FACULTAT DE TRADUCCIÓ I D'INTERPRETACIÓ

GRAU D'ESPANYOL I XINÈS: LLENGUA, LITERATURA Y CULTURA

TREBALL DE FI DE GRAU
Curs 2025-2026

DESARROLLO DE UN PRIMER MODELO DE TRANSCRIPCIÓN AUTOMÁTICA APLICADO A LA COLECCIÓN MANUSCRITA YONGLE DADIAN

Abril Sánchez Reinaldo
1606159

TUTOR/A
ÁLVARO CUELLAR

Barcelona, 31 de mayo del 2026

UAB
Universitat Autònoma
de Barcelona

Datos del TFG

Título:

Desarrollo de un primer modelo de transcripción automática aplicado a la colección manuscrita Yongle Dadian.

Desenvolupament d'un primer model de transcripció automàtica aplicat a la col·lecció manuscrita Yongle Dadian.

Development of a first automatic transcription model applied to the Yongle Dadian manuscript collection.

Autor/a: Abril Sánchez Reinaldo

Tutor/a: Álvaro Cuellar

Centro: Universidad Autónoma de Barcelona

Estudios: Grado en Estudios de Español y Chino: Lengua, Literatura y Cultura

Curso académico: 2025-2026

Palabras clave:

Transkribus, HTR, Inteligencia artificial, Humanidades digitales, Literatura china, Yongle Dadian.

Transkribus, HTR, Intel·ligència artificial, Humanitats digitals, Literatura xinesa, Yongle Dadian.

Transkribus, HTR, Artificial Intelligence, Digital Humanities, Chinese Literature, Yongle Dadian.

Resumen del TFG:

Este Trabajo de Fin de Grado aborda el desarrollo y evaluación de un modelo de reconocimiento de texto manuscrito (HTR) aplicado a la Yongle Dadian (永乐大典), la mayor enciclopedia manuscrita de la China premoderna, compilada durante la dinastía Ming en el siglo XV. A lo largo de esta investigación, se analiza el potencial de la plataforma Transkribus para la transcripción automática de textos manuscritos en chino clásico, una lengua y un sistema de escritura que han sido escasamente abordados en el ámbito de las humanidades digitales. El estudio

combina una aproximación teórica y práctica, con especial atención al desarrollo del modelo de transcripción automática. Asimismo, el trabajo examina las dificultades que presentan las herramientas HTR actuales cuando se aplican a documentos no occidentales, prestando especial atención a los problemas de segmentación, reconocimiento y tratamiento de estructuras textuales verticales. De este modo, se pretende valorar tanto la viabilidad de la transcripción automática de manuscritos chinos clásicos como las limitaciones metodológicas y tecnológicas que todavía condicionan este campo de investigación.

Aquest Treball de Fi de Grau aborda el desenvolupament i avaluació d'un model de reconeixement de text manuscrit (HTR) aplicat a la Yongle Dadian (永乐大典), la major enciclopèdia manuscrita de la Xina premoderna, compilada durant la dinastia Ming al segle XV. Al llarg d'aquesta investigació, s'analitza el potencial de la plataforma Transkribus per a la transcripció automàtica de textos manuscrits en xinès clàssic, una llengua i un sistema d'escriptura que han estat abordats escassament en l'àmbit de les humanitats digitals. L'estudi combina una aproximació teòrica i pràctica, amb una atenció especial al desenvolupament del model de transcripció automàtica. Així mateix, el treball examina les dificultats que presenten les eines HTR actuals quan s'apliquen a documents no occidentals, prestant una atenció especial als problemes de segmentació, reconeixement i tractament d'estructures textuals verticals. D'aquesta manera, es pretén valorar tant la viabilitat de la transcripció automàtica de manuscrits xinesos clàssics com les limitacions metodològiques i tecnològiques que encara condicionen aquest camp de recerca.

This Final Degree Project addresses the development and evaluation of a handwritten text recognition (HTR) model applied to the Yongle Dadian (永乐大典), the largest handwritten encyclopedia of pre-modern China, compiled during the Ming Dynasty in the 15th century. Throughout this research, the potential of the Transkribus platform for the automatic transcription of handwritten texts in Classical Chinese, a language and writing system that have been scarcely addressed in the field of digital humanities, is analysed. The study combines a theoretical and practical approach, with special attention to the development of the automatic transcription model. Likewise, the work examines the difficulties that current HTR tools present when applied to non-western documents, paying special attention to the problems of segmentation, recognition and treatment of vertical textual structures. In this way, the aim is to assess both the feasibility of automatic transcription of classical Chinese manuscripts and the methodological and technological limitations that still condition this field of research.

ÍNDICE

1. Introducción *6*

1.1. Motivación **6**

1.2. Objetivos **7**

1.3. Metodología y estructura **8**

2. Marco teórico *9*

2.1. La tradición literaria manuscrita en la literatura china **9**

2.2. Características del sistema de escritura chino **11**

2.3. Elección del periodo y del corpus textual **13**

2.3.1. *Yongle Dadian* (永乐大典) 14

2.4. Transkribus **15**

2.4.1. Técnicas de aprendizaje automático 16

2.4.2. Funcionamiento de Transkribus: segmentación y reconocimiento 18

2.4.3. El eurocentrismo tecnológico y sus implicaciones 20

2.4.4. Modelo *baseline* y modelo HTR: diferencias y funciones 22

3. Desarrollo del modelo HTR **28**

3.1. Selección y preparación del corpus 28

3.2. Elaboración del *ground truth* 29

4. Resultados y análisis **40**

4.1. Evolución del CER a lo largo del entrenamiento 40

4.2. Análisis cualitativo de los errores 41

4.3. Discusión: viabilidad del HTR para el chino tradicional manuscrito 42

5. Conclusiones **44**

Bibliografía *46*

Bibliografía citada: **46**

Bibliografía consultada: **47**

Webgrafía y vídeos: **48**

Anexos..... *49*

Anexo 1. Yongle Dadian Baseline Model **49**

Anexo 2. Historial de modelos entrenados en Transkribus..... **50**

Anexo 2 a. Tabla de los modelos *baseline* con sus métricas 50

Anexo 2 b. Tabla de los modelos HTR con sus métricas 51

Anexo 3. 52 páginas transcritas manualmente para entrenar el modelo HTR de la colección **52**

Anexo 4. Historial de AI Jobs en Transkribus **53**

Anexo 5. Yongle Dadian HTR Model **56**

1. Introducción

1.1. Motivación

El estudio y la preservación de los manuscritos históricos constituyen una parte fundamental de la investigación en el ámbito de las humanidades. En el caso de la literatura china, que posee una de las tradiciones literarias más extensas y continuadas del mundo, conservamos manuscritos que abarcan siglos de historia y que representan una fuente de conocimiento de enorme valor para investigadores y estudiosos. Sin embargo, la dificultad de acceso a la lectura y transcripción de estos documentos limita en muchos casos su análisis sistemático y su difusión en el ámbito académico.

En los últimos años se han abierto nuevas posibilidades para abordar este reto gracias a la irrupción de la inteligencia artificial[1] y las técnicas de transcripción automática. La digitalización y el tratamiento automatizado de fuentes manuscritas facilitan el acceso a documentos históricos y literarios, mejorando su conservación, análisis y difusión. Como señalan Bazzaco et al. (2022), los avances en el reconocimiento automático de textos "han permitido una migración masiva del patrimonio textual analógico a la web", aunque ello plantea también nuevas responsabilidades en cuanto a la fiabilidad y supervisión de los resultados obtenidos (p. 71).

No obstante, mientras que estas tecnologías han avanzado notablemente en el tratamiento de escrituras latinas, su aplicación a sistemas de escritura no alfabéticos, como el chino clásico, sigue presentando desafíos importantes. Cuéllar (2023), refiriéndose a su experiencia con Transkribus en el contexto del teatro español del Siglo de Oro, señala que "en ese momento no existía ningún modelo para el reconocimiento del español, por lo que fue necesario su construcción desde cero" (p. 106). Esta misma situación se reproduce, de forma aún más acusada, en el caso del chino clásico manuscrito, donde la escasez de

[1] También conocida como IA o AI (*artificial intelligence*) en inglés.

modelos base y la complejidad gráfica del sistema de escritura dificultan considerablemente el punto de partida. En la actualidad, Transkribus cuenta con un primer modelo base para el chino, denominado *Test model Chinese*, con un *Character Error Rate* (CER)[2] del 7,0%, su existencia es reciente y su aplicabilidad al chino clásico manuscrito sigue siendo muy limitada.

Es precisamente ante esta situación que el presente trabajo considera pertinente y necesario explorar la aplicación de herramientas *Handwritten Text Recognition*[3] al patrimonio manuscrito chino, contribuyendo así a reducir la brecha tecnológica que separa las escrituras occidentales de las orientales en el ámbito de las humanidades digitales. (Transkribus, 2025)

1.2. Objetivos

El presente Trabajo de Fin de Grado tiene como objetivo principal entrenar un primer modelo específico de inteligencia artificial capaz de transcribir textos manuscritos originales de una colección concreta de la literatura china clásica. Con ello se pretende evaluar la viabilidad de las tecnologías HTR en el contexto de la escritura china, así como analizar el grado de precisión que puede alcanzarse mediante el entrenamiento supervisado del modelo.

De este objetivo principal se derivan los siguientes objetivos específicos: describir el funcionamiento técnico de Transkribus y los procesos de aprendizaje automático que lo sustentan. Entrenar y evaluar modelos HTR personalizados, documentando su evolución y los resultados obtenidos en términos de CER. Y, por último, reflexionar críticamente sobre las limitaciones tecnológicas y el eurocentrismo implícito en las herramientas HTR actuales.

Este trabajo tiene, además, una doble finalidad: contribuir al debate académico sobre el uso de herramientas digitales en el estudio de lenguas y

[2] El *Character Error Rate* (CER) es el porcentaje de caracteres reconocidos erróneamente por el modelo respecto al total de caracteres del texto de referencia. Se define con detalle en §2.4.1.
[3] De ahora en adelante HTR.

escrituras no occidentales, y explorar metodologías que faciliten la preservación y el acceso a fuentes manuscritas chinas.

1.3. Metodología y estructura

La metodología de este trabajo se divide en dos partes principales. La primera constituye el marco teórico, cuyo objetivo es contextualizar el objeto de estudio y fundamentar académicamente el trabajo. Para ello se han consultado fuentes bibliográficas especializadas en literatura china clásica, paleografía, humanidades digitales y reconocimiento automático de textos, incluyendo artículos académicos, monografías y documentación técnica de la plataforma Transkribus.

La segunda parte aborda el trabajo desde una perspectiva práctica. Se ha empleado Transkribus, plataforma digital desarrollada por READ-COOP [4] que permite la transcripción, el reconocimiento y el análisis de documentos históricos mediante inteligencia artificial. Mediante esta herramienta se ha elaborado el corpus de datos etiquetados (*ground truth*[5]) a través de la transcripción manual de la Yongle Dadian, se han entrenado modelos HTR personalizados y se han evaluado los resultados obtenidos.

El trabajo se estructura de la siguiente manera: en primer lugar, se presenta el marco teórico (§2), que incluye la contextualización de la tradición literaria manuscrita china, el análisis del sistema de escritura chino, la descripción del corpus textual seleccionado y la presentación técnica de Transkribus con sus métodos de aprendizaje automático. A continuación, se describe la metodología empleada (§3), seguida del análisis de los resultados (§4). Finalmente, se formulan las conclusiones del estudio (§5), y se incluyen la bibliografía y los anexos pertinentes.

[4] Cooperativa europea, sin ánimo de lucro, que gestiona y desarrolla Transkribus.
[5] "*Ground truth*" (datos de entrenamiento) hace referencia a los textos originales digitalizados emparejados con sus transcripciones humanas exactas y verificadas.

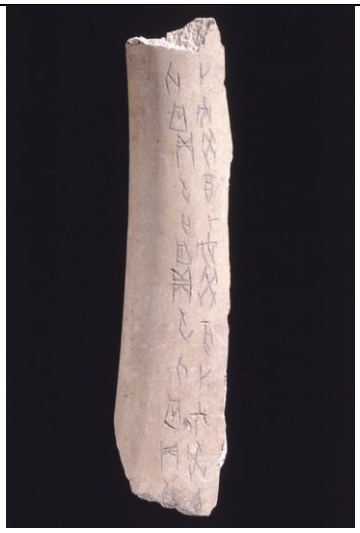
2. Marco teórico

2.1. La tradición literaria manuscrita en la literatura china

La literatura china es considerada una de las tradiciones literarias más extensas y continuadas del mundo, con orígenes orales que se remontan al siglo X a.C. La antología poética más antigua conocida es el *Shijing* (詩經), o Clásico de la Poesía, formada por 305 poemas de la dinastía Zhou (1046-256 a.C.) que abordan la vida, la cultura, la política y la sociedad de la época. Sin embargo, contamos con testimonios escritos aún anteriores, pertenecientes al siglo XIII a.C., grabados sobre huesos oraculares —caparazones de tortuga y escápulas de animales— denominados *jiaguwen* (甲骨文), en los que se consignaban asuntos de carácter histórico, oracular o cotidiano. Junto a estos, las inscripciones sobre bronce o metal en vasijas, conocidas como *jinwen* (金文), constituyen otro de los soportes más antiguos de la escritura china.

Figura 1

Imagen de un hueso oracular



Museum, T. T. o. t. B. (2016, February 24). Oracle Bone. *World History Encyclopedia*. https://www.worldhistory.org/image/4563/oracle-bone/

La escritura mantiene una relación estrecha con la evolución de la literatura. El concepto de literatura en chino se vincula principalmente a dos caracteres: *wen* (文) y *shi* (诗). El primero, *wen*, es lo que en occidente entenderíamos como literatura en sentido amplio, aunque se trata de un concepto históricamente variable y culturalmente situado. Como señala Prado Fonts (2008), el *wen* es un concepto dinámico cuya comprensión varía según cada época debido a sus influencias políticas, culturales y sociales, hasta evolucionar en el término actual *wenxue* (文学). El segundo, *shi*, en los textos más antiguos tenía el significado de *shijing*; según el *Shuowen jiezi* (說文解字) —el primer diccionario chino, compilado por Xu Shen hacia el año 100 d.C.— significa "intención" y se asocia con el acto de comunicar.

