La Inteligencia Artificial al rescate del Siglo de Oro: transcripción y modernización automática de mil trescientos impresos y manuscritos teatrales

Artificial Intelligence to the Rescue of the Spanish Golden Age: Automatic Transcription and Modernization of One Thousand Three Hundred Theatrical Prints and Manuscripts

Álvaro Cuéllar
https://orcid.org/0000-0002-9934-6321
Universität Wien
AUSTRIA
alvaro.cuellar.gonzalez@univie.ac.at

[*Hipogrifo*, (issn: 2328-1308), 11.1, 2023, pp. 101-115]
Recibido: 20-01-2023 / Aceptado: 20-02-2023
DOI: http://dx.doi.org/10.13035/H.2023.11.01.08

Resumen. Un elevado porcentaje de impresos y manuscritos teatrales del periodo aurisecular no ha sido nunca transcrito en un formato analógico ni, por supuesto, digital. Es imposible, por tanto, emplear estos documentos para realizar búsquedas de nuestro interés o para los valiosos análisis informáticos (estilometría, topic modelling, detección de sentimientos, etc.) que se están desarrollando

Este trabajo no hubiera sido posible sin la ayuda de Germán Vega García-Luengos (Universidad de Valladolid), quien rastreó un sinnúmero de fondos en busca de los documentos teatrales necesarios. El artículo se inscribe en el proyecto *Impresos sueltos del teatro antiguo español: base de datos integrada del teatro clásico español*, financiado por el Ministerio de Ciencia, Innovación y Universidades (Ref. PID2019-104045GA-C55) / AEI / 10.13039/501100011033) y Fondos Feder. También se adscribe al proyecto *Sound and Meaning in Spanish Golden Age Literature* (FWF Austrian Science Fund 32563).

en los últimos años. Gracias a la Inteligencia Artificial (Transkribus) y técnicas de HTR (Handwritten Text Recognition) he entrenado tres modelos, públicos ya para la comunidad investigadora, capaces de transcribir y modernizar ortográficamente estos documentos de forma automática con un alto grado de precisión: alrededor del 97% de acierto en impresos y 91% en manuscritos. A través de estos modelos he podido procesar unas 1.300 obras teatrales contenidas en impresos y manuscritos procedentes de numerosas bibliotecas, archivos y otras fuentes digitalizadas. Las transcripciones resultantes forman ahora parte del proyecto ETSO, del buscador TEXORO y, además de suponer un avanzado punto de partida para la edición cuidada de los textos, cuentan por sí mismas con la calidad suficiente para ser sometidas a análisis estilométricos, los cuales están arrojando atribuciones autoriales de interés.

Palabras clave. Transcripción automática; HTR; teatro del Siglo de Oro; impresos; manuscritos, modernización ortográfica.

Abstract. A high percentage of theatrical prints and manuscripts from the aurisecular period have never been transcribed in an analogical or, of course, digital format. It is therefore impossible to use these documents to carry out searches of our interest or for the valuable computer analyses (stylometry, topic modelling, sentiment analysis, etc.) that have been developed in recent years. Thanks to Artificial Intelligence (Transkribus) and HTR (Handwritten Text Recognition) techniques, I have trained three models, already public for the research community, capable of transcribing and orthographically modernizing these documents automatically with a high degree of precision: around 97% of success in prints and 91% in manuscripts. Through these models I have been able to process some 1,300 theatrical plays contained in prints and manuscripts from numerous libraries, archives, and other digitized sources. The resulting transcripts are now part of the ETSO project, of the TEXORO search engine and, in addition to being an advanced starting point for careful editing of the texts, they themselves have sufficient quality to be subjected to stylometric analysis, which is yielding authorship attributions of interest.

Keywords. Automatic transcription; HTR; Spanish Golden Age Theatre; Prints; Manuscripts; Orthographic modernization.

1. Enfrentando un estancamiento

Entre los años 2017 y 2020, el proyecto *ETSO: Estilometría aplicada al teatro del Siglo* de Oro[1] (https://etso.es/) había conseguido reunir unas 1200 obras del periodo aurisecular en formato digital (texto plano). El fin principal de realizar esta recolección era el de someter los textos a pruebas estilométricas de autoría. Si contamos con un corpus de teatro aurisecular de un número suficiente de obras y dramaturgos, la estilometría es capaz, entre sus muchas posibilidades, de apuntar cuáles son los textos con un uso léxico más cercano al de nuestro interés. Estas

