpretar un documento antiguo es haciendo uso de la paleografía, disciplina que permite leer y transcribir correctamente los documentos manuscritos (Silva Prada, 2001, p. 17). Dicha disciplina requiere el uso de las mejores habilidades de los investigadores y sobre todo requiere tiempo, un tiempo extenso para interpretar y clasificar cada folio, transcribiendo su contenido a un idioma y estilo de escritura contemporáneo. El método de trabajo empleado en la transcripción de documentos se desarrolló de manera estandarizada durante la mayor parte de la historiografía moderna, pero no fue sino hasta la adopción masiva de los sistemas digitales de reconocimiento de texto impreso (OCR) y posteriormente manuscrito (HTR) que dicho proceso se hizo más ágil, usando la computación como ayuda (Bazzaco et al., 2022, p. 75).

Los sistemas de reconocimiento de texto digital HTR (Handwritten Text Recognition) y OCR (Optical Character Recognition) ya dominan el reconocimiento de letra imprenta contemporánea (Bazzaco, 2020, p. 544), aunque podría argumentarse que versiones tempranas de los sistemas OCR existen desde los años cuarenta. No obstante, no fue sino entre los años ochenta del siglo veinte y los inicios del siglo veintiuno que se desarrollaron los primeros softwares de reconocimiento de caracteres (Memon et al., 2020, p. 142642). Estos se encuentran presentes en buscadores de texto como Google y Bing, también se encuentran presentes en herramientas como la suite de Adobe para leer textos impresos que han sido digitalizados. Estos sistemas funcionan, a grandes rasgos, reconociendo las formas de cada píxel digitalizado y comparándolos con una base digital inmensa para identificar letra a letra, que luego se transcribe a una hoja de texto o se indexa en un sistema de búsqueda (Nockels et al., 2022, p. 3).

Aunque estos sistemas de reconocimiento de caracteres han llegado a dominar el reconocimiento de texto impreso, se les dificulta enormemente la letra manuscrita, por razones que tienen que ver con la escritura única de cada persona, los defectos del escaneo y los rigores del tiempo, que afectan cada texto de una manera distinta. Cada letra y escritura, tan diferente una de otra, presentan la dificultad de tener un sistema que permita compararlas, se hace necesario un sistema que pueda aprender de la información que se le da, un sistema que pueda ser entrenado agregándole información y luego pudiese aprender de dicha información para interpretar los documentos que se le entreguen y transcribirlos posteriormente a un procesador de texto o indexarlos en un buscador, la existencia de dicho sistema solo es posible a través de la inteligencia artificial (IA) (Muehlberger et al., 2019).

Los esfuerzos por crear máquinas que puedan aprender de sí mismas existen desde mediados del siglo veinte (Nilsson, 2010, p. 71), pero no es hasta la segunda década del siglo veintiuno cuando se crean los primeros sistemas de IA verdadera, los cuales funcionan como un

sistema de redes neuronales alimentadas con una gran cantidad de información. Muehlberger et al. (2019) definen a las redes neuronales de la siguiente manera:

> Las redes neuronales son sistemas computacionales de hardware y software que están vagamente modeladas en las redes biológicas encontradas en los cerebros de los animales. Ellas aprenden y mejoran su desempeño al ser entrenadas con una serie de ejemplos. Las redes neuronales profundas son una clase de algoritmo de aprendizaje de máquinas que utiliza múltiples capas de procesamiento para aprender y analizar una tarea (p. 969).

Tegmark (2018) llama a estas IA basadas en redes neuronales *good old-fashioned AI* (p. 110), son sistemas que requieren ser alimentados para que puedan interpretar la información que se les da y emitir un resultado basado en el aprendizaje, siendo las redes neuronales el medio de procesar la información y tomar decisiones. Estos tipos de IA tienen la capacidad de procesar texto (ChatGPT, Google Bard)[3], imágenes (Midjourney)[4], e incluso ha llegado a fabricar pequeños fragmentos de video usando instrucciones dadas en texto[5].

La herramienta llamada *Transkribus* (Muehlberger et al., 2019, p. 957) mezcla la IA con los sistemas OCR/HTR, siendo un software que utiliza la IA para reconocer diversos tipos de textos, sean impresos o manuscritos. Su principal característica es que para poder reconocer un texto primero debe alimentarse a la IA con información preliminar sobre este, de esta manera pueden crearse modelos de lectura automática que después pueden usarse con tipos de letra similar.

El propósito de este artículo es el de entrenar un modelo de lectura automática de textos usando la IA de Transkribus para letra manuscrita itálica del siglo XVIII español. La finalidad de la transcripción automática de este tipo de letra radica en la falta de disponibilidad a la fecha de un modelo de manuscrita española del siglo XVIII[6].

Este artículo hará un recorrido historiográfico de los esfuerzos de otros investigadores para crear modelos de lectura de textos con el objetivo de determinar los retos y problemas encontrados, la metodología y las buenas prácticas a emplear en aras de obtener un modelo de lectura preciso. Como fuente primaria y texto base para el entrenamiento se ha tomado un proceso de fe de finales del siglo XVIII realizado al fray Joseph Cavadas[7]. Este proceso de fe en particular tiene una escritura itálica diáfana y correctamente digitalizada, siendo factible su uso para entrenar un modelo de lectura IA.

El resultado de esta investigación pretende contribuir al uso y masificación de este tipo de herramientas digitales para la paleografía, facilitar el acceso y lectura de documentos, disminuyendo el tiempo que toma transcribirlos para su uso.

2. LA TRANSCRIPCIÓN DE DOCUMENTOS UTILIZANDO INTELIGENCIA ARTIFICIAL

El uso de la IA para la transcripción de textos antiguos es una combinación entre el futuro y el pasado que podemos materializar en el presente; un claro ejemplo de lo que el campo de las

[3] OpenAi. Accesible desde: https://openai.com/blog/chatgpt/.
[4] Sobre Midjourney, véase: https://www.sciencefocus.com/future-technology/midjourney/.
[5] Sobre Unite, véase: https://www.unite.ai/best-ai-video-generators/.
[6] Sobre los modelos públicos de Transkribus, véase: https://readcoop.eu/transkribus/public-models/.
[7] Puede consultarse este texto desde el Portal de Archivos Españoles. *Proceso de fe de fray José Mariano Cavadas*. http://pares.mcu.es/ParesBusquedas20/catalogo/description/1312347?nm.

Humanidades Digitales puede ofrecer.

Enero de 2016 vio el nacimiento del proyecto READ (Recognition and Enrichment of Archival Documents), enfocado en hacer que el material de archivo sea más accesible a través del uso de herramientas tecnológicas. Dicho proyecto se enfocó en ofrecer una plataforma que permitiese el reconocimiento automático, la transcripción y la búsqueda e indexación de textos antiguos. Esta plataforma, llamada Transkribus, nació a finales del 2015 como inicio del proyecto READ y es producto de una colaboración entre la Universidad de Innsbruck en Austria y la Universidad de Valencia en España[9]. Ese mismo año, investigadores de diferentes instituciones europeas empezaron a hacer uso del software, publicando artículos de investigación que con sus hallazgos que permiten obtener un estado del arte completo y una aproximación a lo que podríamos considerar como buenas prácticas de transcripción utilizando IA.