La literatura clásica china cumple principalmente una función de armonía social, a la vez que posee una dimensión estética articulada a través de géneros como la poesía lírica, la prosa filosófica y la narrativa de corte fantástico o mitológico. Los géneros principales de la literatura china clásica son la historiografía, la filosofía, la poesía y los textos militares y divinatorios; géneros como la ficción o el teatro no adquirieron reconocimiento literario hasta épocas tardías. El propio término de ficción en chino, *xiaoshuo* (小说), formado por los caracteres de "pequeño" y "hablar", refleja esta consideración marginal hasta finales de la dinastía Qin (206 a.C.).

La lengua china ha atravesado distintas etapas evolutivas que condicionan directamente la naturaleza de los manuscritos conservados. Siguiendo la periodización propuesta por Wilkinson (2013), pueden distinguirse cuatro fases principales: el Old Chinese (上古汉语), desde antes de la dinastía Shang hasta la caída de la Han; el Middle Chinese (中古汉语), influenciado por el budismo; el Mandarin (近代汉语), lengua de los funcionarios desde el final de la Tang hasta el fin de la Qing; y el Modern Chinese (现代汉语), lengua actual desde el Movimiento

del 4 de mayo (五四運動)[6], en adelante. Esta estratificación lingüística e histórica determina el grado de accesibilidad e interpretabilidad de los textos manuscritos conservados.

La transmisión física de estos textos se realizó durante siglos principalmente en soporte manuscrito, a través de copias sucesivas que dan cuenta de la extraordinaria importancia que la cultura china otorgaba a la conservación y transmisión del conocimiento. Como señala Bazzaco (2020), la reproducción del contenido de un documento en formato digital implica la transposición del lenguaje escrito a un formato comprensible para la máquina, mientras que la forma concreta del contenido coincide con la fotorreproducción del objeto (p. 537). Esta distinción entre texto digitizado y texto digitalizado es especialmente pertinente cuando se trabaja con manuscritos históricos chinos, cuya complejidad gráfica hace que la conversión automática del segundo al primero no sea en absoluto trivial.

En este sentido, la digitalización del patrimonio manuscrito chino constituye hoy un campo de estudio emergente dentro de las humanidades digitales, en el que el reconocimiento automático de textos se presenta como una herramienta de enorme potencial, aunque todavía con importantes limitaciones cuando se aplica a escrituras no occidentales.

2.2. Características del sistema de escritura chino

Para comprender las dificultades que plantea la transcripción automática del chino clásico manuscrito, es necesario analizar las características específicas de su sistema de escritura, que difiere fundamentalmente de los sistemas alfabéticos para los cuales se desarrollaron las herramientas HTR actuales.

El sistema de escritura chino se basa en caracteres logográficos [7] denominados hanzi (汉字), que en su origen derivaban de pictogramas. La lengua

[6] Movimiento social a raíz de las protestas de los estudiantes a las puertas de Tiananmén en Pekín el 4 de mayo de 1919, donde exigían una transformación de la sociedad china. El movimiento evidenció la necesidad de modernización del país para su prosperidad.

[7] Según la RAE, es un "sistema de escritura en el que los signos (logogramas) representan conceptos o palabras enteras en lugar de sonidos individuales". (Real Academia Española, 2026).

clásica se considera especialmente compleja por varias razones: es principalmente monosilábica, de modo que las palabras están formadas por una única sílaba; es fonéticamente pobre en cuanto a la variedad de sílabas disponibles, lo que genera una elevada cantidad de homófonos; y es aislante, es decir, carece de inflexión morfológica, por lo que la ausencia de marcadores temporales o de persona hace que muchos textos sean sintácticamente ambiguos.

Un rasgo fundamental desde el punto de vista del reconocimiento automático es la enorme variabilidad gráfica de los caracteres chinos. Un solo carácter puede estar compuesto por un número muy variable de trazos —entre uno y más de treinta—, y su forma varía significativamente según el estilo caligráfico, el período histórico, el copista y el soporte. Como apunta Su (2013), el reconocimiento automático de la escritura china se enfrenta a desafíos únicos derivados precisamente de esta complejidad gráfica y de la ausencia de fronteras claras entre caracteres (p. 1).

Otro factor decisivo es la dirección de escritura. La forma tradicional de escribir el chino clásico es en columnas verticales leídas de arriba abajo, ordenadas de derecha a izquierda. Esta orientación es radicalmente opuesta a la convención occidental de escritura horizontal de izquierda a derecha, que es la que asumen por defecto la mayoría de las herramientas de reconocimiento automático de textos, incluyendo Transkribus. Este desajuste entre la orientación del sistema de escritura y la lógica de las herramientas tecnológicas disponibles constituye uno de los obstáculos centrales del presente trabajo, como se desarrollará en el apartado §2.4.3.

Finalmente, el inventario de caracteres es exponencialmente más amplio que el de cualquier alfabeto. El diccionario Kangxi del siglo XVIII recoge más de 47.000 caracteres, de los cuales unos 3.500 son de uso frecuente en chino moderno, y los textos clásicos emplean un subconjunto distinto y específico de ese inventario. Esta amplitud del espacio de caracteres implica que los modelos HTR necesitan haber sido expuestos a una enorme cantidad de ejemplos para poder generalizar adecuadamente el reconocimiento.

2.3. Elección del periodo y del corpus textual

La selección del corpus textual ha constituido uno de los pasos más determinantes del trabajo, en la medida en que condiciona tanto la viabilidad técnica del proyecto como su relevancia académica. Para ello se han tenido en cuenta criterios de carácter metodológico, práctico e histórico-literario.

El primer criterio ha sido la disponibilidad de fuentes digitalizadas de acceso abierto. El trabajo con herramientas de reconocimiento automático de textos requiere el acceso a imágenes digitalizadas de alta calidad, que permitan tanto la anotación manual como el entrenamiento del modelo. La existencia de repositorios digitales que ofrezcan acceso a manuscritos chinos clásicos es, por tanto, un requisito previo indispensable.

El segundo criterio ha sido el equilibrio entre antigüedad del texto y complejidad gráfica. En términos cronológicos, cuanto más antiguo es un manuscrito, menor es la cantidad de ejemplares conservados en buen estado, y mayor es la distancia lingüística y gráfica respecto al chino actual. Por el contrario, textos demasiado recientes pueden encontrarse ya impresos o mecanografiados, lo que reduce el interés del estudio desde la perspectiva del reconocimiento de escritura manuscrita. Este equilibrio sitúa el período óptimo entre la dinastía Ming (1368-1644 d.C.) y la dinastía Qing (1644-1912 d.C.), correspondiente a la etapa del Mandarin (近代汉语) identificada por Wilkinson (2013).

El tercer criterio ha sido el valor académico intrínseco del corpus. Se ha buscado trabajar con un texto que posea relevancia histórica y literaria documentada, de modo que los resultados del trabajo no solo tengan interés metodológico, sino que también contribuyan a facilitar el acceso a un corpus de especial importancia para la investigación sinológica.

A la luz de estos factores, se ha seleccionado como corpus la Yongle Dadian (永乐大典), que reúne todos los criterios establecidos: es un manuscrito de la dinastía Ming con versiones digitalizadas de acceso abierto, posee una escritura

homogénea y de buena calidad gráfica, y representa uno de los proyectos enciclopédicos más ambiciosos de la historia china.

2.3.1. *Yongle Dadian* (永乐大典)

La *Yongle Dadian* (永乐大典), conocida en español como la Enciclopedia Yongle, es considerada la enciclopedia más extensa de la China premoderna y una de las mayores compilaciones enciclopédicas de la historia universal. Fue encargada por el emperador Yongle de la dinastía Ming, quien ordenó a un equipo de aproximadamente 3.000 eruditos la recopilación sistemática de obras procedentes de múltiples ámbitos del conocimiento: filosofía, historia, geografía, literatura, arte y religión. La obra fue completada en torno al año 1408 d.C.

En su versión completa, la enciclopedia constaba de 22.877 *juan*[8] (卷), a los que se añadían un prólogo y un índice de 60 juan, distribuidos en aproximadamente 11.095 volúmenes. Se estima que el conjunto alcanzaba alrededor de 370 millones de caracteres, lo que la convierte en una de las compilaciones textuales más extensas jamás realizadas. Su contenido abarca textos que van desde la Antigüedad china anterior a la dinastía Qin (221-206 a.C.) hasta los inicios de la propia dinastía Ming (1368-1644 d.C.), incorporando tanto obras clásicas como documentos históricos, literarios y científicos.

La obra original sufrió graves pérdidas a causa de la inestabilidad política de finales de la dinastía Ming. Una primera copia se realizó durante el reinado del emperador Jiajing en el siglo XVI, y es de esta versión de la que provienen prácticamente todos los volúmenes conservados actualmente. Con el paso del tiempo y los conflictos posteriores, la mayor parte de estos volúmenes también fue dispersada o destruida. En la actualidad solo se conservan alrededor de 400 volúmenes. La Biblioteca Nacional de China posee más de la mitad del total mundial, con 221 volúmenes en su colección, y desde principios del siglo XXI se ha

[8] Clasificador para tomos de libros o capítulos tradicionales.

encargado de restaurar y digitalizar la obra, con el objetivo de reunificar digitalmente volúmenes dispersados por todo el mundo (Xie et al., s.f.).

Desde el punto de vista paleográfico, la Enciclopedia Yongle presenta características especialmente adecuadas para el presente estudio. Las páginas muestran una escritura en chino tradicional de estilo regular (*kǎishū,* 楷书), con columnas de texto verticales de derecha a izquierda, y una notable homogeneidad caligráfica debida a la naturaleza colectiva y supervisada del proyecto enciclopédico. Esta uniformidad es, como se verá en §3.1, un factor favorable para el entrenamiento de modelos HTR.

Las imágenes utilizadas en este trabajo han sido obtenidas a través de la Biblioteca del Congreso de los Estados Unidos, que pone a disposición del público versiones digitalizadas de alta resolución de varios volúmenes de la copia Jiajing (Xie et al., s.f.). Para este estudio se han seleccionado y trabajado 52 páginas, que han sido transcritas manualmente y utilizadas como corpus de entrenamiento del modelo HTR personalizado.

2.4. Transkribus

Transkribus es una plataforma digital basada en inteligencia artificial desarrollada por la cooperativa READ-COOP, diseñada para la digitalización, transcripción, reconocimiento y análisis de documentos históricos tanto manuscritos como impresos. En el campo de las humanidades digitales, Transkribus se ha consolidado como una herramienta de referencia para proyectos de transcripción automática a gran escala. Como señala Cuéllar (2023): Transkribus pretende ser "un espacio libre para el reconocimiento de textos: no importa la lengua o el alfabeto empleado" (p. 106). Asimismo, Cuéllar (2023, p.106), nos explica que el proceso básico de Transkribus consiste en:

> ... transcribir adecuadamente un número suficiente de palabras, para que, más adelante, a través de un sistema de Inteligencia Artificial con redes neuronales, la máquina aprenda a enlazar cada parte del texto con su transcripción correspondiente. De esta manera, y requiriendo decenas o incluso cientos de mi- les de palabras

correctamente transcritas, la máquina es capaz de aprender y de transcribir correctamente las líneas del nuevo documento que se le ingrese.

El funcionamiento de Transkribus se estructura en tres niveles de uso. El primer nivel (*Quick*) permite obtener resultados rápidamente con modelos ya entrenados, siguiendo el flujo:

Subir documentos → reconocimiento → exportar resultados.

El segundo nivel (*Guided*) permite mejorar la precisión eligiendo el modelo más adecuado y ajustando la detección del *layout*:

Subir documentos → reconocimiento → editar resultados → exportar.

El tercer nivel (*Custom*) es el empleado en el presente trabajo, y consiste en crear o adaptar el propio modelo:

Subir documentos → entrenamiento → reconocimiento → editar → exportar.

2.4.1. Técnicas de aprendizaje automático

Para comprender el funcionamiento de Transkribus es necesario conocer las técnicas de aprendizaje automático que lo sustentan. Las principales son el *machine learning*, el *deep learning*, el HTR y el OCR, desglosadas a continuación.

El *machine learni*ng es una rama de la inteligencia artificial que permite a los sistemas informáticos aprender a partir de datos y mejorar su rendimiento en tareas específicas sin necesidad de ser programados explícitamente para cada una de ellas. A través del análisis de grandes volúmenes de información, los modelos de *machine learning* identifican patrones, realizan predicciones y toman decisiones de manera autónoma.

El *deep learning* es una subdisciplina del *machine learning* basada en el uso de redes neuronales artificiales de múltiples capas. Estas arquitecturas, inspiradas en el funcionamiento del cerebro humano, permiten aprender representaciones complejas y abstractas a partir de grandes conjuntos de datos. Gracias a esta

capacidad, el *deep learning* se emplea especialmente en tareas de reconocimiento de imágenes, procesamiento del lenguaje natural y transcripción automática.

En este contexto, el *Handwritten Text Recognition* (HTR) es una tecnología específica que utiliza técnicas de *deep learning* para procesar imágenes de documentos manuscritos y convertirlos en texto digital editable. A diferencia del *Optical Character Recognition* (OCR), que se centra principalmente en texto impreso y trabaja con un inventario tipográfico relativamente predecible, el HTR tiene que lidiar con la gran variabilidad propia de la escritura manuscrita: diferencias entre copistas, variaciones de tinta, irregularidades del soporte y diversidad de estilos caligráficos. Como señala Ayuso García (2022), la exactitud de los resultados de la transcripción se mide mediante el *Character Error Rate* (CER), definido como "el porcentaje de caracteres reconocidos erróneamente en un conjunto dado" (p. 152).

Respecto a la diferencia entre OCR y HTR, conviene aclarar que, si bien ambos persiguen el mismo objetivo —convertir una imagen de texto en texto digital—, su ámbito de aplicación y sus técnicas son distintos. El OCR, señala Ayuso García (2022, 152):

> … es una cuestión resuelta para los impresos modernos, en los que un CER < 0,5% se consigue con cualquier aplicación comercial … este cometido dista aún de estar resuelto para los impresos anteriores a la invención de la prensa automática en el siglo XIX y presenta generalmente peores resultados con los impresos más antiguos.

El HTR surge precisamente para abordar estos casos más complejos mediante el uso de redes neuronales recurrentes (RNN[9]) capaces de modelar la variabilidad de la escritura manual.

[9] De las siglas en inglés *Recurrent Neural Network*, es un modelo de aprendizaje profundo diseñado para procesar datos secuenciales o series temporales.

2.4.2. Funcionamiento de Transkribus: segmentación y reconocimiento

El proceso de transcripción automática en Transkribus se divide en dos fases técnicas claramente diferenciadas: el análisis de *layout* (segmentación) y el reconocimiento de texto (HTR propiamente dicho). Esta distinción es fundamental para entender tanto el funcionamiento de la herramienta como las dificultades específicas que presenta su aplicación al chino clásico manuscrito.