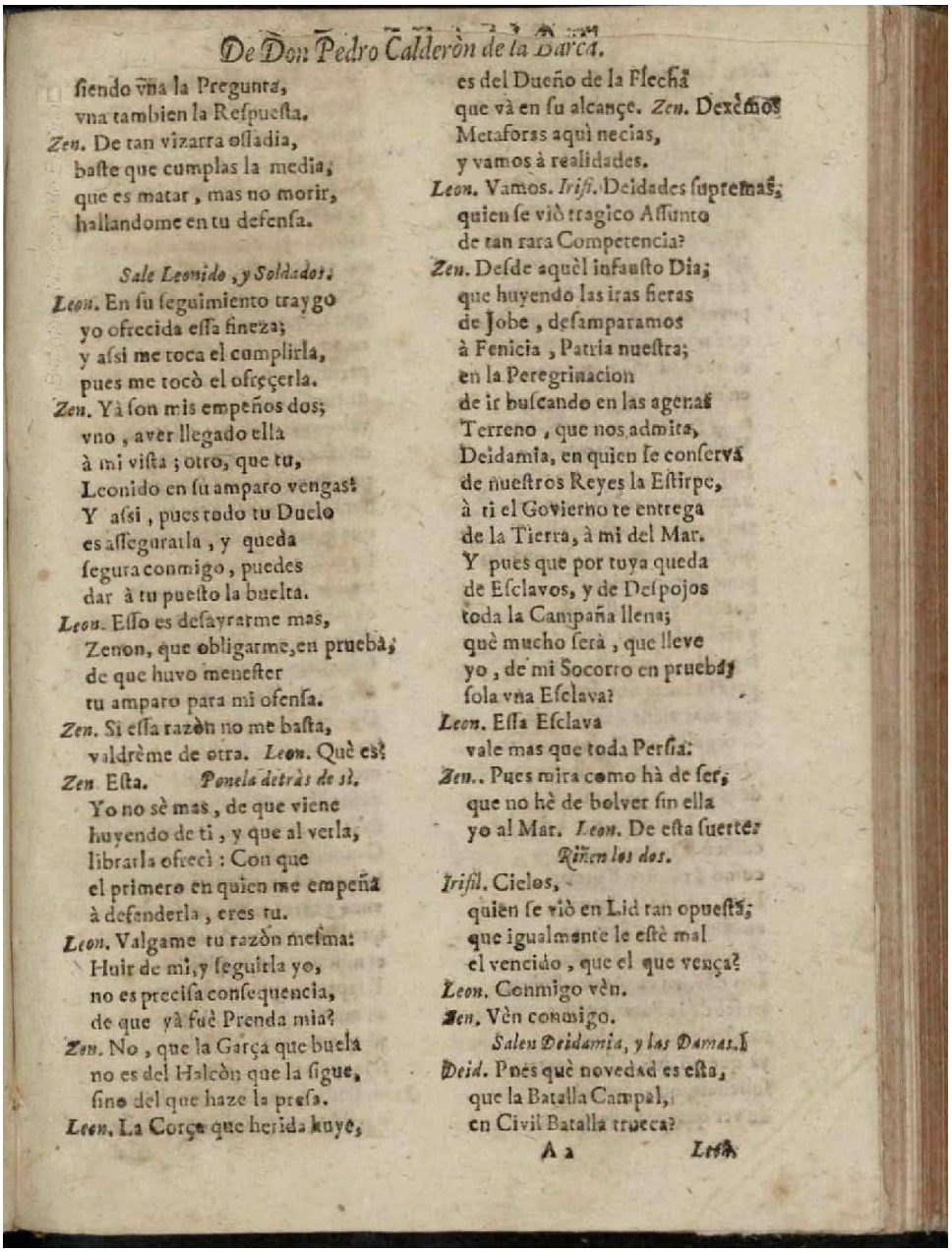
1. Cuéllar y Vega García-Luengos, 2017-2023.

obras suelen corresponder a las del autor de nuestra obra. Cuanto más numeroso sea el conjunto de piezas teatrales y dramaturgos, en general, mejores y más fiables serán estos análisis. El proyecto, por tanto, aspira idealmente a hacerse con todos los textos teatrales del Siglo de Oro que existen[2]. Las 1200 obras recolectadas entre 2017 y 2020, las cuales deben encontrarse en formato digital para poder ser tratadas informáticamente, se obtuvieron a través de numerosos portales abiertos al público, como AHCT, Artelope, Aula Biblioteca Mira de Amescua, BVMC, Moretianos, Teatro de Autores Portugueses do Séc. XVII, etc., así como por la colaboración de equipos de investigación como GRISO, ISTAE, Prolope, etc. y el envío particular de investigadores interesados en los resultados de los textos de los que disponían[3].

El proyecto, sin embargo, empezó a desacelerar por la dificultad de encontrar nuevas obras en un formato apropiado para su ingreso en el corpus. Se produjo un estancamiento por el que era cada vez más complicado que aparecieran nuevos textos o dramaturgos, más allá de las incorporaciones puntuales de la digitalización de alguna edición en papel o por el envío del esforzado trabajo de algún filólogo interesado en una obra concreta. Durante ese tiempo una idea sobrevolaba el proyecto: la gran cantidad de impresos y manuscritos —cientos o incluso miles— de teatro del Siglo de Oro que, aunque digitalizados y contenidos en valiosos portales como la BNE (Biblioteca Nacional de España), o la BVMC (Biblioteca Virtual Miguel de Cervantes), no podían emplearse para los análisis por no estar transcritos, es decir, el texto no podía ser leído ni entendido por el ordenador, por lo que era imposible someter las obras a los análisis estilométricos. Una miríada de piezas teatrales quedaba, por tanto, fuera de nuestro alcance. En la Figura 1 podemos ver un ejemplo de una página habitual de impreso del teatro del Siglo de Oro, en este caso de la obra *Duelos de amor y lealtad* del dramaturgo Calderón de la Barca. Como se puede comprobar, la tipografía, excepto alguna salvedad o característica típica en este tipo de documentos, es clara y legible para cualquier lector moderno, por lo que la transcripción sería sencilla de acometer. Sin embargo, los procesos informáticos habituales no están preparados para este tipo de textos. También podemos ver un ejemplo en la Figura 2 de unas páginas habituales de un manuscrito como este de la anónima *La francesa Laura*. Enseguida salta a la vista la complejidad de los manuscritos, en muchas ocasiones con una dificultad de lectura elevada o con fragmentos imposibles de transcribir hasta para el historiador o filólogo entrenado.