Uno de los artículos más tempranos que mencionan el uso de Transkribus proviene de Tiziana Mancinelli y fue publicado en diciembre de 2016. Allí se presenta una discusión relevante sobre el estado del arte en la lectura de textos digitalizados, alegando que Transkribus, al ser comparado con otros sistemas experimentales de lectura OCR, presenta un menor margen de error al reconocer textos escritos a mano (Mancinelli, 2016, p. 259). Así mismo, Mancinelli pone en discusión el estado de la digitalización diciendo que "no es lo mismo digitalizar un documento que obtener una edición digital de él" (Mancinelli, 2016, p. 256), en el sentido que la mera captura digital de un texto no proporciona un camino directo al uso digital de este, como, por ejemplo, son las actividades de búsqueda e indexación.

Bajo esta óptica, Transkribus se presenta como una alternativa viable para convertir de manera eficiente documentos digitalizados en ediciones digitales. En 2017, Philip Kahle y Guenter Muehlberg, fundadores del proyecto READ, dieron una conferencia enfocada en presentar Transkribus fuera de la comunidad académica ligada al proyecto (Kahle et al., 2017). Dicha conferencia serviría a manera de introducción de Transkribus como software y permitiría que académicos fuera del proyecto READ se familiarizaran con el uso de la herramienta.

2.1. Buenas prácticas de transcripción, dificultades y soluciones

Uno de los proyectos de relevancia académica realizados con Transkribus fuera del proyecto READ es el llamado *The Foucault fiches de lecture* (FFL), enfocado en organizar, digitalizar, transcribir y publicar online toda la biblioteca de notas producida por Michel Foucault a lo largo de su vida académica (Massot et al., 2019, p. 11). El proyecto, que inició en 2013 usando transcripción manual de los documentos se vio fuertemente beneficiado con la aplicación de modelos de inteligencia artificial proporcionados por el entrenamiento en Transkribus. Uno de los aspectos más importantes que podemos obtener de este proyecto de investigación son las dificultades encontradas para su elaboración y como se superaron. La primera de ellas tiene que ver con las abreviaturas. De acuerdo con los autores, "varias abreviaturas son equívocas, lo que

[8] Sobre READ, véase: https://readcoop.eu/the-read-project-has-started/.
[9] Sobre esta colaboración, véase: https://readcoop.eu/a-short-history-of-transkribus-with-gunter-muhlberger/.

puede significar diferentes palabras en diferentes contextos" (Massot et al., 2019, p. 12). Es decir, las abreviaturas de palabras usadas por Foucault podrían significar una cosa en sus cartas, pero algo diferente en sus notas. Para solucionar este inconveniente se consideró establecer un diccionario de abreviaturas que contuviese capturas de imagen de las abreviaturas encontradas en sus textos junto a los posibles significados.

La segunda dificultad encontrada tiene que ver con la capacidad de transcripción de los investigadores al entrenar el modelo de Transkribus. De acuerdo con los autores, cargar y transcribir un folio de dos imágenes toma "entre treinta y cuarenta minutos" (Massot et al., 2019), sin considerar el tiempo que toma descifrar abreviaturas, nombres propios o palabras borrosas. Si bien este tipo de dificultades no son propias de Transkribus, se muestran como comunes ante cualquier proyecto de transcripción.

Para finales de 2018, Muehlberger y otros autores publicarían el primer artículo con la descripción completa del proyecto READ y el funcionamiento interno de Transkribus. Este articulo sugiere la implementación de un modelo colaborativo de entrenamiento de modelos HTR para asegurar la escalabilidad del proyecto (Muehlberger et al., 2019, p. 955). Muestra además información sobre el funcionamiento interno del sistema de IA, el cual funciona a partir de redes neuronales que aprenden constantemente de entrenamientos previos, lo que se ha descrito en este trabajo como *good old-fashioned AI*.

Durante este mismo periodo se publicaron investigaciones relacionadas con crear software que pudiese implementar reconocimiento de texto mediante redes neuronales de manera offline, es decir, sin que el archivo digital se encuentre conectado a internet (Tran et al., 2019, p. 52), además de otros proyectos enfocados en el desarrollo de modelos de transcripción por inteligencia artificial de código abierto. A la fecha de escritura de este artículo, Transkribus solo puede utilizarse bajo una conexión a internet y sus nuevos modelos algorítmicos de IA son propiedad de la plataforma (Reul et al., 2022, p. 416).

Uno de los principales problemas a los que un sistema de reconocimiento de texto se enfrenta es la correcta digitalización de los documentos. A la fecha de escritura de este artículo no existe un acuerdo por la comunidad académica sobre los parámetros correctos de digitalización. Bazzaco lo describe como "*clean transcription* y *dirty* OCR" (Bazzaco, 2020, p. 544), es decir que problemas comunes como el escaneo defectuoso de las fuentes, problemas particulares de una obra impresa y un mal procedimiento de digitalización dificultan enormemente el reconocimiento automático de texto. Este autor menciona además que los problemas comunes al reconocer un texto de manera automática tienen que ver con que el texto se reproduce digitalmente manteniendo de forma intacta su materialidad, es decir, en la digitalización se pueden observar las manchas en las páginas, las deformaciones del escaneo manual y las transferencias de la tinta a vuelta de página (Bazzaco, 2020, p. 545). Por otro lado, se presentan retos particulares en cada texto a reconocer propios de la escritura del texto. En el caso del propuesto por Bazzaco, que es letra gótica del Siglo de Oro, características como la variabilidad gráfica de las letras, abreviaturas, signos tironianos, ligaduras entre caracteres, desgaste de los tipos y defectos de se verán reflejadas en

la digitalización, siendo responsabilidad del investigador a cargo de la transcripción determinar la viabilidad de su lectura sin afectar el modelo HTR. También deberá considerar si algún folio determinado debe intervenirse digitalmente para facilitar su lectura, cosa que aplicaciones como las del Archivo General de Indias ya permiten (aumentar contraste, disminuir brillo, cambios monocromáticos, etc.) para facilitar la lectura del texto.