La primera fase es el *Advanced Layout Anal*ysis (ALA), también denominado análisis de *layout*. Su función es detectar automáticamente las regiones de texto presentes en la imagen (*Text Regions*[10]) y trazar las líneas base (*Baselines*[11]) sobre las que se apoyan los caracteres. Las *baselines* son las líneas imaginarias que marcan la dirección y posición de cada línea de texto, y constituyen la referencia geométrica fundamental sobre la que opera posteriormente el modelo HTR. Este proceso funciona mediante redes neuronales entrenadas para detectar estructuras de texto en imágenes, y su rendimiento depende en gran medida de que el *layout* del documento se ajuste a los patrones aprendidos durante el entrenamiento.

La segunda fase es el reconocimiento de texto. Una vez segmentadas las líneas, el modelo HTR procesa cada una de ellas de forma independiente y genera la transcripción textual correspondiente. El modelo aprende la relación entre los patrones visuales de los caracteres y su representación digital a través del entrenamiento supervisado con datos etiquetados (*ground truth*).

Para entrenar un modelo personalizado, Transkribus requiere de un conjunto de datos de entrenamiento denominado *ground truth*, definido como texto correctamente transcrito emparejado con su imagen correspondiente. Ayuso García (2022, 153) explica:

> A partir del primer resultado se selecciona una parte del texto para corregir la transcripción proporcionada de manera automática y crear una transcripción sin

[10] De ahora en adelante TR.
[11] De ahora en adelante BL.

errores. Con estos datos el sistema se entrena y la máquina aprende esta tarea. Este proceso se puede repetir hasta conseguir la menor tasa de error posible.

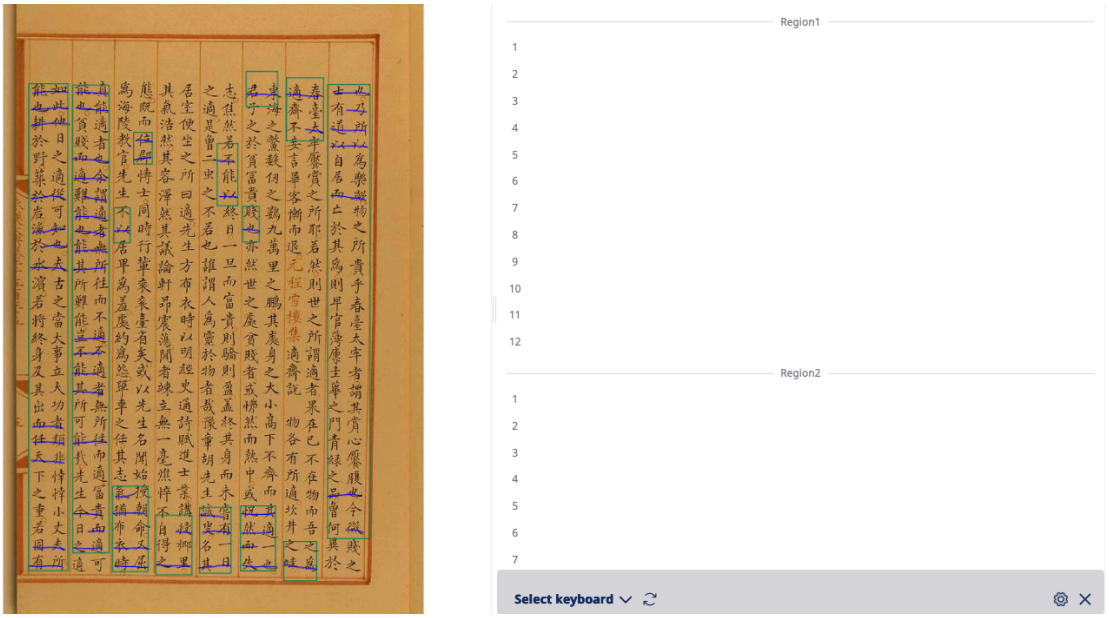
Los parámetros clave del entrenamiento incluyen el número de *epochs* (pasadas completas sobre el conjunto de datos de entrenamiento), que determina cuántas oportunidades tiene el modelo para ajustar sus parámetros y mejorar su rendimiento.

Existen dos vías para elaborar un modelo personalizado. La primera consiste en la transcripción manual de las páginas, que una vez completada se guarda como *ground truth* para iniciar el primer entrenamiento. La segunda vía más eficiente en términos de tiempo y consiste en aplicar primero un modelo público sobre las imágenes, corregir los errores que genera y guardar el resultado corregido como *ground truth*. En ambos casos, las páginas deben haber sido transcritas, corregidas y guardadas antes de iniciar el proceso de entrenamiento. Cuando la transcripción presenta fragmentos ilegibles, la plataforma permite etiquetarlos como *unclear* (poco legible) o *gap* (laguna), evitando que el ruido en los datos perjudique el aprendizaje del modelo (Transkribus, 2025).

En el presente trabajo se ha optado por la primera vía —transcripción íntegramente manual—, dado que ningún modelo público ofrece resultados aceptables para el chino clásico vertical, lo que hacía inviable la segunda opción como punto de partida, como vemos en la figura 2.

Figura 2

Resultado de la segmentación automática aplicada mediante el modelo público Test model Chinese sobre una página del Yongle Dadian.



Nota. En la parte izquierda puede observarse la imagen original del manuscrito con las baselines detectadas automáticamente, cuya delimitación no se corresponde con la estructura real de las columnas verticales del texto. En la parte derecha se muestra el panel de transcripción, dividido en dos regiones (Region 1 y Region 2), en el que puede comprobarse que ninguna de las líneas segmentadas ha generado transcripción alguna: el modelo no ha reconocido ningún carácter. Este resultado evidencia la inviabilidad del modelo público como punto de partida para el corpus trabajado, y justifica la opción de la transcripción íntegramente manual como método de elaboración del ground truth. Fuente: elaboración propia a partir de Transkribus.

2.4.3. El eurocentrismo tecnológico y sus implicaciones

Una de las reflexiones más importantes que emerge del trabajo con Transkribus en el contexto del patrimonio manuscrito chino es la del eurocentrismo implícito en el diseño de las herramientas HTR actuales. Este concepto, ya apuntado en la literatura especializada en humanidades digitales, hace referencia a la tendencia de las plataformas tecnológicas a asumir como universal un conjunto de convenciones propias de las escrituras occidentales, con

consecuencias directas para la investigación sobre lenguas y escrituras no occidentales.

Cuéllar (2023) afirma que Transkribus "no ha sido pensada y creada en atención a una lengua y un alfabeto determinado" y que "no importa la lengua o el alfabeto empleado" ya que “está orientada a ser un espacio libre para el reconocimiento de textos” (p. 106), así pues, la experiencia práctica de este trabajo demuestra que esta afirmación es válida en términos de intención, pese a que, cuando se aplica al chino clásico manuscrito, se encuentran ciertas dificultades. Por ejemplo, las herramientas de análisis de *layout* de Transkribus asumen, por defecto, que el texto está dispuesto en líneas horizontales orientadas de izquierda a derecha, lo que responde a la convención de las escrituras europeas. En el caso de la Enciclopedia Yongle, sin embargo, el texto está organizado en columnas verticales leídas de arriba abajo y ordenadas de derecha a izquierda, una estructura que el analizador de *layout* no es capaz de procesar correctamente sin intervención manual.

Esta limitación no es un error puntual o corregible mediante el ajuste de parámetros: es una consecuencia estructural de que los modelos de segmentación fueron entrenados casi exclusivamente sobre documentos de escritura occidental. Como señalan Bazzaco et al. (2022), los avances en el reconocimiento automático de textos han favorecido "una migración masiva del patrimonio textual analógico a la web", pero lo han hecho de forma asimétrica, privilegiando las escrituras y los documentos para los que existe una mayor demanda y tradición tecnológica (p. 71). “La figura del humanista” que trabaja con patrimonio no occidental "ha quedado inevitablemente fuera de la ecuación", como señalan los mismos autores (p. 71).

La escasa consideración de las escrituras no occidentales en el entrenamiento de estos modelos tiene consecuencias metodológicas directas en el presente trabajo. La detección incorrecta de líneas horizontales donde en realidad hay columnas verticales produce una sobresegmentación masiva: en las pruebas realizadas, el modelo llegó a detectar 100 líneas en páginas que contenían entre 15 y 20 columnas de texto real. Esto implica que el modelo HTR recibe

fragmentos mínimos de caracteres —a veces un único carácter, a veces ninguno— en lugar de columnas completas, lo que invalida el reconocimiento posterior. El resultado es un CER inicial del 100%, que no mejora hasta que se interviene manualmente en la segmentación.

Esta problemática no es exclusiva de Transkribus: afecta de forma general a las plataformas HTR disponibles actualmente, que carecen de modelos base para el chino manuscrito clásico. Mientras que para escrituras latinas existen modelos base entrenados con cientos de miles de palabras —el modelo Noscemus GM 4.0, por ejemplo, alcanza un CER del 0,79% con un conjunto de entrenamiento de 541.611 palabras (Ayuso García, 2022, p. 153, nota 9)—, para el chino tradicional manuscrito no existe ningún modelo de referencia sobre el que construir, ya que el modelo público *Test model Chinese* es para el chino actual, el cual se escribe, al igual que lenguas occidentales, de izquierda a derecha y de arriba abajo. Esta situación pone de manifiesto, en el campo de las humanidades digitales, la desigualdad de recursos tecnológicos para el estudio de lenguas y culturas no occidentales.

2.4.4. Modelo *baseline* y modelo HTR: diferencias y funciones

En el contexto del entrenamiento con Transkribus, es fundamental distinguir entre dos tipos de modelos que cumplen funciones complementarias y actúan en fases distintas del proceso de transcripción: el modelo de *baseline* (o modelo de *layout*) y el modelo HTR (o modelo de reconocimiento de texto).

El modelo de *baseline* es el responsable de la primera fase del proceso: detectar automáticamente las regiones de texto y trazar las líneas de base (*baselines*) sobre las imágenes del documento. Su función es esencialmente geométrica y estructural: determinar dónde está el texto en la página y cuál es la dirección y posición de cada línea. Este modelo trabaja sobre la imagen completa del documento y genera una segmentación que sirve como entrada para el modelo HTR. Un modelo de *baseline* deficiente —que segmente incorrectamente las líneas— invalidará el trabajo del modelo HTR, independientemente de la calidad de este último, ya que operará sobre fragmentos de imagen incorrectos.

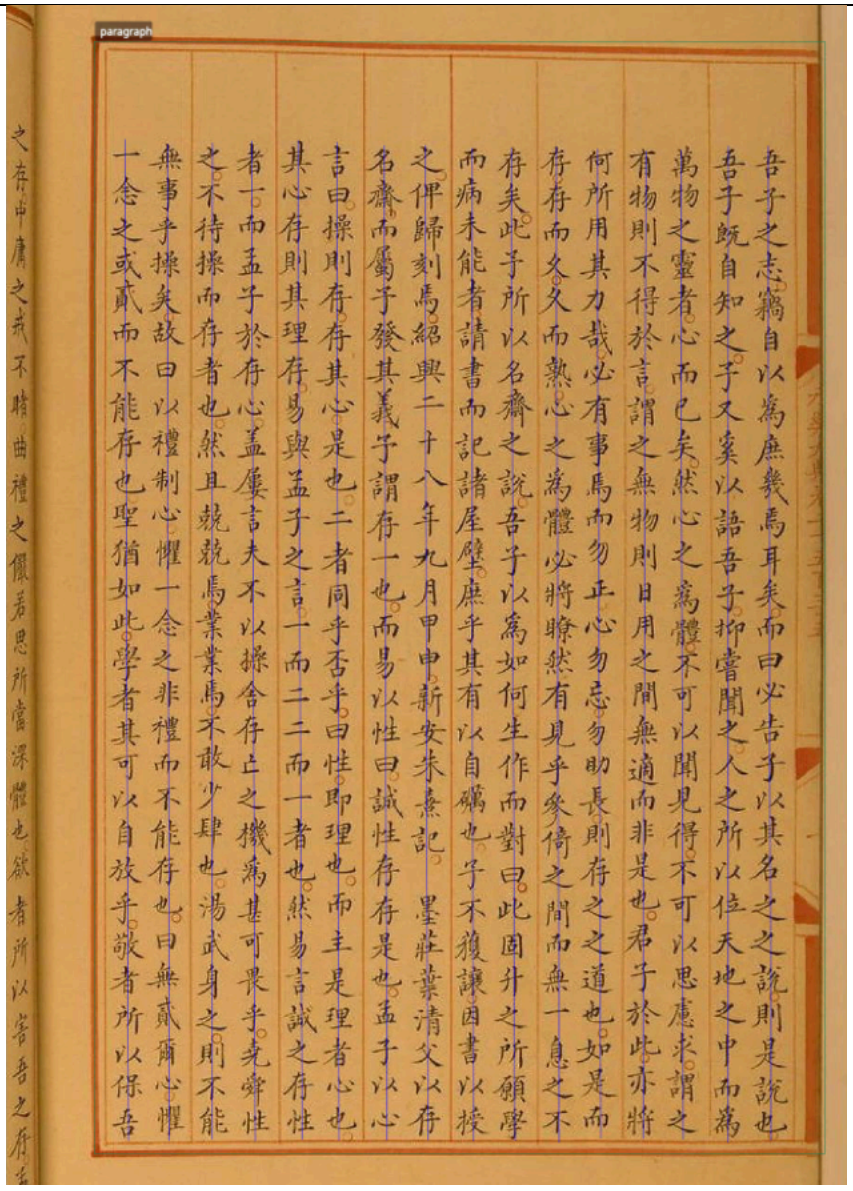
El modelo HTR, por su parte, actúa sobre las líneas ya segmentadas por el modelo de *baseline*. Su función es reconocer los caracteres presentes en cada línea y generar la transcripción textual correspondiente. Aprende la relación entre patrones visuales y representaciones digitales a través del entrenamiento supervisado con *ground truth*, y su rendimiento se mide mediante el CER.

En el presente trabajo, el modelo de baseline empleado es el modelo *Universal Lines* proporcionado por Transkribus, que, como se ha señalado en §2.4.3, presenta dificultades significativas con la escritura vertical del chino clásico. La alternativa es entrenar un modelo de *baseline* personalizado para texto vertical, proceso que requeriría un volumen de datos anotados aún mayor. En consecuencia, los resultados de los modelos HTR quedan condicionados desde la fase de segmentación, de modo que su rendimiento máximo se ve estructuralmente limitado.