2. En estos momentos (finales de 2022) el proyecto cuenta con unas 2700 obras mayores (comedias y autos) y no parece probable poder aumentar este número mucho más de las 3000, aun con los procesos aquí descritos. Las obras supervivientes que componen nuestro teatro aurisecular, por tanto, no son tantas como podíamos imaginar.

3. Puede consultarse la procedencia de cada texto en https://etso.es/cetso.



De Don Pedro Calderòn de la Barca.

ſiendo vna la Pregunta,
vna tambien la Reſpueſta.
Zen. De tan vizarra oſſadia,
baſte que cumplas la media;
que es matar, mas no morir,
hallandome en tu defenſa.

Sale Leonido, y Soldados.

Leon. En ſu ſeguimiento traygo
yo ofrecida eſſa fineza;
y aſsi me toca el cumplirla,
pues me tocò el ofreçerla.
Zen. Yà ſon mis empeños dos;
vno, aver llegado ella
à mi viſta; otro, que tu,
Leonido en ſu amparo vengas:
Y aſsi, pues todo tu Duelo
es aſſeguratla, y queda
ſegura conmigo, puedes
dar à tu puesto la buelta.
Leon. Eſſo es deſayrarme mas,
Zenon, que obligarme, en prueba;
de que huvo menester
tu amparo para mi ofenſa.
Zen. Si eſſa razòn no me baſta,
valdrème de otra. *Leon.* Què es?
Zen. Eſta. *Ponela detràs de sì.*
Yo no sè mas, de que viene
huyendo de ti, y que al verla,
librarla ofrecì: Con que
el primero en quien me empeña
à defenderla, eres tu.
Leon. Valgame tu razòn meſma:
Huir de mi, y ſeguirla yo,
no es preciſa conſequencia,
de que yà fuè Prenda mia?
Zen. No, que la Garça que buela
no es del Halcòn que la ſigue,
ſino del que haze la preſa.
Leon. La Corça que herida huye,
es del Dueño de la Flecha
que và en ſu alcançe. *Zen.* Dexèmos
Metaforas aquì necias,
y vamos à realidades.
Leon. Vamos. *Iriſi.* Deidades ſupremas;
quien ſe viò tragico Aſſunto
de tan rara Competencia?
Zen. Deſde aquèl infauſto Dia;
que huyendo las iras fieras
de Jobe, deſamparamos
à Fenicia, Patria nueſtra;
en la Peregrinacion
de ir buſcando en las agenas
Terreno, que nos admita,
Deidamia, en quien ſe conſerva
de nueſtros Reyes la Eſtirpe,
à ti el Govierno te entrega
de la Tierra, à mi del Mar.
Y pues que por tuya queda
de Eſclavos, y de Deſpojos
toda la Campaña llena;
què mucho ſerà, que lleve
yo, de mi Socorro en prueba
ſola vna Eſclava?
Leon. Eſſa Eſclava
vale mas que toda Perſia:
Zen. Pues mira como hà de ſer;
que no hè de bolver ſin ella
yo al Mar. *Leon.* De eſta ſuerte:
Riñen los dos.
Iriſi. Cielos,
quien ſe viò en Lid tan opueſta;
que igualmente le eſtè mal
el vencido, que el que vença?
Leon. Conmigo vèn.
Zen. Vèn conmigo.
Salen Deidamia, y las Damas.
Deid. Pues què novedad es eſta,
que la Batalla Campal,
en Civil Batalla trueca?

Figura 1. Una página del impreso *Duelos de amor y lealtad* de Calderón de la Barca