La constitución del proyecto READ en 2019 como una cooperativa que abarca una mezcla de instituciones, archivos y personas privadas otorgó la posibilidad de expandir y autogestionar el uso de Transkribus a través de la comunidad académica. Este hecho permitió la expansión del uso del software y la apertura del sistema de Transkribus al público en general gracias al desarrollo de una aplicación online a través del navegador web[10]. Producto de esta etapa encontramos proyectos de transcripción como el de los diarios de Eugène Wilhelm, donde su autor sugiere que una de las principales dificultades encontradas en la transcripción de documentos que abarcan varias décadas (los diarios Wilhelm van de 1885 a 1951) son las dificultades de lectura del texto, que se va deformando cada vez más a medida que la persona envejece (Schlagdenhauffen, 2020, p. 10), un reto similar al descrito en las notas de Foucault (Massot et al., 2019). Adicionalmente, existe la necesidad de comprobar que la transcripción realizada por la IA sea en efecto correcta, y esto puede probar ser un reto para el ojo humano no entrenado (Schlagdenhauffen, 2020, p. 10). Otra dificultad descrita por el autor es la falta de conocimiento de los transcriptores en los contextos históricos, políticos y sociales propios del documento original, proponiendo la supervisión del proceso de transcripción por un grupo de expertos (Schlagdenhauffen, 2020, p. 10). Así mismo, se cuestiona la necesidad de crear un diccionario para las abreviaturas y términos, confirmando los planteamientos del proyecto Foucault.

Dadas estas circunstancias, es importante saber que un proceso de transcripción automática nunca podrá ser perfecto, sin embargo, utilizar criterios fijos de transcripción tales como la conservación de los signos de interpunción en la transcripción, los acentos y el desarrollo o mantenimiento de las abreviaturas según la necesidad, permiten mantener un margen de error bajo en el modelo de entrenamiento de Transkribus (Bazzaco et al., 2022, p. 94). Otros autores como Alemayehu dan el nombre de "buenas prácticas de transcripción" al hecho de establecer un protocolo riguroso de trabajo previo al proceso de transcripción manual de documentos (2022, p. 4).

De acuerdo con la documentación de Transkribus, un margen de error inferior al 10% puede considerarse como bueno[11] (conocido como *character error rate* o CER), por ende, actividades tales como crear un protocolo de digitalización de documentos, establecer criterios previos de transcripción y crear un diccionario de abreviaturas son factores determinantes en un proceso de transcripción automática de documentos. Sin embargo, dichos procesos no reemplazan la labor de un editor o director de proyecto en verificar que los resultados obtenidos vayan

[10] Más sobre Readcoop en: https://readcoop.eu/our-story.
[11] Sobre el margen de error en Transkribus, véase: https://help.transkribus.org/character-error-rate-and-learning-curve.

acorde al documento original y en organizar el equipo de trabajo para que logren obtener resultados consistentes de manera conjunta.

Dentro de este trabajo editorial, Deslandres et al. destacan la importancia de crear un comité de transcripción y tomarse el tiempo de planificar el trabajo inicial, pues un protocolo de transcripción riguroso se refleja directamente en el margen de error de los modelos (2022, p. 9). Además, este protocolo de transcripción puede ayudar a circunvenir errores humanos como los diferentes criterios de transcripción que un equipo diverso de paleógrafos pudiese llegar a tener; factores como implementar reglas de transcripción para las abreviaturas y homogeneizar los términos al transcribirlos han probado ser útiles en disminuir el margen de error de las transcripciones (Deslandres et al., 2022, pp. 7-8). Para ayudar en este tema, READ ha planteado unas convenciones de transcripción[12] que sirven como protocolo base para los trabajos de transcripción.

Utilizando estas referencias de investigaciones previas, se ha planteado la transcripción automática de un documento escrito en letra itálica española del siglo XVIII. A continuación, se elaborará el proceso de transcripción digital junto a los resultados obtenidos bajo el concepto de buenas prácticas de transcripción.

3. EL MODELO CAVADAS 1762

Existen diferentes aproximaciones a entrenar un modelo de IA en Transkribus, diversos autores han planteado métodos de transcripción distintos que se adaptan a las necesidades del resultado deseado que se pretende obtener (Alemayehu, 2022; Bazzaco et al., 2022; Deslandres et al., 2022). Sin embargo, pueden tomarse las instrucciones de uso, junto a la metodología de transcripción y reconocimiento presentada por Transkribus, como base para realizar el proceso[13].

3.1. Metodología y origen de los datos

Hemos tomado como referencia los trabajos de Massot et al. (2019) y Milioni (2020) para estructurar la metodología de presentación de resultados. Dichos resultados están estructurados por origen de datos, criterios de transcripción, entrenamiento del modelo, dificultades encontradas y resultados. Como documento base para el entrenamiento del modelo hemos tomado un proceso de fe contra Joseph Mariano Cavadas realizado en 1762 (Archivo General de Indias, Inquisición,1730, Exp.35), por ende, los motivos para escoger dicho documento para entrenar un modelo de escritura itálica española son los siguientes:

1. El documento se encuentra correctamente digitalizado y tiene buena calidad de captura de imagen, la letra es legible a lo largo de todo el documento.
2. La letra manuscrita es clara y se encuentra escrita de manera consistente a lo largo del

[12] Sobre las convenciones de transcripción en Transkribus, véase: https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/.
[13] Un breve tutorial de Transkribus en: https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/.

documento, la tinta no ha perdido su intensidad ni la mano del escribano perdió su estilo ni deformó el texto a lo largo del documento.

3. El estilo de escritura itálica manuscrita es consistente y representativa con la realizada durante el siglo XVIII, podría decirse que el escribano tiene un estilo limpio de escritura.

Como paso preliminar se realizó una limpieza de información innecesaria del pdf antes de cargarlo a la plataforma de Transkribus, pues el texto adicional que no se va a transcribir podría afectar el reconocimiento automático. Se cortó la información de la portada superior y de la barra inferior o *footer* utilizando una aplicación web gratuita llamada Sejda[14] (Figura 1):