A lo largo del proceso se entrenaron seis modelos de segmentación, (en el anexo 1 encontramos el sexto modelo), partiendo de un primer modelo con 64 líneas anotadas y una pérdida (*loss*) del 41,00%, hasta alcanzar el modelo denominado Yongle Dadian Baseline 6 (ID: 568449), entrenado con 414 líneas y una pérdida del 6,60%. Este modelo se entrenó con 25 páginas, de las cuales 23 fueron *ground truth* y 2 *Validation pages*. En la figura 3 vemos un ejemplo de *baseline* manual.

Figura 3

Ejemplo del modelo de segmentación Yongle Dadian baseline.



Fuente: elaboración propia a partir de Transkribus.

La reducción drástica de la pérdida —de más del 40% hasta menos del 7%—, que se muestra a continuación, en la figura 4, indica que el modelo ha aprendido a detectar correctamente las columnas verticales del manuscrito, constituyendo un avance fundamental para las fases posteriores del proyecto.

Figura 4

Tabla de modelos Baseline.

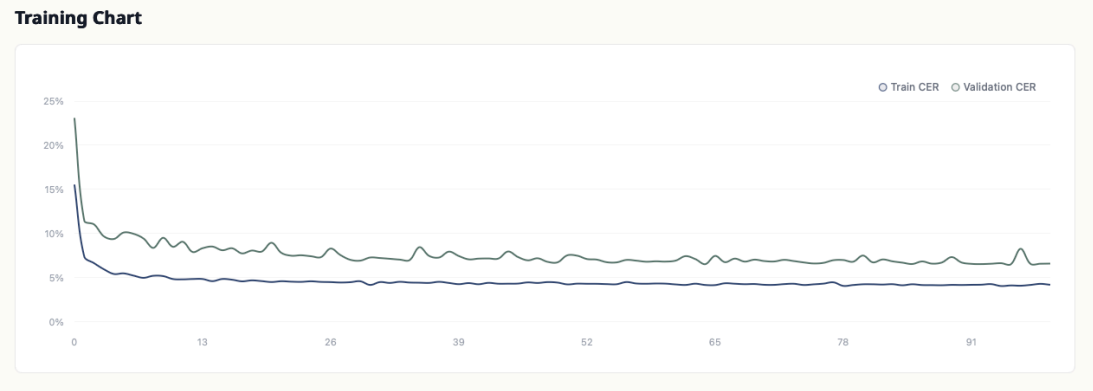
Modelo	***Líneas***	***Loss***	***ID***
Yongle Dadian Baseline 1	64	41,00%	548137
Yongle Dadian Baseline 2	64	41,57%	549337
Yongle Dadian Baseline 3	80	59,74%	550817
Yongle Dadian Baseline 4	321	46,76%	565669
Yongle Dadian Baseline 5	321	58,53%	565689
Yongle Dadian Baseline 6	**414**	**6,60%**	**568449**

Tabla comparativa de los 6 modelos distintos de Baseline entrenados. Fuente: elaboración propia.

La métrica *loss* utilizada para los modelos de *baseline* mide el error del modelo en la detección y delimitación de las líneas de texto, y no debe confundirse con el CER, que evalúa el reconocimiento de caracteres. Un *loss* del 6,60% implica que el modelo *Yongle Dadian Baseline 6* segmenta correctamente aproximadamente 9 de cada 10 líneas del corpus, lo que representa una base sólida sobre la que continuar el entrenamiento HTR en trabajos futuros. Los modelos HTR entrenados en el presente estudio han operado sobre una segmentación subóptima generada por el modelo universal, pero la existencia de este modelo de *baseline* personalizado abre la posibilidad de obtener resultados significativamente mejores en investigaciones posteriores que combinen ambos modelos.

Figura 5

Curva de entrenamiento del modelo Yongle Dadian Baseline 6 (ID: 568449) en Transkribus.



La curva de Train CER (azul oscuro) alcanza el 4,61% en el epoch 14, mientras que la Validation CER (verde) se sitúa en el 8,53% en el mismo punto, estabilizándose ambas en los epochs posteriores. Fuente: elaboración propia a partir de Transkribus.

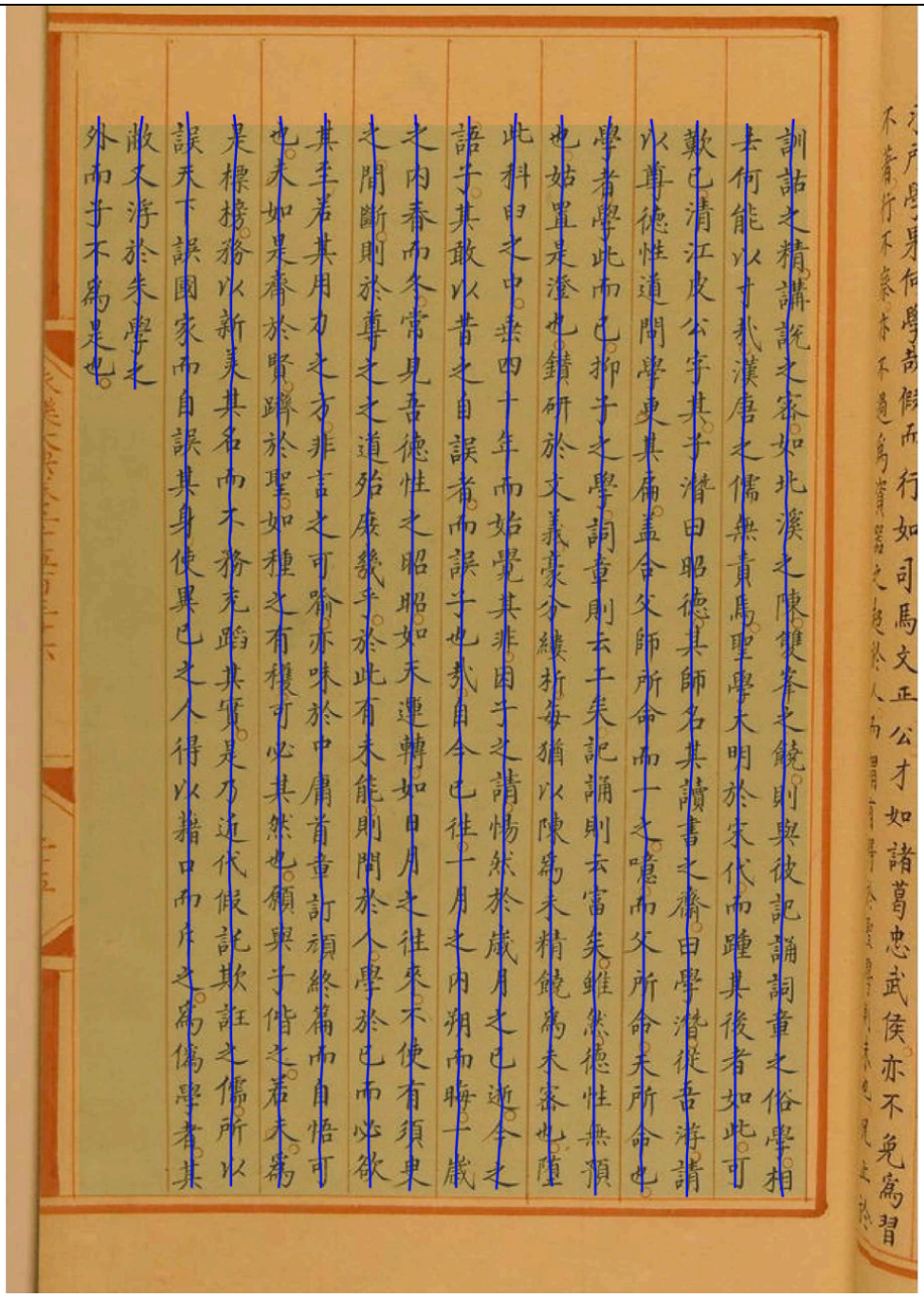
La figura 5 muestra la curva de entrenamiento del modelo *Yongle Dadian Baseline 6* a lo largo de aproximadamente 100 *epochs*. El eje vertical representa el CER aplicado a la detección de líneas y el eje horizontal el número de *epochs* completados, con dos curvas: el *Train CER* (azul oscuro), que mide el error sobre los datos de entrenamiento, y el *Validation CER* (verde), que lo hace sobre el conjunto de validación.

El comportamiento de la curva es académicamente significativo por varias razones. En primer lugar, ambas líneas parten de valores elevados —en torno al 22% y al 15% respectivamente— y descienden de forma pronunciada durante los primeros *epochs*: en el *epoch* 14, el *Train CER* ya se sitúa en el 4,61% y el *Validation CER* en el 8,53%, una reducción superior al 50% respecto a los valores iniciales en muy pocas iteraciones. En segundo lugar, a partir del *epoch* 14 ambas curvas se estabilizan y oscilan de forma moderada en torno a sus valores finales —aproximadamente un 4-5% para el entrenamiento y un 6-8% para la validación— sin mostrar signos de sobreajuste (*overfitting*) severo, lo que evidencia que el modelo mantiene un rendimiento consistente sobre datos no vistos durante el entrenamiento. En tercer lugar, la brecha entre *Train CE*R y *Validation CER* se

mantiene contenida a lo largo de todo el proceso, lo que indica que el conjunto de datos, pese a su tamaño limitado (414 líneas), es suficientemente representativo del corpus.

Figura 6

Baseline realizado por el modelo propio "Yongle Dadian Baseline".



Fuente: elaboración propia a partir de Transkribus.

3. Desarrollo del modelo HTR

La parte aplicada se ha llevado a cabo íntegramente mediante la plataforma Transkribus. A continuación, se detalla el proceso seguido en cada una de las etapas.

3.1. Selección y preparación del corpus

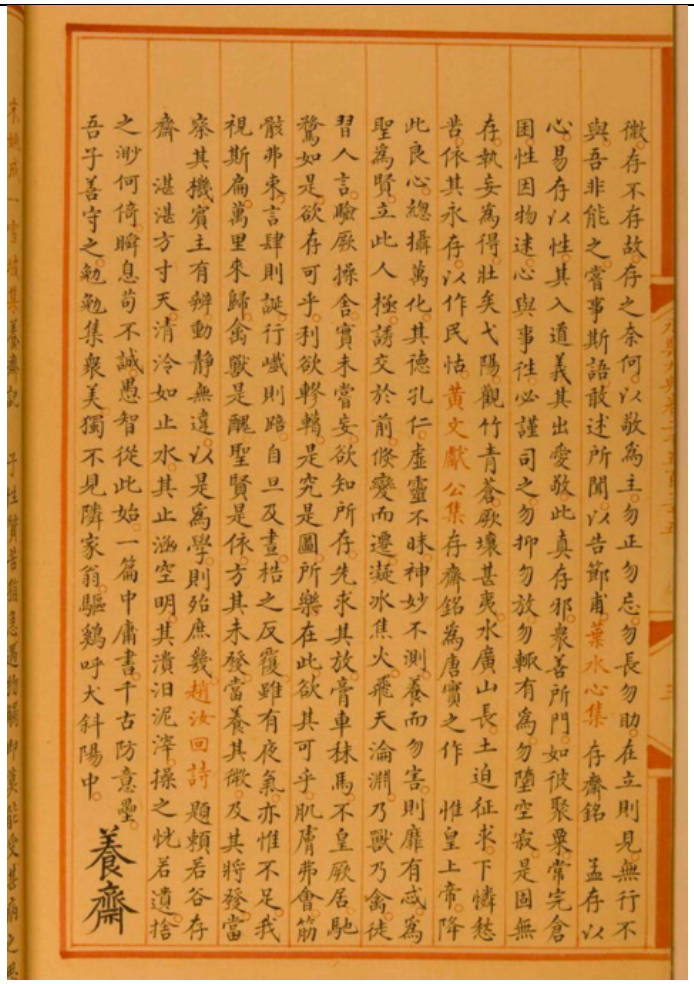
El primer paso consistió en la selección del corpus. Se optó por trabajar con páginas procedentes del volumen 2530 de la enciclopedia Yongle (永樂大典卷之二千五百三十), enciclopedia manuscrita compilada durante el reinado del emperador Yongle (1402-1424), en la dinastía Ming. Como ya hemos mencionado previamente, esta obra presenta características paleográficas especialmente adecuadas para un proyecto de transcripción automática: escritura en chino clásico de estilo regular (*kǎishū*[12]), organización en columnas verticales de texto homogéneo, y acceso digital de calidad a través de la Biblioteca del Congreso de Estados Unidos.

Para este trabajo se seleccionaron y cargaron en Transkribus un total de 107 páginas digitalizadas, organizadas en la colección denominada Yongle Dadian. Las imágenes presentan una calidad aceptable para el proceso de digitalización, aunque algunas páginas muestran leves variaciones de contraste o sombras que pueden afectar al análisis de *layout*. La homogeneidad de la escritura entre páginas —consecuencia del carácter institucional y supervisado del proyecto enciclopédico original— constituye un factor favorable para el entrenamiento, ya que reduce la variabilidad del corpus de entrenamiento.

[12] Conocido como escritura regular o estándar, es el estilo de caligrafía china más utilizado hoy en día.

Figura 7

Muestra de la homogeneidad de las páginas seleccionadas para el estudio.



Página 8 del archivo, observamos la homogeneidad de las páginas. Fuente: elaboración propia a partir de Transkribus.

3.2. Elaboración del *ground truth*

Dado que Transkribus requiere de un conjunto de datos etiquetados para el entrenamiento supervisado del modelo, ha sido necesario elaborar transcripciones manuales que sirvan como *ground truth*, es decir, como referencia base para evaluar el rendimiento del modelo. Este proceso es el más laborioso y determinante de todo el flujo de trabajo, ya que la calidad y cantidad del *ground truth* condicionan directamente la precisión final del modelo. Finalmente, se han transcrito manualmente 52 de las 107 páginas, ya que, al finalizarlas, los resultados han sido suficientes para cumplir los objetivos del presente trabajo.

Cada página ha sido anotada manualmente siguiendo el orden del trabajo propio de Transkribus: en primer lugar, se definieron las regiones de texto[13].

Figura 8

Herramienta de selección TR.



Se utiliza la herramienta TR (Text Region) para delimitar la región correspondiente al texto. Fuente: elaboración propia a partir de Transkribus.

[13] *Text Regions o TR.*

A continuación, se marcan manualmente las líneas base[14] de cada columna de texto.