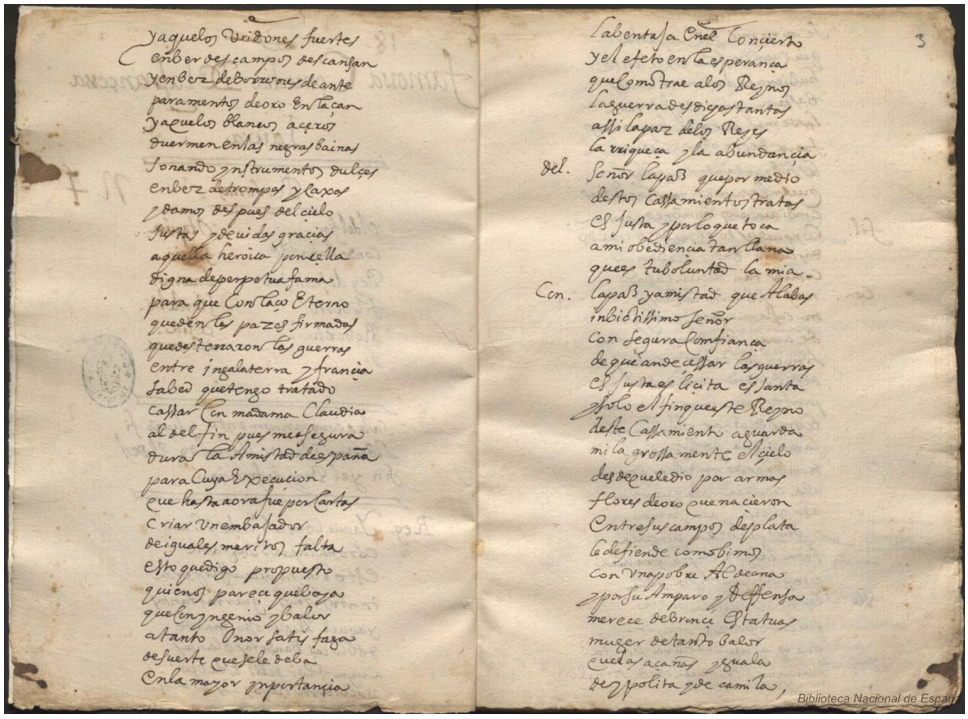


Figura 2. Unas páginas del manuscrito de *La francesa Laura*, anónima

Para intentar transcribir automáticamente documentos existe un proceso cotidiano conocido como OCR (Optical Character Recognition), que es el utilizado para la conversión digital de textos modernos con una elevada eficacia, pero que no sirve para nuestros propósitos. Los tipos de la imprenta no guardan relación con la tipografía empleada en la actualidad y los procesos de OCR, entrenados para reconocer tipografías modernas, no funcionan con el acierto adecuado. En el caso de los manuscritos los procesos tradicionales de OCR no consiguen apenas transcribir ni una palabra con acierto. Trabajar a partir de estas transcripciones automáticas, conllevaría, en definitiva, más esfuerzo que transcribir directamente desde cero, de forma manual.

Por otra parte, es necesario considerar el problemático asunto de la ortografía. Para nuestros intereses particulares, necesitamos que la ortografía de las obras esté modernizada, pues solo de esta manera pueden ser comparadas con el resto del corpus, cuya ortografía estaba modernizada a las normas actuales en todos los casos. Es decir, aunque el proceso de transcripción automática fuera perfecto y consiguiera transcribir las letras y las palabras de forma exacta, aun así, el texto no serviría para nuestros propósitos concretos, pues quedaría en un estadio en el

que no puede compararse con el resto de las obras del corpus. Tampoco existen, por el momento[4], modernizadores de ortografía automáticos para el español que cumplan nuestros propósitos.

En el proyecto ETSO, por tanto, enfrentábamos un estancamiento que parecía irresoluble. Cientos de impresos y manuscritos aguardaban digitalizados, pero no podían ser utilizados para los análisis por la incapacidad de los ordenadores para leerlos y convertirlos en formato digital apropiado para las pruebas.

2. Transkribus

A principios de 2020 descubrimos la herramienta Transkribus[5], la cual pretende ser una solución para la transcripción de manuscritos antiguos con un alto grado de acierto a través de *machine learning*. La herramienta, desarrollada por el grupo READ-COOP, cuenta con un potente sistema de Inteligencia Artificial para realizar el entrenamiento en el reconocimiento de texto y para, finalmente, acometer las transcripciones automáticas[6]. Se trata, además, de una herramienta con una interfaz amigable y de fácil aprendizaje, que resulta muy útil incluso para la transcripción manual de los textos.

Esta herramienta no ha sido pensada y creada en atención a una lengua y un alfabeto determinado, en el que se educa a la máquina a reconocer unas formas concretas. La herramienta, sin embargo, está orientada a ser un espacio libre para el reconocimiento de textos: no importa la lengua o el alfabeto empleado. El proceso básico 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 miles de palabras correctamente transcritas, la máquina es capaz de aprender y de transcribir correctamente las líneas del nuevo documento que se le ingrese.

Desafortunadamente, en ese momento no existía ningún modelo para el reconocimiento del español[7], por lo que fue necesario su construcción desde cero. Vea-

4. Javier de la Rosa, Jörg Lehmann y yo estamos trabajando en *The Modernisa Project: Orthographic Modernization of Spanish Golden Age Dramas with Language Models*, un proceso para modernizar la ortografía habitual en los Siglos de Oro de forma automática, con prometedores resultados.
5. Muehlberger *et al.*, 2019.
6. Cuando se realizó la mayor parte de los trabajos de transcripción comentados en este artículo, Transkribus era completamente gratuito. Ahora (finales de 2022) funciona con un sistema de pagos. Transcribir, por ejemplo, 10.000 páginas de manuscrito o 60.000 páginas de impreso cuesta 2.160 euros. Bien es cierto que ofrecen becas para proyectos y jóvenes investigadores.
7. Ahora (finales de 2022) existen también los modelos *Charlos V/Charles V*, *SpanishRedonda_XVI-XVII_extended* y *SpanishGothic_XV-XVI_extended*. Al abordar un nuevo documento, merece la pena probar todos los modelos existentes para comprobar cuál se ajusta mejor a nuestros propósitos. Como se puede apreciar en la bibliografía (Aranda García, Ayuso García, Bazzaco, Bazacco et al., Blasut y Fradejas Rueda), Transkribus ya se está empleando en distintos proyectos de interés.

mos, en primer lugar, cómo se realizó este entrenamiento para el reconocimiento de impresos y manuscritos para pasar, en las siguientes secciones, a su aplicación al conjunto del teatro del Siglo de Oro que no había sido transcrito con anterioridad.

3. Entrenamiento en el reconocimiento de impresos

El primer paso fue entrenar a la herramienta en el reconocimiento de impresos. Para ello transcribí a mano respetando la ortografía de la época un total de 74.129 palabras (16.024 líneas). El resultado es el que se puede apreciar en la Figura 3, generado por la propia herramienta, en la que se puede observar cómo la máquina aprende a través de realizar intentos (Epochs) y se perfecciona hasta alcanzar un CER[8] (Character Error Rate) del 0,91% en el set de validación. Por tanto, el 99,09% de los caracteres son certeramente reconocidos y transcritos, lo cual es una tasa de acierto muy elevada y útil para el tratamiento de este tipo de documentos. Este modelo, denominado *Spanish Golden Age Prints 1.0*[9], se encuentra disponible para la comunidad investigadora. Cualquiera puede emplearlo en sus investigaciones a través de la plataforma Transkribus y aplicarlo a los documentos de su interés.