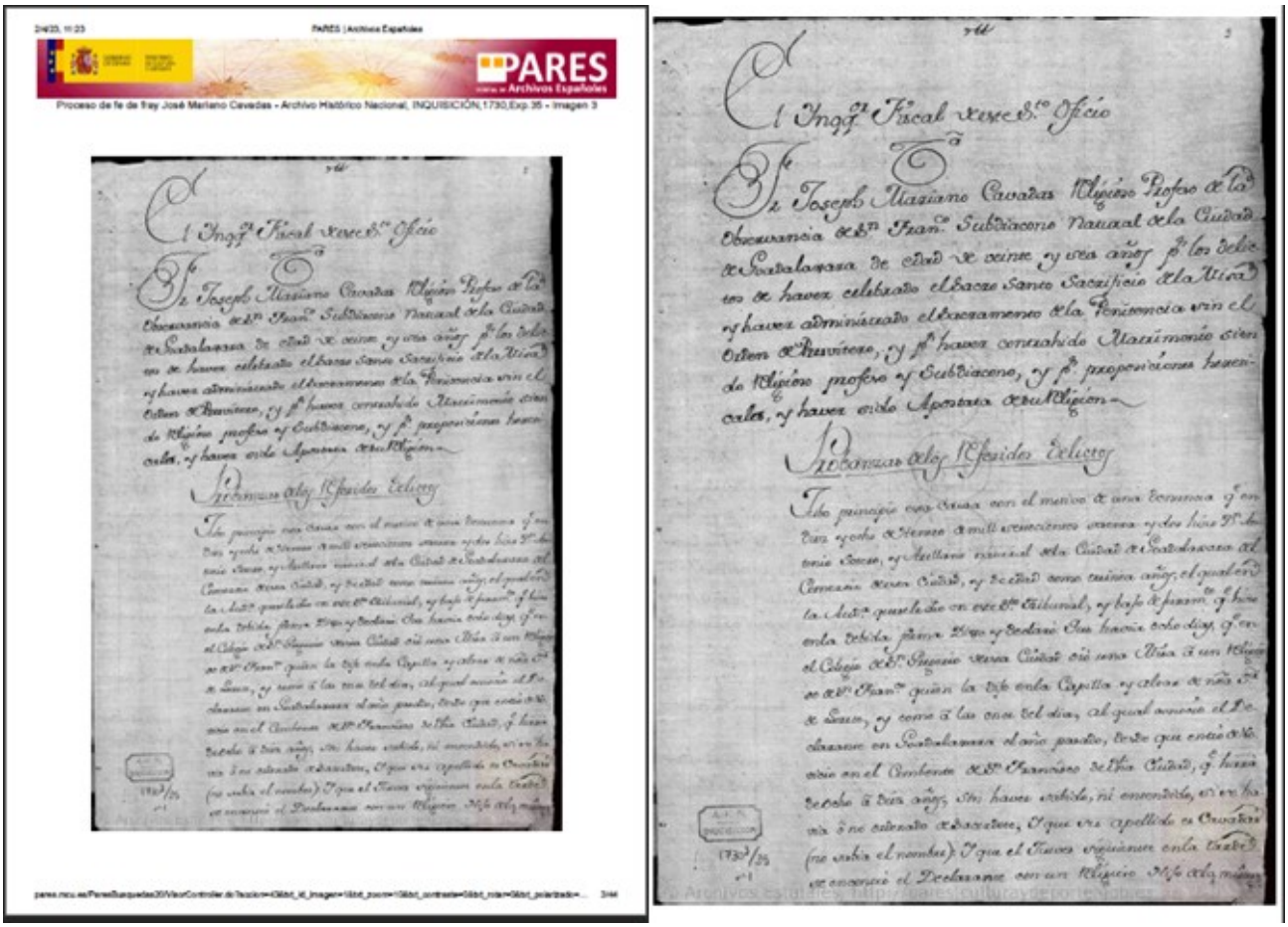

Figura 1. Remoción de áreas de texto no sujetas de transcripción. Fuente: elaboración propia.

Una vez cargado el documento en la biblioteca de Transkribus, se procede a identificar la región de transcripción o *layout*, como indica la aplicación (Figura 2). Existen diferentes maneras de identificar el *layout*, una de ellas es hacerlo de manera automática gracias al sistema de reconocimiento de texto de Transkribus, que probó funcionar perfectamente para reconocer el área de texto de la fuente a transcribir, sin embargo, también es posible crear modelos propios de reconocimiento de regiones de transcripción utilizando la herramienta *baseline model*[15].

[14] Sejda. *Crop Pdf*. Accesible desde: https://www.sejda.com/es/crop-pdf.
[15] Sobre los *baseline models*, véase: https://readcoop.eu/transkribus/howto/how-to-train-baseline-models-in-transkribus/.

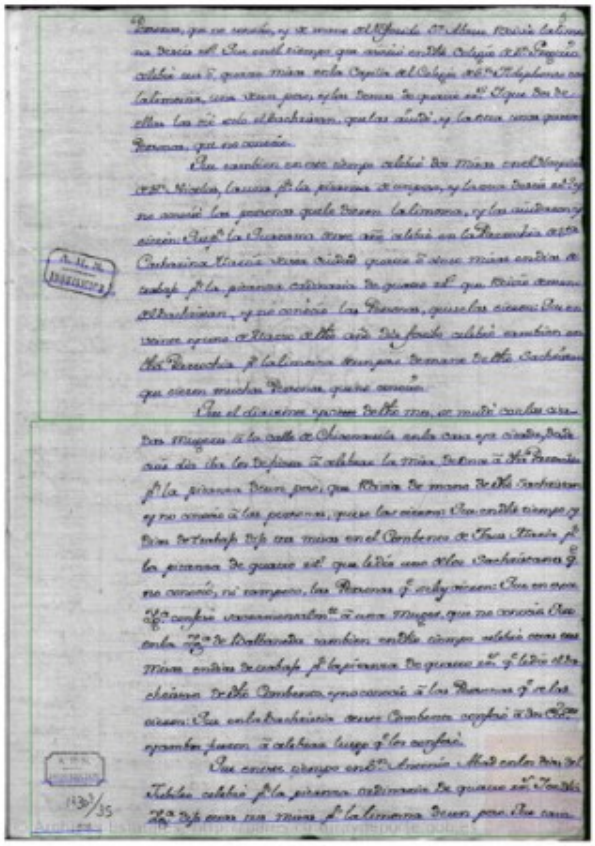

Figura 2. Reconocimiento de layout. Los bloques en verde contienen las áreas de texto (en morado) a reconocer delimitadas automáticamente por Transkribus. Fuente: elaboración propia.

Una vez se ha identificado la región de transcripción de todo el texto es necesario revisar y limpiar el área reconocida de texto no sujeto de ser transcrito. Es decir, el texto que se encuentra en el documento pero que no hace parte integral de este: sellos o anotaciones del archivista o del archivo que se agregaron de manera posterior como referencia bibliográfica (Figura 3). Al realizar este paso adicional se considera que el texto se encuentra listo para la etapa de transcripción. Transkribus dispone de una interfaz adaptable para ubicar el documento digitalizado a transcribir junto con el procesador de texto dividido en líneas. Cada página del documento tiene un set de líneas diferentes, pero no es necesario transcribirlas en orden (Figura 4).

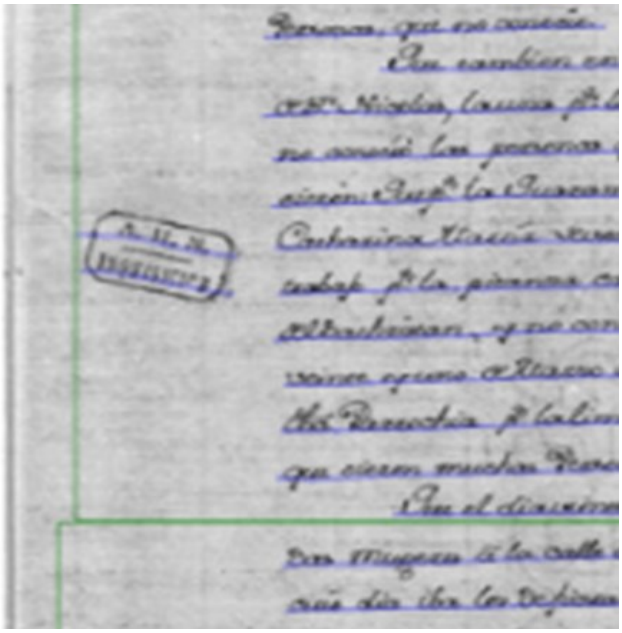

Figura 3. Remoción de áreas de texto no sujetas de transcripción. Fuente: elaboración propia.