Figura 9

Herramienta de segmentación BL (Base Line)



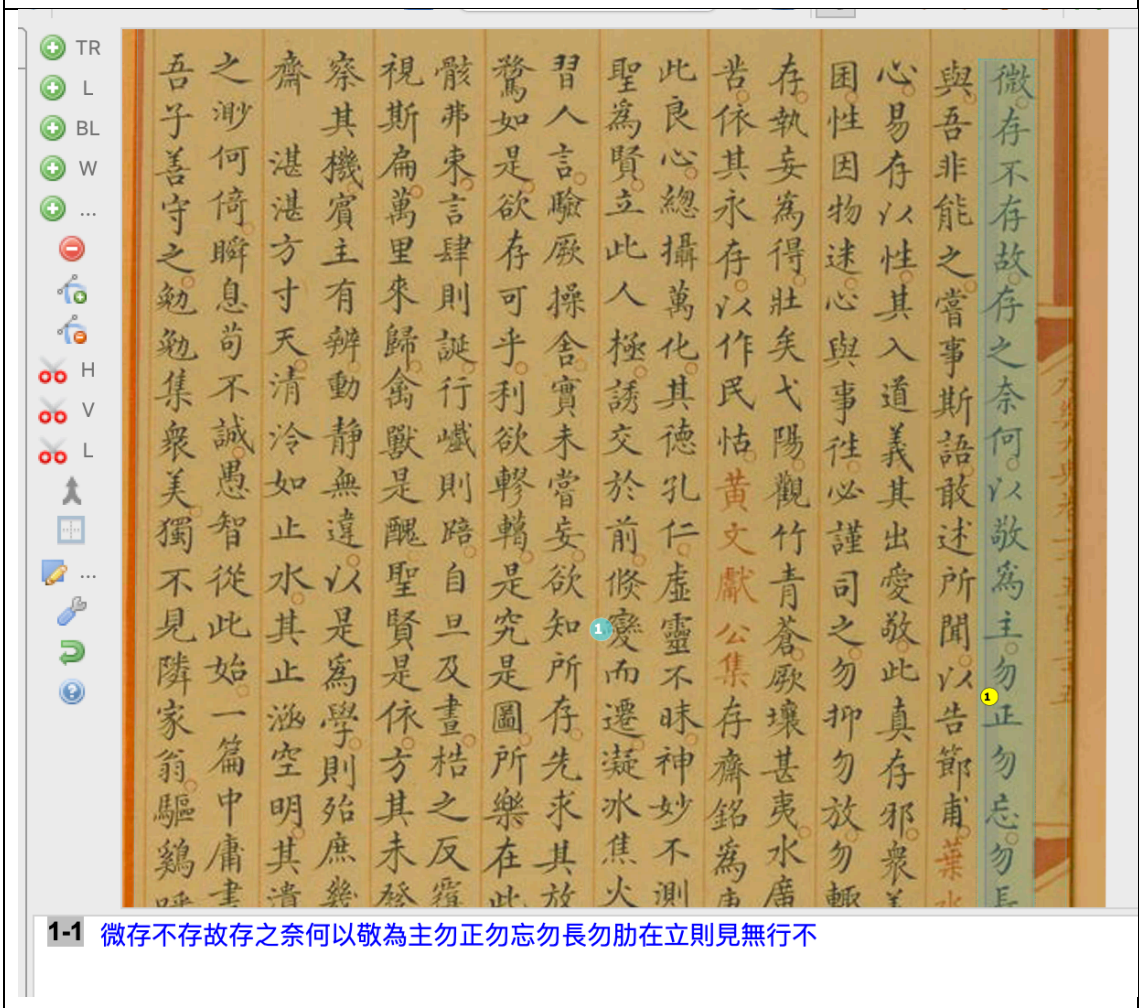
Se sigue la línea correspondiente a la primera frase y se seleccionan todos los caracteres que la componen en vertical. Fuente: elaboración propia a partir de Transkribus.

[14] *Baseline o BL.*

Finalmente se introduce la transcripción carácter a carácter de cada línea. La orientación vertical del texto chino supone un conflicto. Se han probado distintas opciones: ajustar diversos parámetros, realizar diferentes pruebas de *baseline*, e incluso, girar la orientación del documento noventa grados a la izquierda manualmente para que las *baselines* siguieran la dirección horizontal que propone el analizador automático de *layout*.

Figura 10

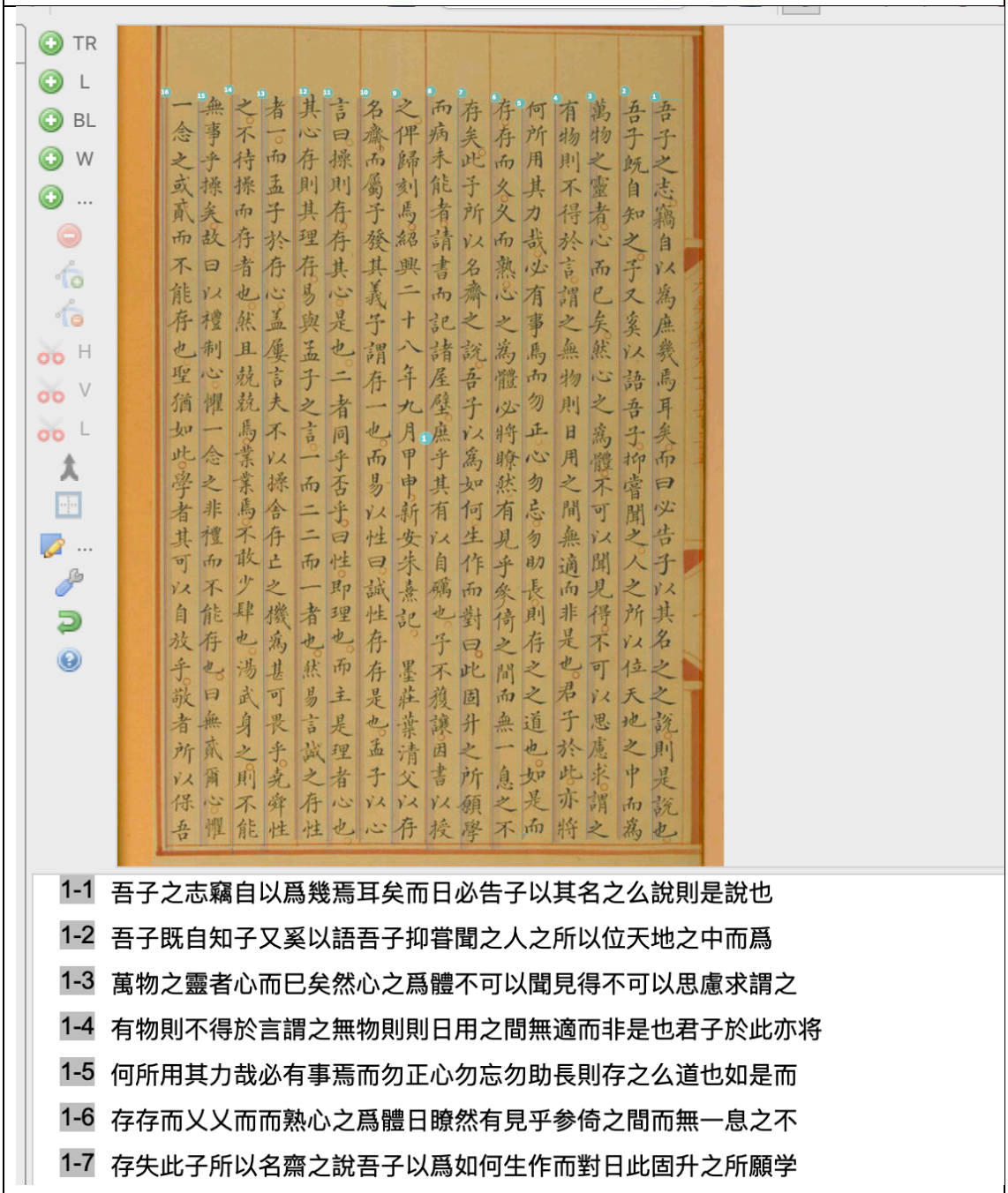
Transcripción manual carácter por carácter.



En la parte inferior de la imagen puede observarse, en azul, la frase transcrita manualmente, correspondiente a la primera línea del texto. Fuente: elaboración propia a partir de Transkribus.

Figura 11

Ejemplo de segmentación y transcripción manual de una página completa.

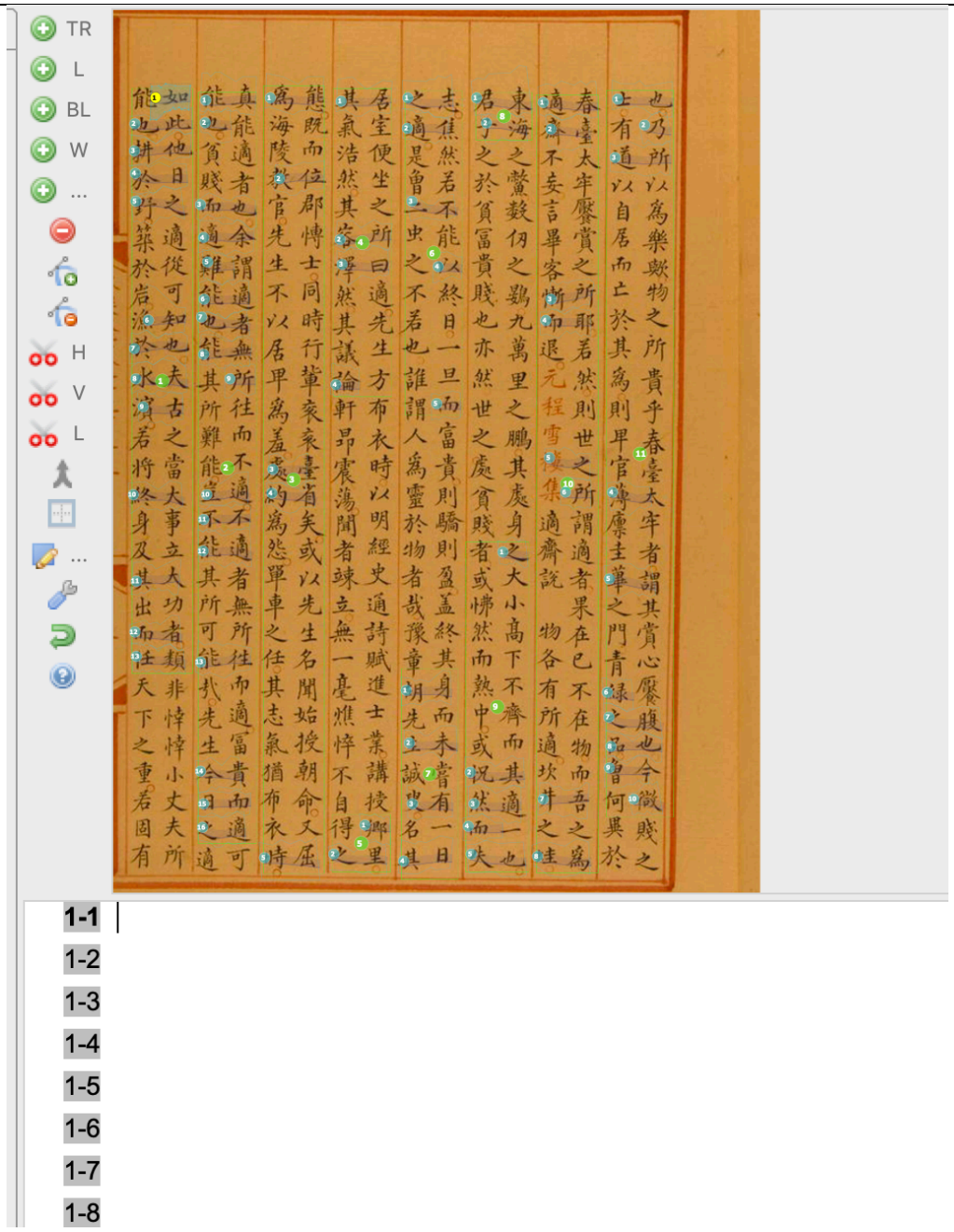


Como puede observarse, todas las frases se encuentran delimitadas mediante las baselines. Debajo de estas, aparece la transcripción numerada, cuyos números se corresponden con los identificadores de cada línea, visibles en forma de círculos azules en la parte superior del texto segmentado. Fuente: elaboración propia a partir de Transkribus.

Una vez realizado este proceso en las cinco primeras páginas correspondientes, pudo llevarse a cabo la primera prueba del modelo HTR. El objetivo de esta fase fue evaluar el porcentaje de error obtenido y corregir posteriormente los fallos detectados, de manera que el modelo pueda aprender a partir de ellos. Se realizó la primera prueba, y como se puede observar en la figura 12, la prueba arroja un porcentaje de error del 100 %, ya que el modelo no consigue detectar correctamente las líneas segmentadas, como hemos mencionado antes, un problema que, si no se corrige, puede afectar a todo el estudio.

Figura 12

Primera prueba del modelo HTR.



En esta imagen se observa la segmentación realizada por el primer modelo de prueba HTR. Puede apreciarse que los marcadores numéricos azules aparecen distribuidos a lo largo de toda la imagen. Esto se debe a que el modelo no ha sido capaz de detectar correctamente las baselines correspondientes a las oraciones,y, por consiguiente, el proceso de transcripción no ha podido llevarse a cabo. Fuente: elaboración propia a partir de Transkribus.

Así pues, a medida que se realizaron diversas pruebas, se fueron incorporando nuevas páginas de *ground truth* y corrigiendo los errores detectados en los resultados obtenidos. Llegando a ampliar progresivamente el volumen de *ground truth* disponible a lo largo del proceso, pasando de un primer conjunto de 71 palabras, con el que se realizó la primera prueba de entrenamiento, hasta alcanzar un corpus total de 1.011 palabras distribuidas en 52 páginas transcritas. El proceso completo de la transcripción manual de las 52 páginas supuso un trabajo de varias semanas, en el que se priorizó la coherencia y la fidelidad a los caracteres originales del manuscrito.