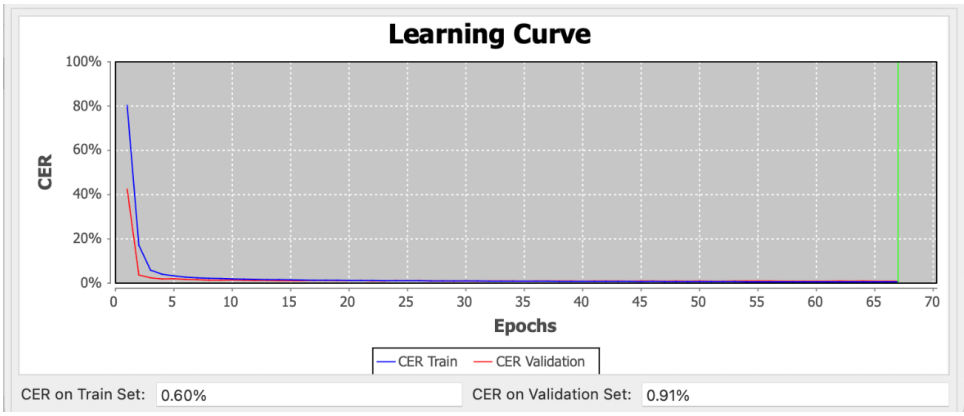


Figura 3. Entrenamiento del modelo *Spanish Golden Age Prints 1.0*.

Recordemos, sin embargo, que para los fines del proyecto ETSO de nada sirven los textos con ortografía sin modernizar, por lo que decidí entrenar a la máquina a través de textos con ortografía modernizada. Se trata de una decisión en extremo arriesgada, pues puede generar problemas en el reconocimiento de los caracteres. En una transcripción que respeta la ortografía de la época cada carácter se trans-

8. Conviene reflexionar y ser muy precavidos respecto a los CER obtenidos. Un modelo entrenado con un conjunto muy reducido y homogéneo de documentos puede conseguir CER casi perfectos, pero eso no significa que vaya a comportarse de forma adecuada al intentar transcribir un nuevo texto ligeramente ajeno a este grupo de entrenamiento. Por ello, según mi experiencia, considero que, en general, cuantas más palabras empleemos para el entrenamiento mejores serán los resultados, pues más léxico y contextos aprenderá la herramienta.

9. Cuéllar, 2021b.

cribe de la misma forma. Por ejemplo, si nos fijamos en el verbo haber: *auia* se transcribe como *auia, e hexo* como *e hexo*, etc. Sin embargo, ahora le vamos a demandar a la máquina que transcriba *auia* como *había* o *e hexo* como *he hecho*. No obstante, en otras ocasiones la *e* será la conjunción copulativa o la *x* será correcta en el español actual. Por tanto, ahora debe aprender en cada caso a reconocer contextualmente el tipo de palabra a la que se está enfrentando y tomar decisiones al respecto, no a realizar una traslación directa de las letras. Esto puede generar, a priori, un sinnúmero de problemas en nuestro proceso.

Conseguir este objetivo es mucho más complejo, por lo que, en vez de recurrir a transcripciones manuales emprendí un proceso masivo a través de la conexión de ediciones de un conjunto de textos con sus documentos de época correspondientes. Así, para el entrenamiento ya no tuve que transcribir manualmente cada línea y cada palabra, sino que conecté ediciones de las obras con los documentos de forma automática utilizando la función Text2Image de Transkribus, la cual ha sido clave en este proceso. Entrené a la máquina, pues, con la nada despreciable cantidad de 2.757.908 palabras (904.457 líneas), lo cual supone un número varios órdenes superior al necesario en el proceso anterior para conseguir un resultado adecuado. La curva de aprendizaje del entrenamiento fue la que puede apreciarse en la Figura 4, donde se ve como, lógicamente, el resultado es peor que con la transcripción que respetaba la ortografía, debido a lo complicado de la situación, pero se alcanza un CER (Character Error Rate) en el set de validación del 3,10%, por lo que, de cada 100 caracteres, 96,90% son transcritos y modernizados correctamente. Se trata de una tasa muy buena debido a la complejidad de la tarea requerida. El modelo, denominado *Spanish Golden Age Prints (Spelling modernization) 1.0*[10], también se encuentra a disposición de la comunidad investigadora a través de la plataforma Transkribus.

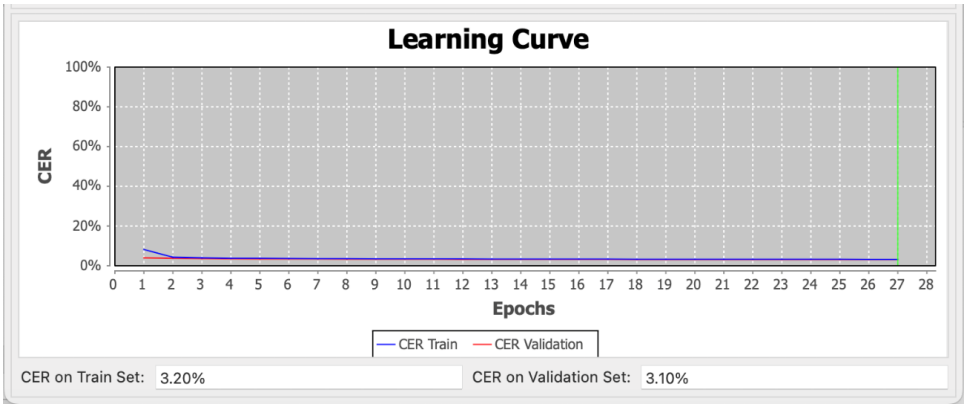


Figura 4. Entrenamiento del modelo *Spanish Golden Age Prints (Spelling modernization) 1.0*.