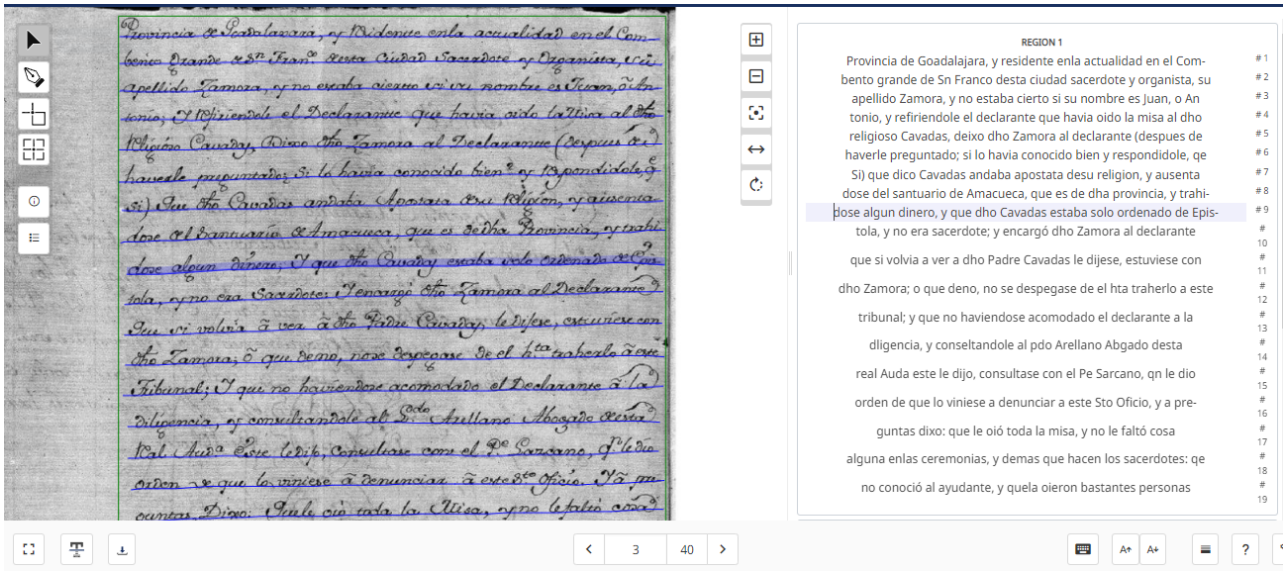

Figura 4. Transcripción manual inicial. Fuente: elaboración propia.

3.2. Criterios de transcripción y buenas practicas

El criterio de transcripción se establece de acuerdo con el resultado deseado, es decir, que mientras algunos autores pueden resolver las abreviaturas dentro del proceso de transcripción manual (Bazzaco et al., 2022, p. 95), las convenciones de transcripción de Transkribus indican lo contrario, aconsejando escribir las abreviaturas como se presentan en el texto y etiquetarlas como tal[16].

Para el proceso de transcripción de este texto en específico, hemos definido los siguientes criterios:

1. Se escribe cada línea de texto en la misma secuencia, si una palabra queda incompleta y se termina en la siguiente línea, se deja como tal.
2. La transcripción del texto será paleográfica, es decir, se utilizarán convenciones de transcripción como las establecidas por (Silva Prada, 2001)
3. Las abreviaturas no se resuelven, se transcriben de la misma manera que aparece en el texto.
4. Si hay una diferencia clara entre mayúsculas y minúsculas, estas se agregarán de forma diferencial.
5. Se usarán etiquetas para resolver las abreviaturas, diferenciar los nombres propios y enumerar las fechas para no intervenir la transcripción.
6. Se aplicará una transcripción no lineal para entrenar el modelo, es decir, se harán transcripciones en diferentes folios a lo largo del documento para obtener una mayor cercanía al estilo de letra del escribano a lo largo de su proceso de escritura, transcribir solo el principio podría no notar una escritura cansada y más alargada al final.
7. Las rubricas sí se transcriben y se agrega una etiqueta con el nombre propio del firmante si está disponible o se conoce con antelación.

[16] Véase nota 13.

8. Los signos de puntuación que se detecten serán transcritos con los comandos existentes en un teclado convencional, manteniendo la mayor cercanía posible con el texto original.
9. Si se encuentran ligaduras en el texto, estas se resolverán y no se marcarán como abreviaturas (ejemplo: æ se transcribe como ae).

El documento original empleado para esta transcripción está compuesto de cuarenta y tres folios (AGI, Inquisición,1730, Exp.35). Sin embargo, tres de ellos son páginas en blanco que se removieron desde la aplicación de Transkribus al cargar el documento en las colecciones.

De acuerdo con las recomendaciones de transcripción dadas por Transkribus, se realizó la transcripción de un aproximado de trece folios, de este proceso pudo evidenciarse que el tiempo de transcripción promedio de un folio es de treinta a cuarenta minutos, confirmando lo afirmado por (Massot et al., 2019). Este tiempo varía dependiendo del estado del documento original, de la cantidad de abreviaturas y nombres propios, así como de la calidad de digitalización para diferenciar una letra de otra.

3.3. Entrenamiento del modelo

Una vez realizada la transcripción, se designaron los folios con la categoría de *ground truth* (transcripciones de las que se está completamente seguro de su veracidad) y se designaron como set de entrenamiento (*training set*). Luego se cargó a las colecciones una copia del mismo documento para usarla como set de validación (*validation set*) (Figura 5). El set de validación es aquel conjunto de folios del mismo documento que no se usa para entrenar el modelo de reconocimiento de texto, sino que se usa para evaluar la precisión del modelo realizado.

Las páginas de validación fueron seleccionadas de manera manual, aunque Transkribus permite realizar una selección de manera automática, escogiendo el 2%, el 5% o el 10% del documento para comparar los resultados.