Como puede observarse en la figura 12, a la hora de segmentar mediante el *Advanced Layout Analysis* [15] se pone de manifiesto uno de los problemas fundamentales del trabajo: la herramienta detecta correctamente la presencia de texto en la página, pero la segmentación no se ajusta a la estructura real de las columnas verticales del manuscrito. El modelo interpreta los caracteres de una misma columna como si fueran líneas independientes, generando una sobresegmentación sistemática. En el ejemplo que tenemos, el modelo ha segmentado la página en 100 líneas, cuando la transcripción manual contemplaba 16 líneas en la página. Ha sido necesario, por tanto, corregir manualmente la segmentación en la mayoría de las páginas antes de proceder al entrenamiento. Por ello, en el historial de trabajos (*AI Jobs*) de la plataforma se registraron un total de 54 trabajos completados, de los cuales una parte significativa corresponde a sucesivas ejecuciones del análisis de *layout* sobre distintas versiones de la colección. Véase anexo 5.

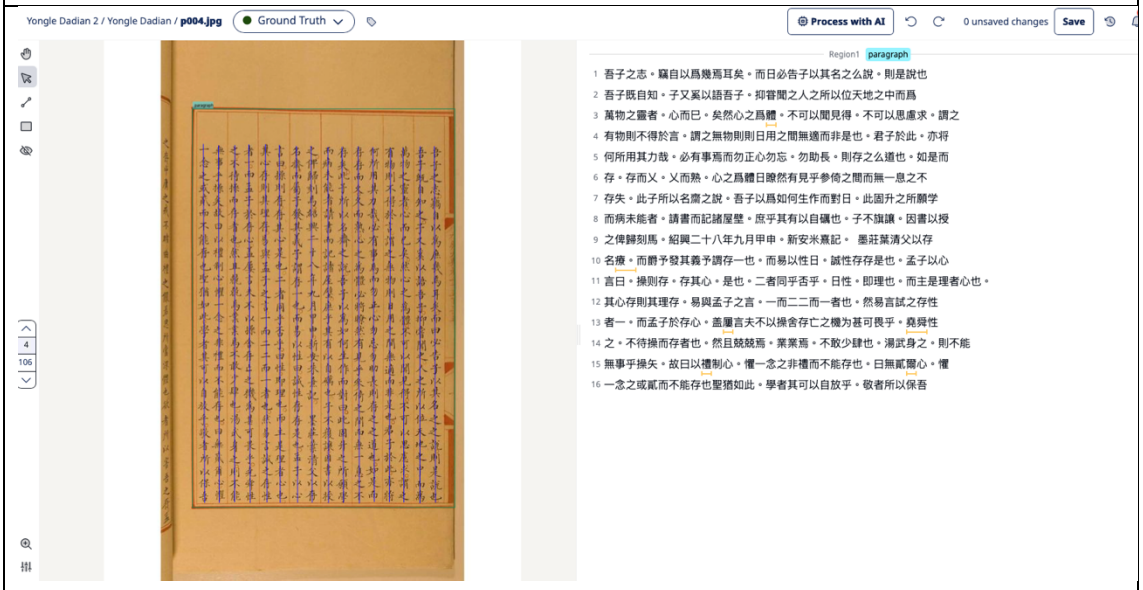
Con el *ground truth* realizado, se procede al entrenamiento del modelo personalizado. El proceso se realiza a través del módulo de entrenamiento de Transkribus, en el que se configura el conjunto de datos de entrenamiento, el conjunto de validación y los parámetros del modelo, incluyendo el número de

[15] De ahora en adelante ALA

epochs. Las transcripciones manuales de entrenamiento pueden encontrarse en el anexo 6, pero a continuación pueden ver una muestra.

Figura 13

Ejemplo de página segmentada y transcrita.



Fuente: elaboración propia a partir de Transkribus.

A lo largo del proceso se entrenaron un total de 13 modelos propios, con variaciones en el volumen de datos y en los parámetros de configuración. La Tabla 1 recoge la evolución de los principales modelos entrenados, ordenados cronológicamente por su identificador en la plataforma.

Tabla 1

Evolución de los modelos HTR entrenados para el Yongle Dadian en Transkribus.

Modelo	Palabras entrenadas	CER (%)	ID Transkribus
Yongle Dadian 1-2	71	100,00	564969
Yongle Dadian 3	315	93,11	565649
Yongle Dadian 5-6	329	93,26	565749
Yongle Dadian 7	659	92,38	568609
Yongle Dadian 8	820	92,04	568689
Yongle Dadian 9	975	82,67	568789

Yongle Dadian 10	**2.079**	96,76	568849
Yongle Dadian 12	1.011	85,18	568909

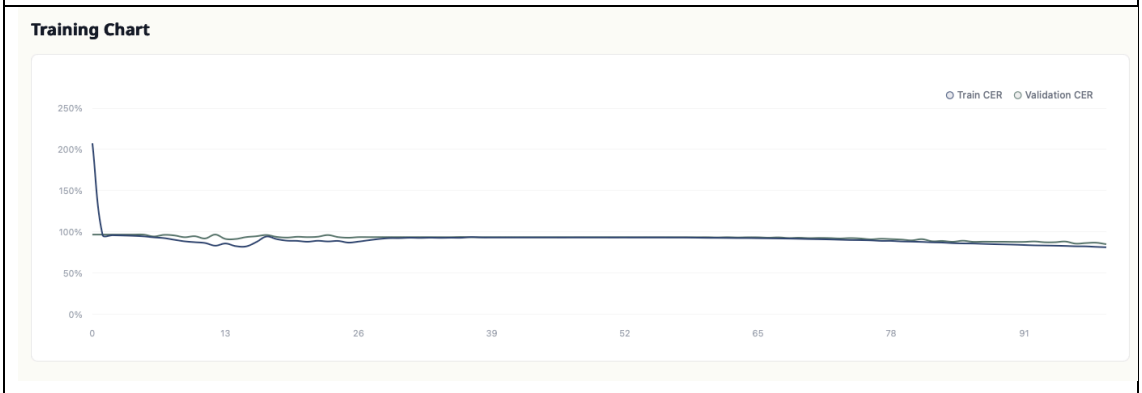
Fuente: elaboración propia a partir de los datos de la plataforma Transkribus.

El modelo Yongle Dadian 10 ha sido entrenado con una variante del corpus en la que se introdujeron espacios entre los caracteres chinos, con la hipótesis de que podría facilitar la segmentación. Este experimento resultó contraproducente, elevando el CER al 96,76%, por lo que se descartó esta estrategia.

Si tomamos como referencia el último modelo realizado, el Yongle Dadian 12 (figura 14), vemos que la gráfica de evolución del CER generada por Transkribus durante el entrenamiento muestra dos curvas: el *Train CER*, que mide el error sobre los datos de entrenamiento, y el *Validation CER*, que lo hace sobre un subconjunto de datos que el modelo no ha visto durante el aprendizaje. En un entrenamiento satisfactorio, ambas curvas descienden progresivamente a lo largo de los *epochs*. Como referencia orientativa, un CER inferior al 5% indica un modelo de alta precisión; entre el 10 y el 15% el modelo es utilizable con revisión manual; a partir del 20-30% los errores son ya muy numerosos; y valores superiores al 90% indican que el modelo prácticamente no está aprendiendo (Transkribus, 2025).

Figura 14

Gráfica de evolución del CER del modelo Yongle Dadian 12.



Fuente: elaboración propia a partir de Transkribus.

En los modelos entrenados en el presente trabajo, ambas curvas permanecieron estancadas en torno al 90-93% durante la mayor parte del proceso. Este patrón apunta a varias causas concurrentes: volumen insuficiente de datos de entrenamiento, segmentación incorrecta de las líneas de texto por la orientación

vertical del corpus, y ausencia de un modelo base adecuado para la escritura objetivo. Las tres causas están presentes en este caso y se analizan con detalle en el siguiente apartado.

4. Resultados y análisis

4.1. Evolución del CER a lo largo del entrenamiento

El indicador principal utilizado para evaluar el rendimiento de los modelos es el *Character Error Rate* (CER), que mide el porcentaje de caracteres incorrectamente reconocidos respecto al total de caracteres del texto de referencia. Un CER del 0% indicaría una transcripción perfecta, mientras que un CER del 100% implica que el modelo no reconoce ningún carácter correctamente.

Los resultados obtenidos muestran una progresión clara. El primer modelo, entrenado con únicamente 71 palabras de *ground truth*, alcanzó un CER del 100%, lo que confirma que con un volumen de datos tan reducido el modelo no es capaz de generalizar ningún patrón de reconocimiento. A medida que se incrementó el corpus de entrenamiento, el error fue disminuyendo progresivamente: con 315-329 palabras el CER se situó en torno al 93%; con 659-820 palabras bajó al 92%; y con 975-1.011 palabras se alcanzó el mejor resultado registrado, un CER del 82,67%.

Sin embargo, estos valores siguen siendo extremadamente elevados en comparación con los modelos disponibles en Transkribus para escrituras latinas. Como señalamos en §2.4.3, los CER de los modelos de referencia para escrituras occidentales oscilan entre el 2% y el 8% para colecciones de tamaño similar. La brecha es aún más evidente si la comparamos con el 91% de acierto (CER ~9%) que Cuéllar (2023) alcanzó con manuscritos españoles del Siglo de Oro, o con el CER del 0,79% que reporta Ayuso García (2022) para el modelo Noscemus GM entrenado con más de 541.000 palabras (p. 153).

Esta brecha de rendimiento no es atribuible únicamente a la falta de datos de entrenamiento, sino que responde a factores estructurales que se analizan en el apartado siguiente.

4.2. Análisis cualitativo de los errores

Más allá de las métricas cuantitativas, el análisis de los resultados de transcripción automática sobre páginas concretas del corpus revela patrones de error de gran interés analítico.

El error más visible y determinante es el de la sobresegmentación. Al aplicar el modelo sobre páginas no incluidas en el entrenamiento, el analizador de *layout* divide las columnas de texto en decenas de fragmentos mínimos. En el ejemplo analizado anteriormente (colección Yongle Dadian, página p004.jpg), el modelo generó 100 líneas segmentadas en una página que el *ground truth* contemplaba 16 columnas. Esta segmentación incorrecta invalida el reconocimiento posterior: aunque el modelo HTR identifique algún carácter aislado, la correspondencia entre las líneas segmentadas y las columnas reales del texto es tan inexacta que el texto resultante carece de coherencia.

El análisis de la transcripción generada por el mejor modelo (CER 82,67%) sobre esa misma página muestra que, de las 100 líneas segmentadas, únicamente dos contenían caracteres reconocidos: 齋 (línea 36) y 以 (línea 42). El resto de las líneas aparecen vacías. Este resultado ilustra de forma directa la causa del CER elevado: el modelo no es capaz de asociar los patrones gráficos de los caracteres chinos con una representación codificada.

Un segundo tipo de error observado son los artefactos en el propio *ground truth*. En el análisis de la página p012.jpg, la línea 11 de la transcripción manual muestra la secuencia «胡針屏优真狀 L 影象光», donde la letra latina L aparece intercalada en medio de caracteres chinos. Este tipo de artefacto, probablemente introducido inadvertidamente durante la anotación manual, constituye ruido en los datos de entrenamiento y puede haber afectado negativamente al proceso de aprendizaje del modelo.

Finalmente, se observa una correlación positiva clara entre el volumen del *ground truth* y la reducción del CER, aunque la velocidad de mejora es considerablemente más lenta que la reportada en trabajos equivalentes con

escrituras latinas. Esto es consistente con los hallazgos de Wang et al. (2012), quienes documentaron que el reconocimiento automático del chino manuscrito se beneficia de la adaptación de los modelos de lenguaje, pero presenta límites estructurales relacionados con la complejidad gráfica y la ausencia de separación entre caracteres (p. 267).

4.3. Discusión: viabilidad del HTR para el chino tradicional manuscrito

Los resultados obtenidos permiten extraer conclusiones relevantes sobre la viabilidad actual de las herramientas HTR para la transcripción automática de manuscritos chinos tradicionales, así como sobre las condiciones necesarias para mejorar los resultados en proyectos futuros.

Por un lado, los datos confirman que el entrenamiento supervisado mejora progresivamente el rendimiento del modelo: a mayor volumen de *ground truth*, menor CER. Esta tendencia es coherente con los hallazgos de la literatura especializada. Bazzaco (2020) documenta, en un contexto de escritura gótica española, cómo el proceso iterativo de corrección y reentrenamiento permite mejorar gradualmente la precisión del modelo, y Cuéllar (2023) alcanza resultados del 91-97% de acierto en manuscritos e impresos del Siglo de Oro tras el entrenamiento con decenas de miles de palabras. La misma lógica se aplica al presente caso, aunque el punto de partida es considerablemente menos favorable.

Por otro lado, los resultados ponen de relieve una limitación estructural de las plataformas HTR actuales: su diseño y sus modelos base están orientados hacia escrituras occidentales. Esta asimetría, que hemos denominado eurocentrismo tecnológico, tiene consecuencias directas para la investigación en humanidades digitales sobre lenguas y escrituras no occidentales. Como señalan Bazzaco et al. (2022), la figura del humanista "ha quedado inevitablemente fuera de la ecuación" en el proceso de digitalización masiva del patrimonio textual global (p. 71), y el presente trabajo constituye un ejemplo concreto de esta exclusión: para transcribir el chino clásico manuscrito con herramientas HTR no solo es necesario construir el

modelo desde cero, sino que además hay que superar limitaciones de diseño previas relacionadas con la orientación del texto.

Dicho esto, el presente trabajo también demuestra que la aplicación de herramientas HTR al patrimonio manuscrito chino no es inviable, sino que requiere condiciones específicas: un volumen significativamente mayor de *ground truth* que el empleado aquí, el desarrollo o adaptación de modelos de *baseline* para texto vertical, y posiblemente la colaboración con equipos especializados tanto en sinología, como en informática. En este sentido, la aportación principal de este trabajo no ha residido en haber alcanzado un CER bajo, sino en haber documentado el proceso, los obstáculos y las condiciones de partida de un proyecto de este tipo, contribuyendo así, a construir la infraestructura de conocimiento y datos, que proyectos futuros necesitarán.

5. Conclusiones

El presente trabajo ha tenido como objetivo principal el desarrollo de un primer modelo de transcripción automática aplicado a la colección manuscrita Yongle Dadian, con el fin de evaluar la viabilidad de las herramientas HTR para el chino clásico manuscrito. A partir del trabajo realizado, es posible formular las siguientes conclusiones.

En primer lugar, la aplicación de Transkribus al chino clásico manuscrito es técnicamente posible, pero enfrenta obstáculos estructurales significativos. El principal de ellos no es la complejidad de los caracteres tradicionales chinos en sí, sino la inadecuación del modelo de segmentación (*layout analysis*) a la orientación vertical del texto. Este problema, de naturaleza previa al reconocimiento de caracteres, limita drásticamente los resultados que pueden obtenerse con las herramientas actuales sin intervención manual intensiva.