10. Cuéllar, 2021c.

4. Entrenamiento en el reconocimiento de manuscritos

El siguiente paso natural para conseguir reunir el mayor número posible de textos teatrales era el de transcribir también los manuscritos del Siglo de Oro. Existe una gran cantidad de obras cuyos testimonios solo encontramos en manuscritos de la época o copias de posteriores siglos. Este fin es mucho más complicado que el anterior. Si bien los impresos guardan una cierta homogeneidad y los tipos, aunque con variaciones, son muy parecidos y estables, los manuscritos responden a copistas y a letras distintas. Cada manuscrito presenta un desafío para su interpretación incluso para los entrenados ojos del filólogo paleógrafo. Es, por tanto, difícil, que la máquina llegue a transcribir este tipo de documentos con una precisión adecuada.

Con la experiencia adquirida en el reconocimiento de impresos entrené un modelo a través de la función Text2Image, combinando documentos teatrales con los textos del proyecto ETSO. En este caso no fue posible crear un modelo de transcripción que conservara la ortografía propia del Siglo de Oro, puesto que no he podido encontrar suficientes textos con estas características para el entrenamiento[11]. Por tanto, el aprendizaje se hizo directamente con textos de ortografía modernizada. Entrené el modelo con 3.250.116 palabras (1.046.287 líneas) y su curva de aprendizaje puede observarse en la Figura 5. Se alcanza en este caso un CER (Character Error Rate) en el set de validación del 9,20%, por lo que, de cada 100 caracteres, 90,8 son transcritos y modernizados correctamente. El modelo, denominado *Spanish Golden Age Manuscripts (Spelling modernization) 1.0*[12], está también disponible para la comunidad investigadora a través de Transkribus.

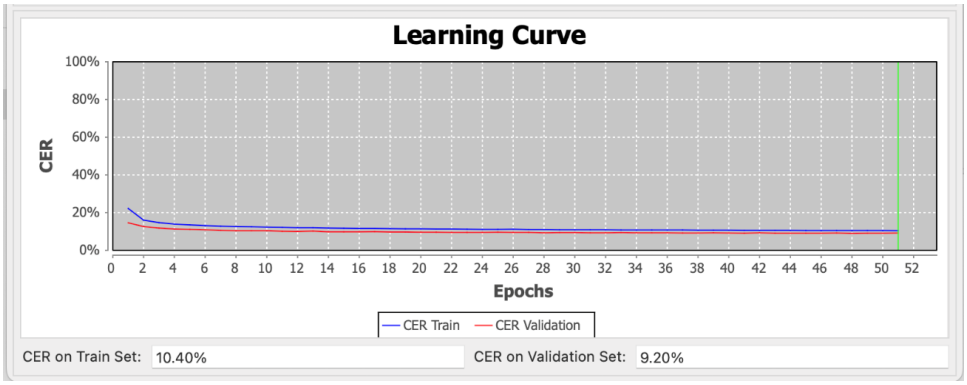
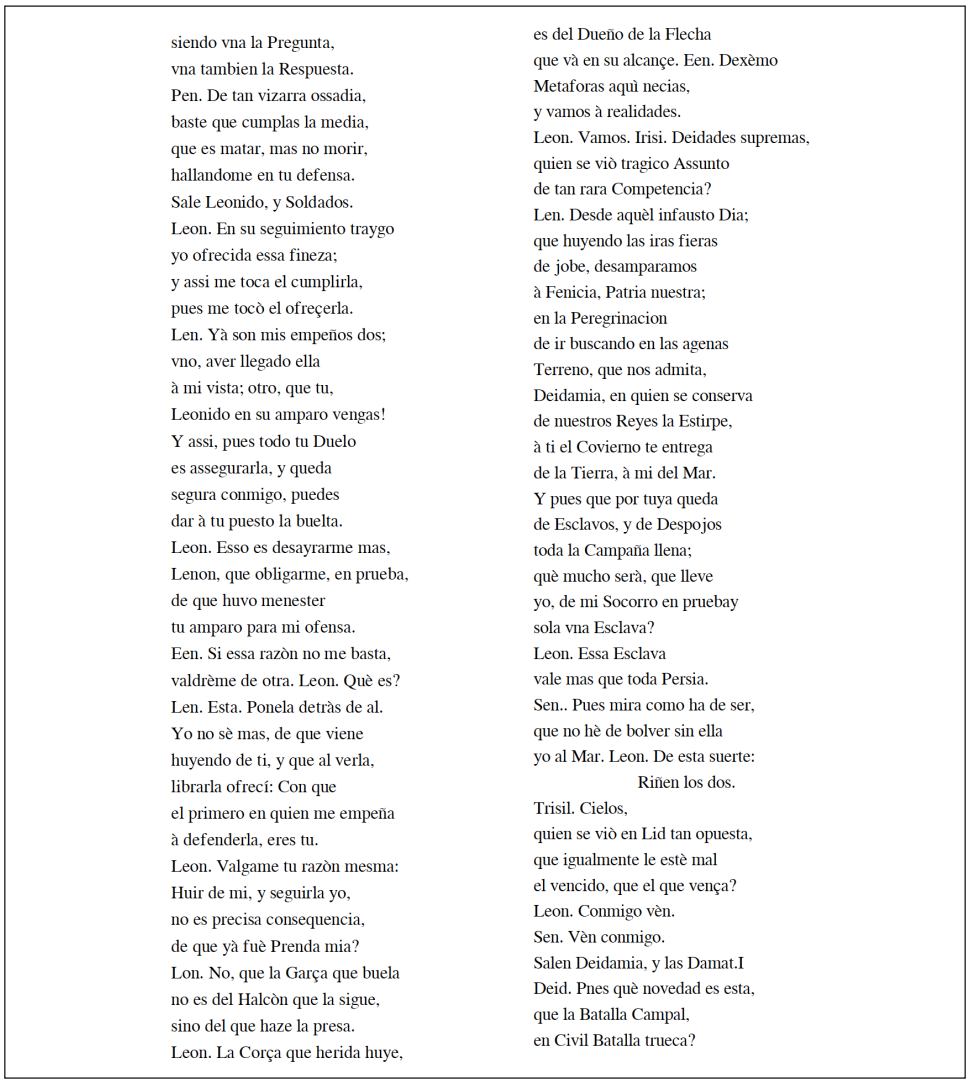