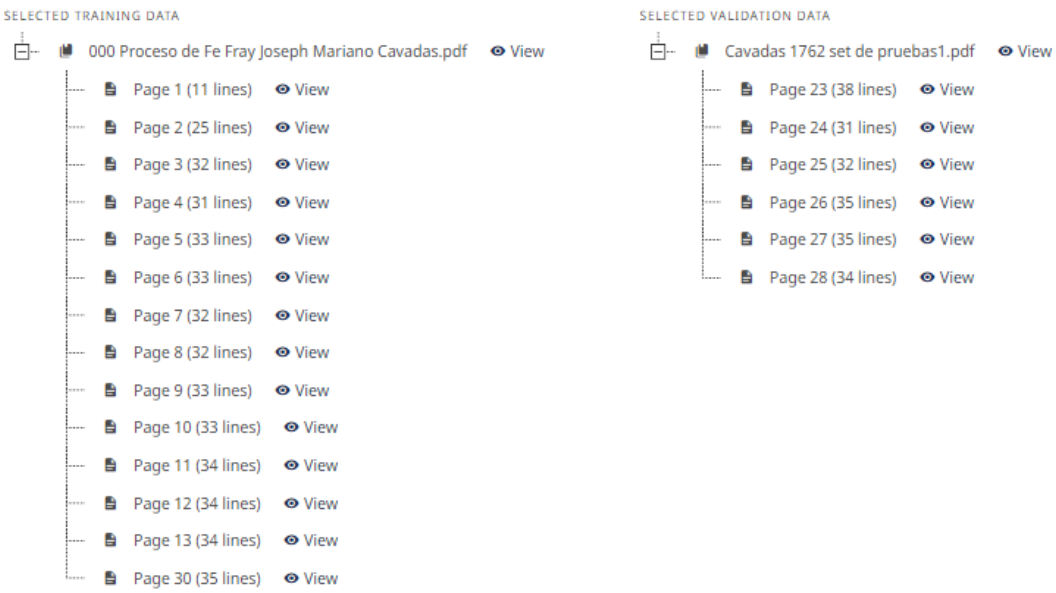

Figura 5. Modelo de entrenamiento y modelo de validación. Fuente: elaboración propia.

Con estos parámetros se realizó una serie de experimentos para comprobar el funcionamiento del modelo de inteligencia artificial.

El primer experimento utilizó trece folios como *ground truth* y usó los mismos folios como set de validación, como resultado se obtuvieron márgenes de error (CER) inferiores al 4%. Sin embargo, bajo este proceso, el modelo de transcripción automática probó ser hábil para transcribir los folios designados como *ground truth*, pero presentaba diversos errores para transcribir otros folios a pesar de su margen de error bajo. Las versiones del modelo Cavadas 1762 con números 1, 3, 4, 5 y 6 proporcionan resultados de este tipo (Figura 6).

Cavadas 1762 V6	17 243	Spanish; Castilian	2.70%
Cavadas 1762 V5	17 154	Spanish; Castilian	0.50%
Cavadas 1762 V4	17 018	Spanish; Castilian	2.50%
Cavadas 1762 V3	16 972	Spanish; Castilian	18.30%
Cavadas 1762 V1	2 914	Spanish; Castilian	0.10%

Figura 6. Versiones del modelo Cavadas1762 que se validaban sobre ellas mismas al no incorporar un set de validación con que comparar. Fuente: elaboración propia.

De este proceso puede concluirse que es recomendable e incluso necesario tener una copia del documento para ser usada como set de validación y el seleccionar folios fuera de los ya transcritos. En lugar de mezclar ambos para realizar el proceso de validación, de esta manera, el margen de error será mayor pero la transcripción de un folio en blanco tenderá a ser completamente legible. Este fue el segundo experimento realizado. La versión 7 del modelo Cavadas 1762 proporciona un entrenamiento de solo 4.522 palabras y un CER de 25.80%, sin embargo, probó ser completamente legible al transcribir los folios 26, 27 y 28 (Figura 7).

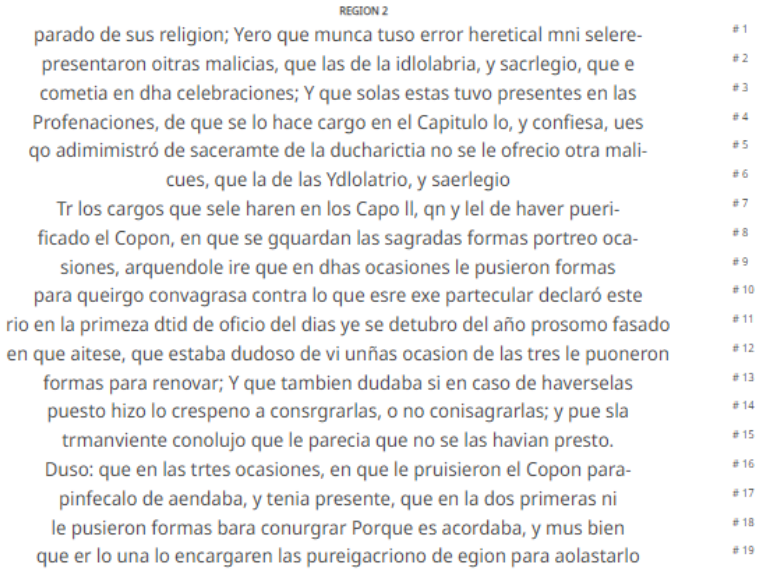

Figura 7. Resultado de reconocimiento del folio número 28 usando el modelo Cavadas1762 V7. Fuente: elaboración propia.

Aunque un CER de 25.8% sigue considerándose alto, el desarrollo del modelo de transcripción ha funcionado correctamente, ahora bien, para mejorar este margen de error se han realizado los siguientes procedimientos:

1. Entrenar nuevas versiones (V8 y V9) utilizando mayor número de *epochs* o veces que el modelo compara las palabras para procesar el resultado, aumentando de 100 a 200 *epochs*.
2. Transcribir de manera manual más folios para alimentar el modelo, se ha realizado la transcripción del folio número 30.

Con la versión 9 usada como modelo base, se realizó el entrenamiento de un nuevo modelo de reconocimiento (versión 10) que tiene un CER de 20.20%, es decir que de cada diez palabras un aproximado de dos pueden requerir corrección, el modelo mostró ser eficiente en el reconocimiento de itálica, probando que puede disminuirse el margen de error de un modelo entrenándolo con modelos previos como base sin hacer cambio alguno adicional.

Finalmente, gracias a los criterios de transcripción usados y a la alimentación del modelo con información adicional, la última versión del modelo Cavadas 1762 (versión 11) logró un CER de 8.9%. Esto quiere decir que de cada 100 palabras solo nueve podrían tener algún tipo error. Este margen de error es calculado al comparar el *ground truth* con el set de validación (Figura 8).