En segundo lugar, los resultados cuantitativos confirman que el entrenamiento supervisado produce mejoras progresivas: el CER descendió del 100% inicial hasta el 82,67% con un corpus de 975 palabras de *ground truth*. Sin embargo, esta tasa de error sigue siendo inaceptablemente alta para usos académicos directos, y la comparación con modelos equivalentes para escrituras latinas evidencia la brecha tecnológica que separa las escrituras occidentales de las orientales en el ecosistema de herramientas HTR disponibles actualmente.

En tercer lugar, el trabajo ha permitido documentar y conceptualizar el problema del eurocentrismo tecnológico en el campo de las humanidades digitales. Las plataformas HTR actuales, incluyendo Transkribus, asumen implícitamente convenciones propias de la escritura occidental —dirección horizontal, inventario limitado de caracteres, separación clara entre palabras— que no son universales. Por ello, su aplicación a escrituras como el chino clásico presenta importantes limitaciones, lo que evidencia la necesidad de crear herramientas adaptadas a este tipo de patrimonio documental.

En cuarto lugar, el presente trabajo contribuye, de forma modesta pero concreta, a la construcción de la infraestructura necesaria para proyectos de transcripción automática del chino clásico: el corpus de *ground truth* elaborado, con 52 páginas del Yongle Dadian con transcripción manual, constituye un recurso que puede ser ampliado y reutilizado en investigaciones futuras, y los modelos entrenados ofrecen un punto de partida sobre el que continuar trabajando.

En cuanto a los objetivos planteados en la introducción, puede afirmarse que han sido alcanzados en su totalidad. Se ha descrito el funcionamiento de Transkribus y sus técnicas de aprendizaje automático; se ha elaborado un corpus de *ground truth* mediante la transcripción manual de 52 páginas; se han entrenado y evaluado 13 modelos HTR documentando su evolución; y se ha reflexionado críticamente sobre las limitaciones tecnológicas y el eurocentrismo de las herramientas disponibles. La hipótesis de trabajo, que es posible entrenar un modelo HTR para chino clásico manuscrito con Transkribus, aunque con limitaciones significativas, ha quedado corroborada.

Como líneas de investigación futuras, sería especialmente relevante explorar tres vías. La primera es el desarrollo o la adaptación de modelos de *baseline* específicos para texto vertical chino, lo que requeriría una colaboración estrecha con los desarrolladores de Transkribus o el trabajo con herramientas alternativas como eScriptorium o Kraken, que ofrecen una mayor flexibilidad en la configuración de la dirección del texto. La segunda es la ampliación del corpus de *ground truth*, idealmente hasta las 5.000-10.000 palabras que los estudios de referencia señalan como el umbral mínimo para obtener resultados aceptables con HTR en escrituras complejas. La tercera es la comparación sistemática entre distintas herramientas HTR para el chino manuscrito, que permita identificar las fortalezas y debilidades de cada una, y orientar las decisiones metodológicas de futuros proyectos.

En definitiva, este trabajo abre un camino que otros podrán recorrer con mejores herramientas, más datos y, esperemos, la conciencia de que el patrimonio manuscrito de la humanidad no se limite a las fronteras de occidente.

Bibliografía

Bibliografía citada:

Ayuso García, M. (2022). 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, 151–173. https://doi.org/10.13136/2284-2667/1102

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

Bazzaco, S., Jiménez Ruiz, A. M., Martín Molares, M., & Torralba Ruberte, Á. (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

Cuéllar, Á. (2023). La inteligencia artificial al rescate del Siglo de Oro: transcripción y modernización automática de mil trescientos impresos y manuscritos teatrales. Hipogrifo, 11(1), 101–115. https://doi.org/10.13035/H.2023.11.01.08

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., … Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. Journal of Documentation, 75(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114

Prado Fonts, C., & Martínez Robles, D. (2008). Narrativas chinas: ficciones y otras formas de no-literatura. Editorial UOC.

Shen, X. (100 d.C.). Shuowen jiezi (說文解字). Chang'an.

Su, T. (2013). Chinese handwriting recognition: An algorithmic perspective. Springer. https://doi.org/10.1007/978-3-642-31812-2

Transkribus. (2025). Training text recognition models with Transkribus [Material de formación]. READ-COOP.

Wang, Q.-F., Yin, F., & Liu, C.-L. (2012). Improving handwritten Chinese text recognition by unsupervised language model adaptation. En 2012 10th IAPR International Workshop on Document Analysis Systems (pp. 267–271). IEEE.

Wilkinson, E. (2013). Chinese history: A new manual. Harvard University Asia Center.

Xie, J., Yao, G., Gao, G., & Sima, G. (s.f.). Yongle da dian [Reproducción digital de la copia Jiajing, Ming Jiajing 41 nian–Ming Longqing yuan nian, 1562–1567]. Library of Congress, Chinese Rare Book Collection. https://www.loc.gov/resource/asianyongle1403.2018433742/

Bibliografía consultada:

Hill, N. W. (2015). Proposal for a transcription of Chinese characters in the study of early Chinese language and literature. Bulletin of Chinese Linguistics, 8(1), 1–24. https://doi.org/10.1163/2405478X-00801003

Kiessling, B. (2019). Kraken – an universal text recognizer for the humanities. En Digital Humanities 2019: Conference Abstracts. Alliance of Digital Humanities Organizations. https://dev.clariah.nl/files/dh2019/boa/0673.html

Real Academia Española. (s.f.) En Diccionario de la lengua española. Recuperado el 2 de febrero de 2026. https://www.rae.es

Schirokauer, C., & Brown, M. (2006). Breve historia de la civilización china. Cengage Learning.

Webgrafía y vídeos:

Transkribus. (2025, 10 de diciembre). Advanced Model Training in Transkribus [Vídeo]. YouTube. https://www.youtube.com/watch?v=x9xcjbR8vmQ

Transkribus. (s.f.). Baseline Models & Complex Layouts Webinar [Vídeo]. YouTube. https://www.youtube.com/watch?v=igseuPPfAdU

Transkribus. (s.f.). Training a custom Transkribus model [Vídeo]. YouTube. https://www.youtube.com/watch?v=DZtjK3DMa3A

Transkribus. (s.f.). Getting started with Transkribus in 90 seconds [Vídeo]. YouTube. https://www.youtube.com/watch?v=dvpxMdAskVI

READ-COOP. (s.f.). Automatic layout recognition. Transkribus Help Center. https://help.transkribus.org/automatic-layout-recognition

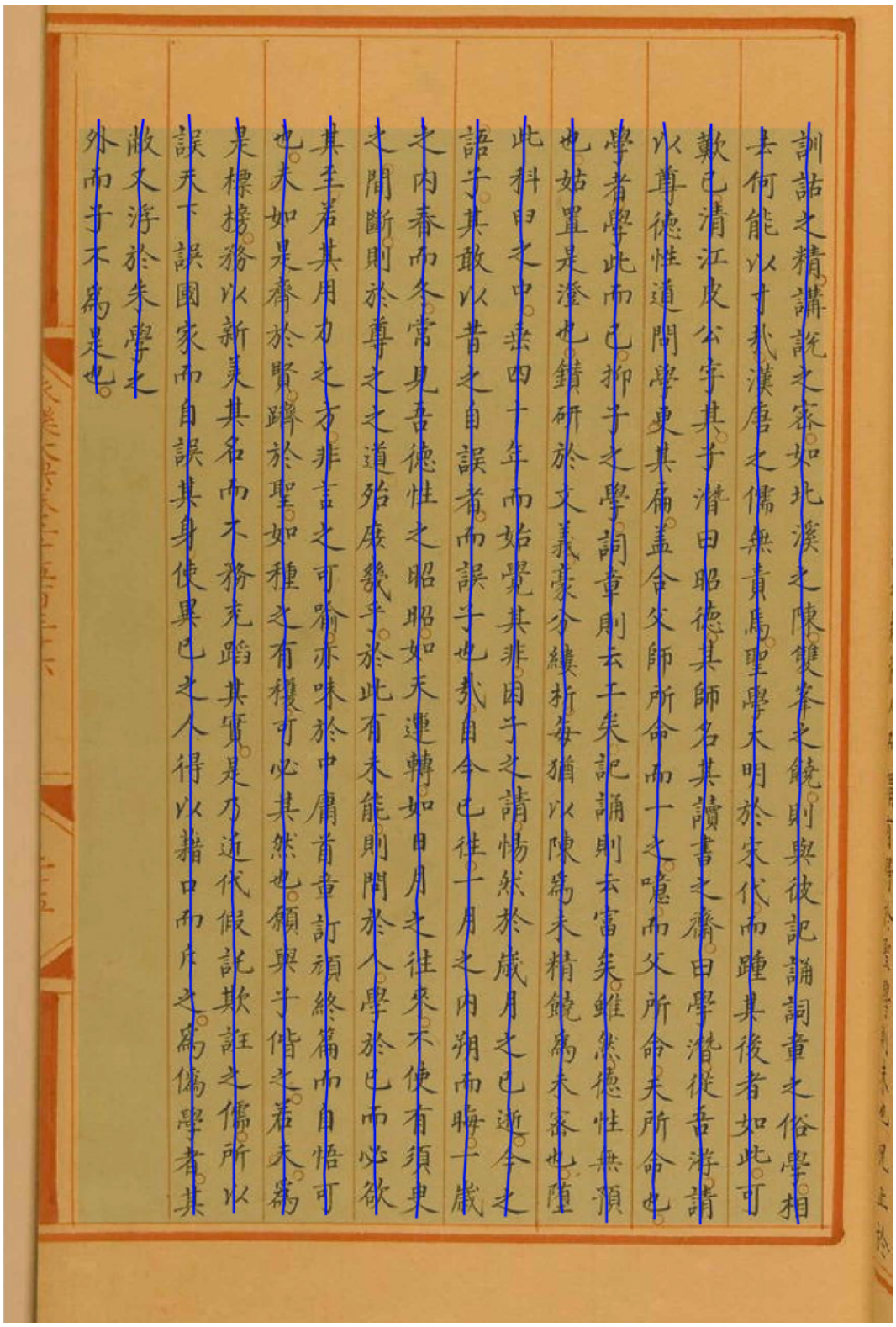
READ-COOP. (s.f.). Training text recognition models. Transkribus Help Center. https://help.transkribus.org/training-text-recognition-models

READ-COOP. (s.f.). Advanced layout configuration settings. Transkribus Help Center. https://help.transkribus.org/advanced-layout-configuration-settings

READ-COOP. (s.f.). Baseline models. Transkribus Help Center. https://help.transkribus.org/baselines-models

Anexos

Anexo 1. Yongle Dadian Baseline Model



Anexo 2. Historial de modelos entrenados en Transkribus

Anexo 2 a. Tabla de los modelos *baseline* con sus métricas

MODEL TITLE	TASK	LINES	LOSS	ID	
Yongle Dadian Baseline 6		414	6.60%	568449	☆
Yongle Dadian Baseline 5		321	58.53%	565689	☆
Yongle Dadian Baseline 4		321	46.76%	565669	☆
Yongle Dadian Baseline 3		80	59.74%	550817	☆
Yongle Dadian Baseline 2		64	41.57%	549337	☆
Yongle Dadian Baseline 1		64	41.00%	548137	☆

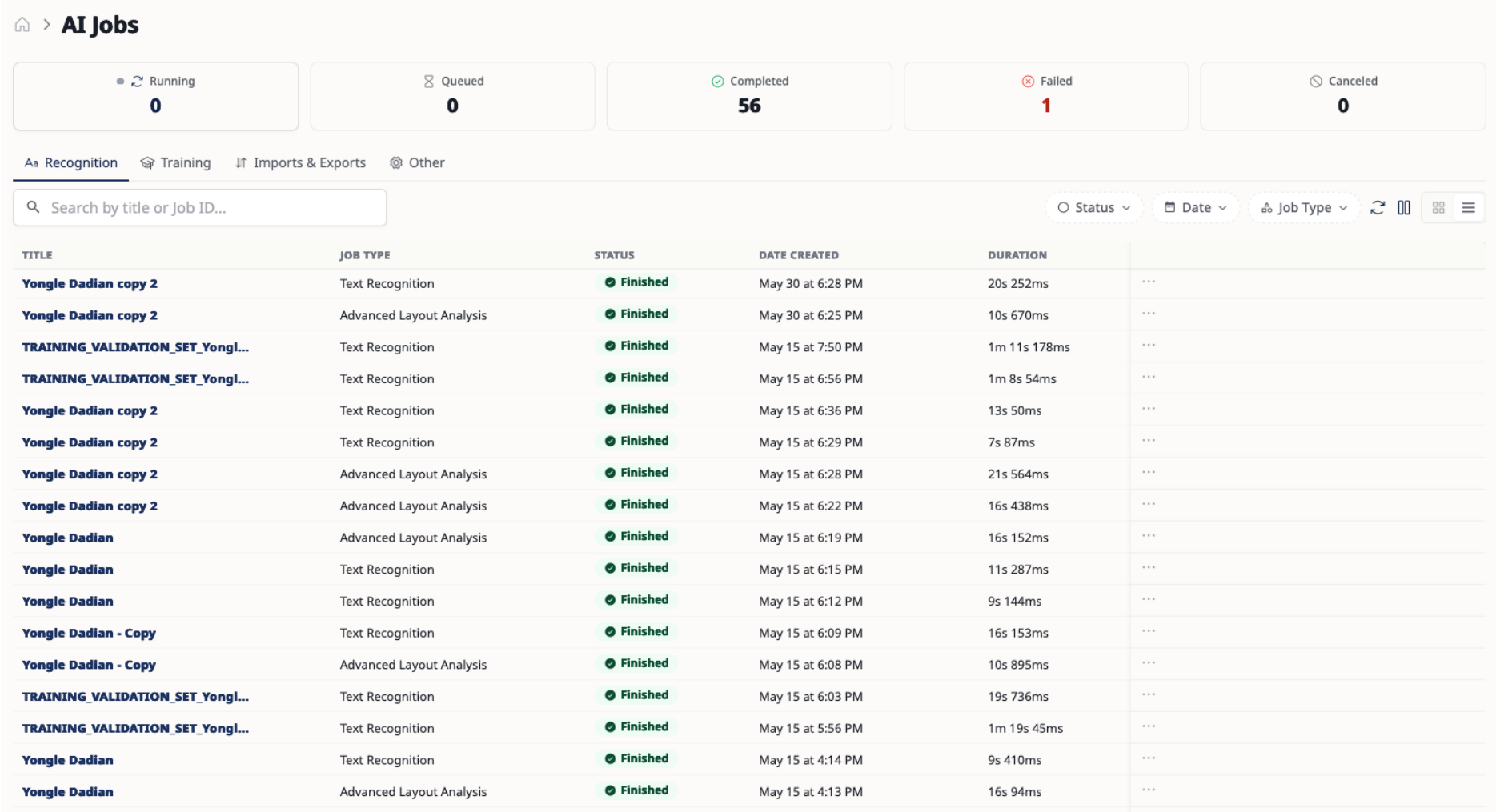
Anexo 2 b. Tabla de los modelos HTR con sus métricas

MODEL TITLE	TASK	SIZE	ACCURACY	ID	
Yongle Dadian 12 Chinese		1 011 words	CER 85.18%	568909	☆
Yongle Dadian 11 Chinese		1 011 words	CER 93.64%	568889	☆
Yongle Dadian 10 Chinese		2 079 words	CER 96.76%	568849	☆
Yongle Dadian 9 Chinese		975 words	CER 82.67%	568789	☆
Yongle Dadian 8 Chinese		820 words	CER 92.04%	568689	☆
Yongle Dadian 7 Chinese		659 words	CER 92.38%	568609	☆
Yongle Dadian 6 Chinese		329 words	CER 93.26%	565749	☆
Yongle Dadian 5 Chinese		329 words	CER 93.26%	565729	☆
Yongle Dadian 4 Chinese		329 words	CER 100.00%	565709	☆
Yongle Dadian 3 Chinese		315 words	CER 93.11%	565649	☆
Yongle Dadian 2 Chinese		71 words	CER 100.00%	564969	☆
Yongle Dadian 1 Chinese		71 words	CER 100.00%	518601	☆

Anexo 3. 52 páginas transcritas manualmente para entrenar el modelo HTR de la colección



Anexo 4. Historial de AI Jobs en Transkribus



AI Jobs

Running	Queued	Completed	Failed	Canceled
0	0	56	1	0

Recognition | Training | Imports & Exports | Other

Search by title or Job ID...