Figura 5. Entrenamiento del modelo *Spanish Golden Age Manuscripts (Spelling modernization) 1.0*.

11. Un buen corpus para este objetivo parecería el de TESO, sin embargo, la ortografía de sus ediciones no coincide exactamente con la de los documentos manuscritos de la época, por lo que tampoco sirve para nuestros propósitos.
12. Cuéllar, 2021a.

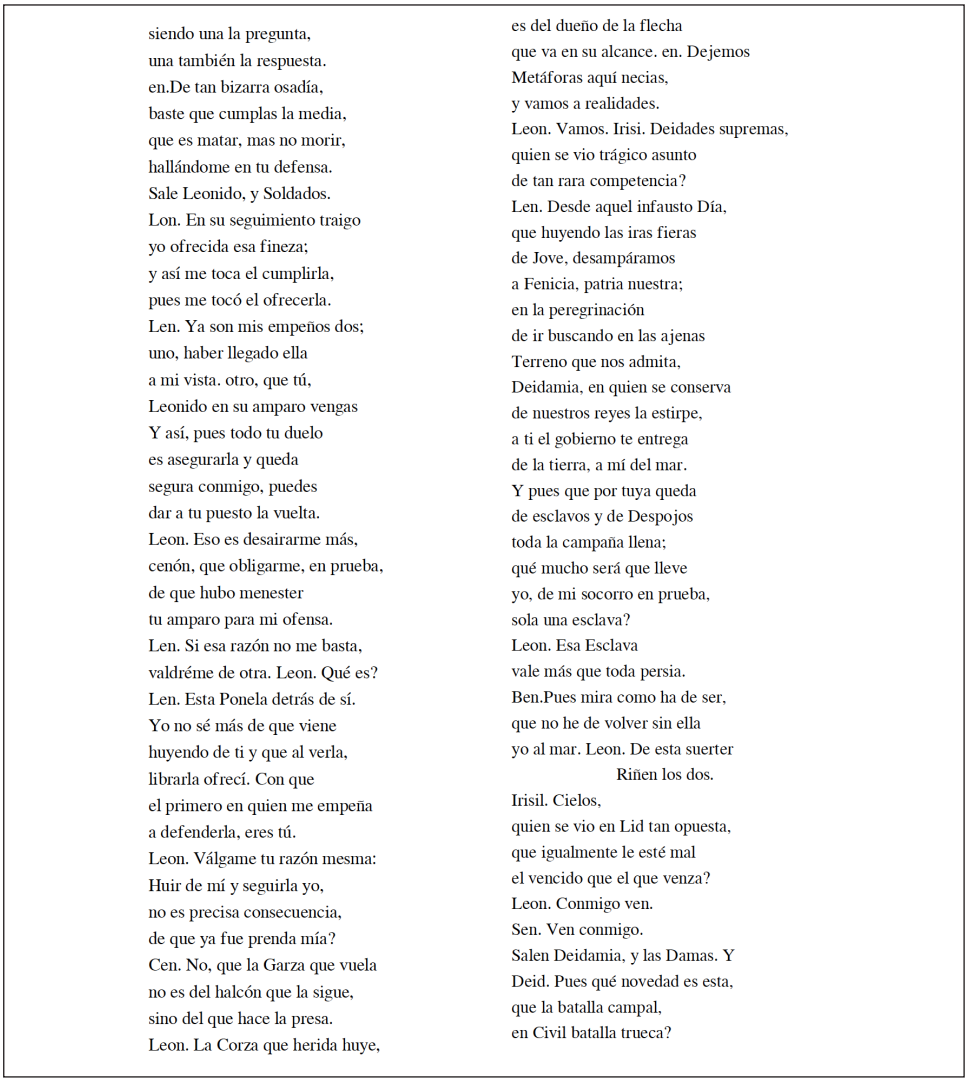
5. Transcripción automática de impresos

Una vez entrenados los modelos para la transcripción de impresos tanto con ortografía de la época como con ortografía modernizada, podemos comprobar su eficacia con algunos ejemplos. En la Figura 6 aplico el modelo *Spanish Golden Age Prints 1.0* a nuestro documento de partida. Aquí se puede apreciar que el resultado es bastante aceptable, como corresponde al éxito porcentual alcanzado en el entrenamiento. La gran mayoría de palabras son transcritas de forma correcta. En la Figura 7 aplico el modelo *Spanish Golden Age Prints (Spelling modernization) 1.0*. Se puede comprobar como el resultado también es bastante ajustado a nuestros propósitos.