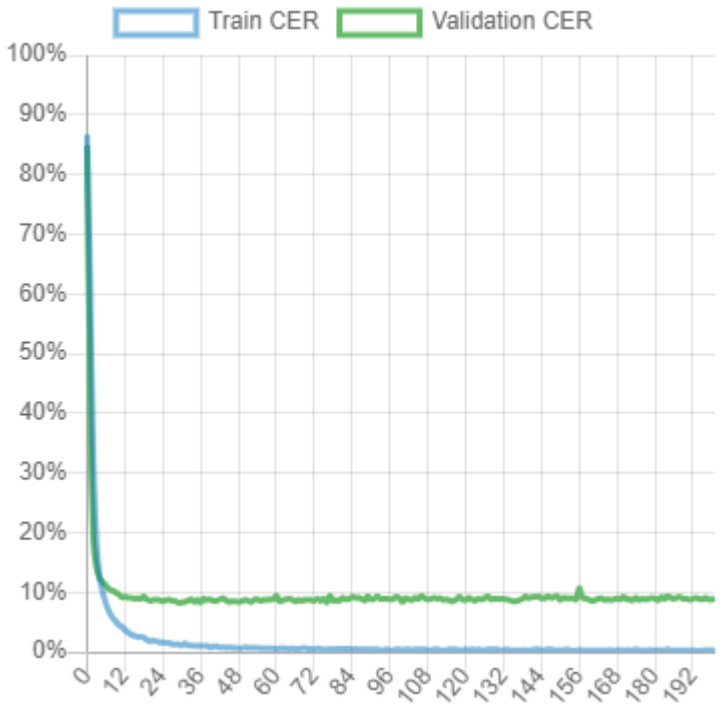

Figura 8. Resultados del modelo de reconocimiento Cavadas 1762, versión 11. Fuente: elaboración propia.

La versión 11 del modelo Cavadas 1762 es lo suficientemente consistente para ser utilizada en la totalidad del documento y realizar así su reconocimiento automático. Al comparar la versión 7 del modelo (Figura 7) con la versión 11 (Figura 9). Los resultados observables a simple vista son competentes en demostrar la mejoría del modelo de inteligencia artificial.

En resumen, los datos técnicos del modelo Cavadas 1762, versión 11, son los siguientes:

- Descripción Dataset: Cavadas 1762 V11
- Tipo de documentos: Manuscrito
- Nr. de palabras: 5.564
- Nr. de Caracteres: 53.691
- CER Training Set: 8.90%

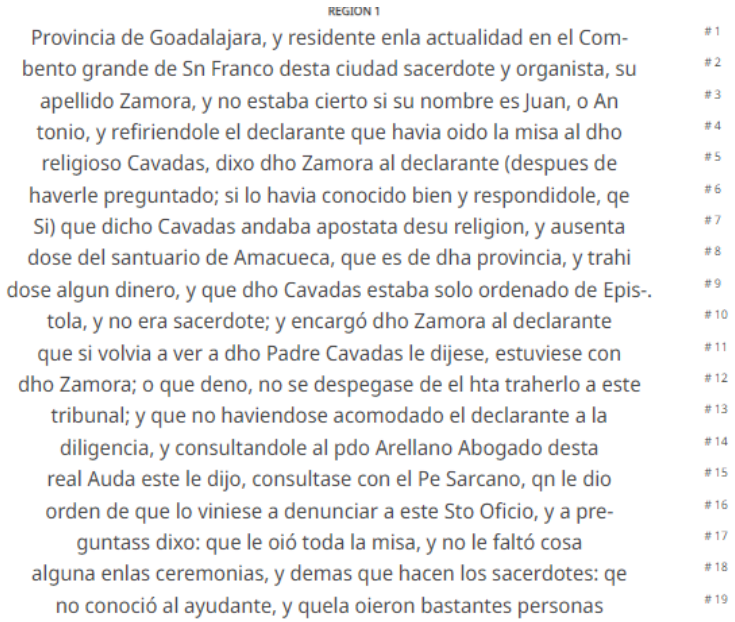

Figura 9. Ejemplo de reconocimiento automático por el modelo Cavadas 1762, versión 11. Fuente: elaboración propia.

3.4. Dificultades encontradas

Una de las principales dificultades encontradas en este proceso de reconocimiento automático de texto tiene que ver con la capacidad de procesamiento de Transkribus, al utilizar su suscripción gratuita, esta tiene un tiempo de espera para entrenar un modelo de inteligencia artificial superior al de las versiones con crédito pagos de la herramienta.

El tiempo para entrenar un modelo de IA como el utilizado en esta investigación puede llegar a ser de dos a seis horas, dependiendo de la capacidad de la herramienta y de si hay otros modelos a entrenar en cola. Es posible que, en un futuro, Transkribus aumente su capacidad de computación, disminuyendo así los tiempos de espera para obtener resultados del entrenamiento.

La segunda dificultad encontrada tiene que ver con los nombres propios. Si un documento tiene una gran variedad de nombres propios, similar al proceso de fe que utilizamos como documento base, el modelo de reconocimiento tendrá problemas para Transcribirlos si no se ha hecho una transcripción manual previa de estos. Por ejemplo, nombres que se repiten mucho como el de Joseph Cavadas (personaje principal del texto, a quien se le realiza el proceso de fe) pueden ser reconocidos con facilidad en folios no transcritos, pero nombres propios que no hayan sido transcritos previamente van a arrojar algún tipo de error, haciendo que sea complicado leerlos.

La tercera dificultad encontrada tiene que ver con la capacidad de la navegabilidad de la plataforma, pasar de un texto a otro para comprobar una palabra requiere salir de diferentes

secciones de la herramienta y los tiempos de carga pueden dificultar este proceso. Por último, al intentar hacer el reconocimiento individual de una página, el sistema también reconoce de nuevo el *layout* en su totalidad, sobrescribiendo cualquier configuración manual de *layout* que se haya hecho previamente. Así, el modelo intenta reconocer áreas del documento que no se suponía debía reconocer, afectando negativamente los resultados del reconocimiento.

4. CONCLUSIONES

La IA es una herramienta al servicio de la humanidad. Usarla para la transcripción de textos antiguos permite acelerar su comprensión y uso. Esta investigación permitió comprender el funcionamiento de Transkribus como plataforma y explorar sus capacidades dentro de las Humanidades Digitales. Así mismo, nos permitió establecer un protocolo de buenas prácticas para la transcripción automática de documentos que podemos utilizar al entrenar nuevos modelos acorde a los resultados requeridos.

Producto de esta investigación obtuvimos el modelo Cavadas 1762 que otorga la posibilidad de transcribir y reconocer letra itálica española manuscrita, una letra muy común en la segunda mitad del siglo XVIII.

Para preparar el modelo recurrimos a fuentes secundarias de proyectos previos que nos permitieron determinar lo que llamamos buenas prácticas de transcripción, es decir, pautas previas que nos permitieron dar rigor al proceso de transcripción del manuscrito en aras de obtener un reconocimiento automático efectivo. Los resultados demuestran que el modelo Cavadas 1762 es consistente y puede utilizarse como modelo base para entrenar otros modelos de reconocimiento global.