Status | Date | Job Type

TITLE	JOB TYPE	STATUS	DATE CREATED	DURATION
Yongle Dadian copy 2	Text Recognition	Finished	May 30 at 6:28 PM	20s 252ms
Yongle Dadian copy 2	Advanced Layout Analysis	Finished	May 30 at 6:25 PM	10s 670ms
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 15 at 7:50 PM	1m 11s 178ms
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 15 at 6:56 PM	1m 8s 54ms
Yongle Dadian copy 2	Text Recognition	Finished	May 15 at 6:36 PM	13s 50ms
Yongle Dadian copy 2	Text Recognition	Finished	May 15 at 6:29 PM	7s 87ms
Yongle Dadian copy 2	Advanced Layout Analysis	Finished	May 15 at 6:28 PM	21s 564ms
Yongle Dadian copy 2	Advanced Layout Analysis	Finished	May 15 at 6:22 PM	16s 438ms
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 6:19 PM	16s 152ms
Yongle Dadian	Text Recognition	Finished	May 15 at 6:15 PM	11s 287ms
Yongle Dadian	Text Recognition	Finished	May 15 at 6:12 PM	9s 144ms
Yongle Dadian - Copy	Text Recognition	Finished	May 15 at 6:09 PM	16s 153ms
Yongle Dadian - Copy	Advanced Layout Analysis	Finished	May 15 at 6:08 PM	10s 895ms
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 15 at 6:03 PM	19s 736ms
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 15 at 5:56 PM	1m 19s 45ms
Yongle Dadian	Text Recognition	Finished	May 15 at 4:14 PM	9s 410ms
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 4:13 PM	16s 94ms

TITLE	JOB TYPE	STATUS	DATE CREATED	DURATION	
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 4:12 PM	10s 798ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 4:11 PM	10s 924ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 4:10 PM	10s 697ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 4:08 PM	12s 90ms	···
Yongle Dadian	Text Recognition	Finished	May 15 at 4:00 PM	14s 186ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 3:58 PM	15s 917ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 3:57 PM	10s 649ms	···
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 15 at 3:54 PM	1m 12s 463ms	···
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 15 at 2:23 PM	54s 421ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 12:16 PM	15s 914ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 12:15 PM	16s 165ms	···
YongleDadian - Copy	Text Recognition	Finished	May 15 at 11:26 AM	15s 906ms	···
YongleDadian - Copy	Advanced Layout Analysis	Finished	May 15 at 11:25 AM	10s 894ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 15 at 7:47 AM	15s 959ms	···
Yongle Dadian	Text Recognition	Finished	May 14 at 12:08 PM	14s 351ms	···
Yongle Dadian	Text Recognition	Finished	May 14 at 12:06 PM	14s 286ms	···
Yongle Dadian	Text Recognition	Finished	May 14 at 11:56 AM	20s 83ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 14 at 11:55 AM	13s 218ms	···
Yongle Dadian	Advanced Layout Analysis	Finished	May 14 at 11:22 AM	15s 883ms	···
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 11 at 1:19 AM	29s 681ms	···
YongleDadian - Copy	Text Recognition	Finished	May 11 at 12:54 AM	17s 191ms	···

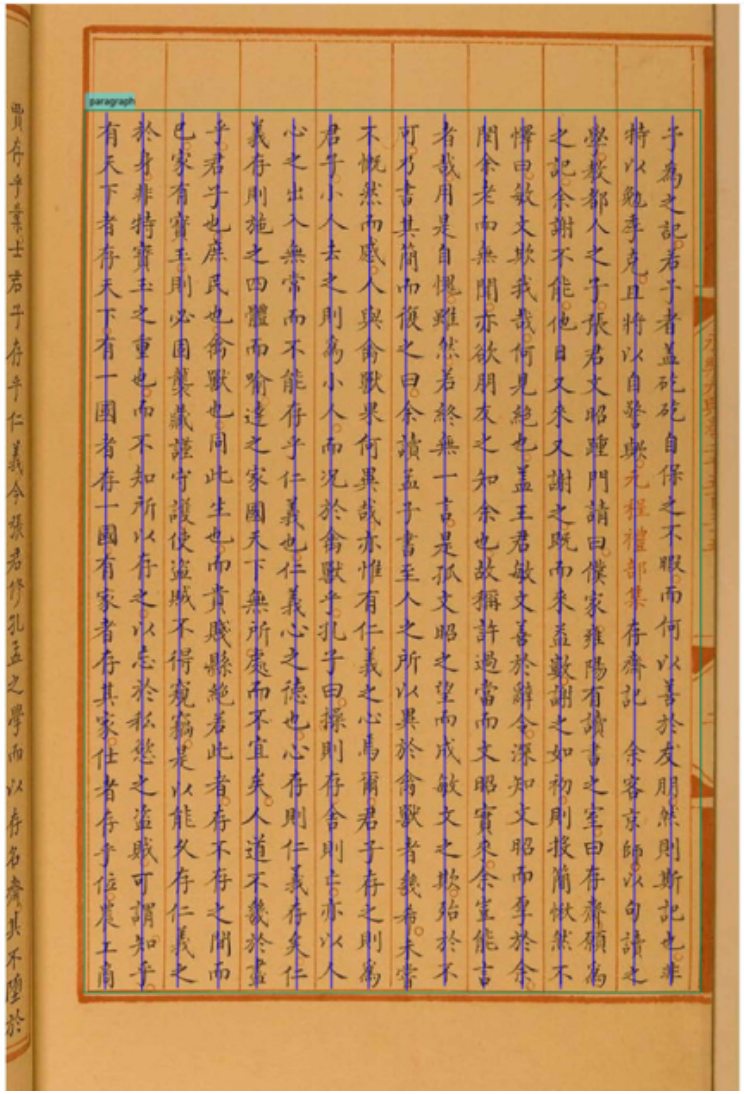
YongleDadian - Copy	Advanced Layout Analysis	Finished	May 10 at 11:00 PM	10s 696ms	···
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 10 at 10:35 PM	28s 300ms	···
YongleDadian	Text Recognition	Finished	May 10 at 5:49 PM	40s 158ms	···
YongleDadian	Advanced Layout Analysis	Finished	May 10 at 5:45 PM	16s 109ms	···
YongleDadian	Text Recognition	Finished	May 10 at 5:37 PM	19s 97ms	···
YongleDadian	Text Recognition	Finished	May 10 at 5:29 PM	17s 823ms	···
YongleDadian	Text Recognition	Finished	May 10 at 5:17 PM	19s 499ms	···
YongleDadian	Advanced Layout Analysis	Finished	May 10 at 5:15 PM	16s 209ms	···
TRAINING_VALIDATION_SET_Yongl...	Text Recognition	Finished	May 9 at 11:50 AM	14s 859ms	···
Deleted Document	Text Recognition	Finished	Apr 14 at 3:41 PM	30s 432ms	···
YongleDadian	Text Recognition	Finished	Feb 24 at 4:53 PM	19s 451ms	···
YongleDadian	Advanced Layout Analysis	Finished	Feb 24 at 4:51 PM	23s 87ms	···
Deleted Document	Text Recognition	Finished	Feb 23 at 11:40 PM	23s 734ms	···
Deleted Document	Text Recognition	Finished	Dec 13, 2025 at 2:42 PM	13s 137ms	···

Anexo 5. Yongle Dadian HTR Model

Aquí se encuentra una muestra de 3 páginas transcritas manualmente para entrenar el modelo HTR. Las transcripciones completas de las 52 páginas que componen el *ground truth* del presente trabajo están disponibles en la plataforma Transkribus bajo la colección Yongle Dadian 2 (ID de colección: 2404298).

Yongle Dadian 2 / Yongle Dadian / **p006.jpg** Ground Truth

Process with AI



Region1 paragraph

1 子為之記。若子者盖矻矻自保之不暇而何以善於友朋然則斯記也。非
2 物以勉李克。且將以自警歟。元稹禮部集存齋記　余客京師。以句讀之
3 學。教都人之子。張君文昭踵門請曰。僕家雍陽有讀書之室。曰存齋願為
4 之記。余謝不能。他日又來又謝之。既而來益數謝之如初。則投簡愀然不
5 悌曰。敏文欺我哉。何見絕也。盖王君敏文善於辭令。深知文昭而厚於余
6 閉余老而無聞。亦欲朋友之知余也。故稱許過當而文昭實來。余宜能言
7 者哉。用是自愧雖然若終無一言。是孤文昭之望而成敏文之欺。殆於不
8 可。乃書其簡而復之曰。余讀孟子書至人之所以異於禽獸者幾希。未嘗
9 不慨然而感。人與禽獸果何異哉亦惟有仁義之心焉爾。君子存之則為
10 君子。小人去之則為小人。而況於禽獸乎。孔子曰。操則存含則亡。亦以人
11 心之出入無常而不能存乎仁義也。仁義心之德也。心存則仁義存矣。仁
12 義存則施之四體而喻達之家國天下無所處而不宜矣。人道不幾於盡
13 乎。君子也。庶民也禽獸也。同此生也。而貴與賤絕者此者。存不存之間而
14 已。家有寶王。則必固襲藏謹守護使盜賊不得窺竊。是以能火存仁義之
15 於身。非特寶玉之重也而 不知所以存之。以忘於私慾之盜賊。可謂知乎。
16 有天下者存天下。有一國者存一國。有家者存其家。仕者存乎。位農工商

Process with AI



Region1 paragraph

宋姚成一雪坡集養齋記　子性質苦狷急。遇物觸即莫能受。甚病之思
自藥。一日靜坐得一字曰養。榜名齋。自警也。養之義大矣哉。養則於。不
養則溢。養則充。不養則簸。德以養紙。才以養裕。學問以養涵深。器識以養
廣大。孰為養蓋有要道。養其大者為大人。養拭辣而含捂攪。養一指而失
肩背。非養之善。養之善莫着養氣。心者氣之主。莫着養心。心之不能養。怎
與慾敗之也。忿不廖吾戈矛。態未淨吾陷穽。忿必懲慾必窒。然後可以持
志而養氣。氣善養則充且私失。充且弘則所養正而聖功寓矣。豈止藥吾
性質之病而已哉。雖然。懲忿窒慾養之善也。吳而懲。奚而室。曰。獨不聞周
夫子之言乎。澹刖慾
心平。和則躁心釋
宜養齋
元蕭勳齋集直養齋銘　萬物斤資
均乎一氣。孰其尸之。日維上帝帝命人
極。健順五常。氣則偕行。業稟莫量。順理為直。得養則大。襲正助忘。匪理斯
害。凡中有主。所向無前。拔山蓋世。古亦有馬。敬義夾持。俯仰無愧。上下同
流。勇何足議。不憂不懼。不疑所行。何物事功。據吾靈局。矯椿裴君。炎黃其
學。從事于斯。受說先覺。緊余小子。顏學未能。敢銘君齋。窴維自銘。
適齋
宋華鎮雲溪居士集道州錄事廳適齋記　元右壬申歲。余來
為營道郡督郵。越明年冬十月。葺舍館之西頹廡。設戶牖以為



Region1 paragraph

室。既成目之為適。齋客有過而疑之。日吾聞適當也物有當於心而無迕於意。然後適。差夫資雅材。抱道街遊乎可用之世。行既久而步不進。日向晚而道則遠。主郡簿書於放逐傲人之地。俛眉高拱。與衆進退庭下。捷步不異昏史組幾不足以為六尺之華廩捕不足飽數口之腹續短裡薄。關構核桶。方風兩飄漂之慮。曾何物之有當而無迕也而謂之適耶。予起而。應之曰。客聞求廉氏之二子乎。昔求氏與廉氏同里開也。求氏家富而官達。曰患失其所有而務得無已。終年戚戚以憂。廉氏農圃也食其力。僅免其凍餒不羨涯。分之外。終年熙熙以遊君吾子之所謂。則高位孕禄。金朱華屋為可適。而早官薄廩。草圭之門青緣之品。爲不可適已。吾恐特未定也。吾之所謂適者適於已之謂也。非適人之所適也。人之所適者物也。物不足則有時而不適。已之所適者道也。道無往而不在。則已無人而不適。世之道不足者。視名位禄養車服室宇之麗。切切然以為已之所宜保而不得去沒其樂也。乃所以為苦歟。人之所。惡夫行獄牛繫者。謂其錮繁而不可去也。今貴富之人。無道以處之而累於物。則高位厚禄華眼美宗。曾何異犴獄牢緊錮繁之地。耶至於欣捧一艘。折腰斗升。所養在中浩然自得者。曾不知甲飛之寫勞。數粒之為薄。搶榆挫地。時止時行。則其受