Dichas buenas prácticas de transcripción pueden condensarse en la siguiente lista:

1. Comprobar la correcta digitalización del documento antes de subirlo a la plataforma.
2. Recortar áreas extras del documento no sujetas de transcripción (encabezados y códigos automáticos) antes de subirlo a la plataforma.
3. Usar la herramienta de reconocimiento automático del área de transcripción (*layout recognition*) y luego borrar manualmente zonas del *layout* no sujetas a transcripción, como sellos o anotaciones del archivista.
4. Establecer unos criterios de transcripción previos que se adapten a las condiciones del documento.
5. Realizar un trabajo de transcripción consistente con los criterios establecidos.
6. Entrenar el modelo de IA con al menos 20% del texto transcrito de forma manual.
7. Usar como set de pruebas una copia en blanco del documento original, se recomienda que reconozca al menos el 5% como set de validación.
8. Entrenar nuevas versiones del modelo utilizando versiones previas como base hasta alcanzar resultados consistentes.

En cuanto a las dificultades encontradas, estas son valiosas como medio de aprendizaje

para futuros proyectos de transcripción, tanto las encontradas por autores citados (Massot et al., 2019, Bazzaco et al., 2022, Bazzaco, 2020, Schlagdenhauffen, 2020), como los hallazgos propios de este artículo. Ambos permiten aproximarse de mejor manera a la tarea de reconocer un texto de manera automática, una labor que podemos esperar se masifique en el futuro gracias a la tecnología.

Por último, los criterios de transcripción expuestos previamente en este artículo permitieron obtener un modelo de reconocimiento que pasó de un CER del 25.8% (Cavadas 1762, versión 7) a un modelo con un CER del 8.9% (Cavadas 1762, versión 11), siendo esta versión la más consistente con el reconocimiento del texto y permitiendo una lectura sencilla del documento.

Al comparar este modelo desarrollado para letra manuscrita del siglo XVIII español con otros modelos de letra manuscrita en Transkribus, se evidencia que el CER del modelo Cavadas 1762, versión11, se encuentra dentro de un rango cercano a modelos manuscritos públicos, como el de letra manuscrita francesa, que tiene un CER del 7.8%[17] o el desarrollado para la letra manuscrita de Carlos V, que tiene un CER del 8.3%[18].

Este modelo y el protocolo de buenas prácticas definido en esta investigación pueden ser usados como base para el entrenamiento de letras manuscritas similares, con el propósito de entrenar un modelo global de reconocimiento de itálica manuscrita y hacerlo disponible al público.

REFERENCIAS BIBLIOGRÁFICAS

Alemayehu, H. (2022). Handwritten Text Recognition Best Practice in the Beta maṣāḥəft workflow. *Journal of the Text Encoding Initiative*, 1-16. https://doi.org/10.4000/jtei.4109

Baucom, E. (2019). A Brief History of Digital Preservation. *Mansfield Library Faculty Publications*, 3-19. https://scholarworks.umt.edu/ml_pubs/31

Bazzaco, S. (2020). El reconocimiento automático de textos en letra gótica del Siglo de Oro: creación de un modelo HTR basado en libros de caballerías del siglo XVI en la plataforma Transkribus. *Janus. Estudios sobre el Siglo de Oro 9*, 534-561. https://www.janusdigital.es/articulo.htm?id=160

Bazzaco, S., Jiménez Ruiz, A., Torralba Ruberte, Á., & Martín Molares, M. (2022). Sistemas de reconocimiento de textos e impresos hispánicos de la Edad Moderna. La creación de unos modelos de HTR para la transcripción automatizada de documentos en gótica y redonda (s. XV-XVII). *Historias Fingidas*. Número especial 1, 67-125. https://doi.org/10.13136/2284-2667/1190

Deslandres, D., Couture, B., Farah, V., & Maxime, G. (2022). The Challenges of HTR Model Training: Feedback from the Project Donner le gout de l'archive a l'ere numerique. *arXiv*, 1-14. https://doi.org/https://doi.org/10.48550/arXiv.2212.11146

Kahle, P., Colutto, S., Hackl, G., & Muhlberger, G. (2017). Transkribus - A Service Platform for

[17] *French General Model*. Accesible desde: https://readcoop.eu/model/french-general-model/.
[18] Transkribus. *Carlos V/ Charles V*. Accesible desde: https://readcoop.eu/model/carlos-v-charles-v/.

Transcription, Recognition and Retrieval of Historical Documents. *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 19-24). IEEE. https://doi.org/10.1109/ICDAR.2017.307

Mancinelli, T. (2016). 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*, 255-260. https://doi.org/10.13136/2284-2667

Massot, M.-L., Sforzini, A., & Ventresque, V. (2019). Transcribing Foucault's handwriting with Transkribus. *Journal of Data Mining and Digital Humanities, Atelier Digit_Hum*, 1-17. https://doi.org/10.46298/jdmdh.5043

Memon, J., Sami, M., Khan, R., & Uddin, M. (2020). Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access, 8*, 142642-142668. https://doi.org/10.1109/ACCESS.2020.3012542

Milioni, N. (2020). Automatic Transcription of Historical Documents. Transkribus as a Tool for Libraries, Archives and Scholars. *Tesis de maestría*. Uppsala University. https://uu.diva-portal.org/smash/record.jsf?pid=diva2:1437985

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., & Colutto, S. (2019). Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study. *Journal of Documentation, 75*(5), 954-976. https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html#sec006

Nilsson, N. (2010). *The quest for artificial intelligence: a history of ideas and achievements*. Cambridge University Press.

Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research. *Archival Science, 22*(3), 367-392. https://doi.org/10.1007/s10502-022-09397-0

Readcoop. (s.f.). *Readcoop*. We revolutionise Access to Historical Documents: https://readcoop.eu

Readcoop. (s.f.). *Transkribus*. Unlock historical documents with AI: https://readcoop.eu/transkribus/

Reul, C., Stefan, T., Langhanki, F., & Springmann, U. (2022). Open Source Handwritten Text Recognition on Medieval Manuscripts Using Mixed Models and Document-Specific Finetuning. En S. Uchida, E. Barney, & V. Eglin (Ed.), *Document Analysis Systems. 13237* (pp. 414-428). Springer International Publishing. https://doi.org/10.1007/978-3-031-06555-2_28

Schlagdenhauffen, R. (2020). Optical Recognition Assisted Transcription with Transkribus: The Experiment Concerning Eugène Wilhelm's Personal Diary (1885-1951). *Journal of Data Mining & Digital Humanities, Atelier Digit_Hum*, 1-14. https://doi.org/10.46298/jdmdh.6249

Silva Prada, N. (2001). *Manual de paleografía y diplomática hispanoamericana, siglos XVI, XVII y XVIII. Libros de texto, manuales de prácticas y antologías*. Universidad Autónoma

Metropolitana, Unidad Iztapalapa.

Tegmark, M. (2018). *Vida 3.0: Qué significa ser humano en la era de la inteligencia artificial* (ebook ed.). Taurus.

Tran, H.-P., Smith, A., & Dimla, E. (2019). Offline Handwritten Text Recognition using Convolutional Recurrent Neural Network. *2019 International Conference on Advanced Computing and Applications (ACOMP)* (pp. 51-56). IEEE. https://doi.org/https://doi.org/10.1109/ACOMP.2019.00015