siendo vna la Pregunta,
vna tambien la Respuesta.
Pen. De tan vizarra ossadia,
baste que cumplas la media,
que es matar, mas no morir,
hallandome en tu defensa.
Sale Leonido, y Soldados.
Leon. En su seguimiento traygo
yo ofrecida essa fineza;
y assi me toca el cumplirla,
pues me tocò el ofreçerla.
Len. Yà son mis empeños dos;
vno, aver llegado ella
à mi vista; otro, que tu,
Leonido en su amparo vengas!
Y assi, pues todo tu Duelo
es asegurarla, y queda
segura conmigo, puedes
dar à tu puesto la buelta.
Leon. Esso es desayrarme mas,
Lenon, que obligarme, en prueba,
de que huvo menester
tu amparo para mi ofensa.
Een. Si essa razòn no me basta,
valdrème de otra. Leon. Què es?
Len. Esta. Ponela detràs de al.
Yo no sè mas, de que viene
huyendo de ti, y que al verla,
librarla ofrecí: Con que
el primero en quien me empeña
à defenderla, eres tu.
Leon. Valgame tu razòn mesma:
Huir de mi, y seguirla yo,
no es precisa consequencia,
de que yà fuè Prenda mia?
Lon. No, que la Garça que buela
no es del Halcòn que la sigue,
sino del que haze la presa.
Leon. La Corça que herida huye,
es del Dueño de la Flecha
que và en su alcançe. Een. Dexèmo
Metaforas aquì necias,
y vamos à realidades.
Leon. Vamos. Irisi. Deidades supremas,
quien se viò tragico Assunto
de tan rara Competencia?
Len. Desde aquèl infausto Dia;
que huyendo las iras fieras
de jobe, desamparamos
à Fenicia, Patria nuestra;
en la Peregrinacion
de ir buscando en las agenas
Terreno, que nos admita,
Deidamia, en quien se conserva
de nuestros Reyes la Estirpe,
à ti el Covierno te entrega
de la Tierra, à mi del Mar.
Y pues que por tuya queda
de Esclavos, y de Despojos
toda la Campaña llena;
què mucho serà, que lleve
yo, de mi Socorro en pruebay
sola vna Esclava?
Leon. Essa Esclava
vale mas que toda Persia.
Sen.. Pues mira como ha de ser,
que no hè de bolver sin ella
yo al Mar. Leon. De esta suerte:
Riñen los dos.
Trisil. Cielos,
quien se viò en Lid tan opuesta,
que igualmente le estè mal
el vencido, que el que vença?
Leon. Conmigo vèn.
Sen. Vèn conmigo.
Salen Deidamia, y las Damat.I
Deid. Pnes què novedad es esta,
que la Batalla Campal,
en Civil Batalla trueca?

Figura 6. Transcripción automática respetando la ortografía de impreso



siendo una la pregunta,
una también la respuesta.
en.De tan bizarra osadía,
baste que cumplas la media,
que es matar, mas no morir,
hallándome en tu defensa.
Sale Leonido, y Soldados.
Lon. En su seguimiento traigo
yo ofrecida esa fineza;
y así me toca el cumplirla,
pues me tocó el ofrecerla.
Len. Ya son mis empeños dos;
uno, haber llegado ella
a mi vista. otro, que tú,
Leonido en su amparo vengas
Y así, pues todo tu duelo
es asegurarla y queda
segura conmigo, puedes
dar a tu puesto la vuelta.
Leon. Eso es desairarme más,
cenón, que obligarme, en prueba,
de que hubo menester
tu amparo para mi ofensa.
Len. Si esa razón no me basta,
valdréme de otra. Leon. Qué es?
Len. Esta Ponela detrás de sí.
Yo no sé más de que viene
huyendo de ti y que al verla,
librarla ofrecí. Con que
el primero en quien me empeña
a defenderla, eres tú.
Leon. Válgame tu razón mesma:
Huir de mí y seguirla yo,
no es precisa consecuencia,
de que ya fue prenda mía?
Cen. No, que la Garza que vuela
no es del halcón que la sigue,
sino del que hace la presa.
Leon. La Corza que herida huye,
es del dueño de la flecha
que va en su alcance. en. Dejemos
Metáforas aquí necias,
y vamos a realidades.
Leon. Vamos. Irisi. Deidades supremas,
quien se vio trágico asunto
de tan rara competencia?
Len. Desde aquel infausto Día,
que huyendo las iras fieras
de Jove, desampáramos
a Fenicia, patria nuestra;
en la peregrinación
de ir buscando en las ajenas
Terreno que nos admita,
Deidamia, en quien se conserva
de nuestros reyes la estirpe,
a ti el gobierno te entrega
de la tierra, a mí del mar.
Y pues que por tuya queda
de esclavos y de Despojos
toda la campaña llena;
qué mucho será que lleve
yo, de mi socorro en prueba,
sola una esclava?
Leon. Esa Esclava
vale más que toda persia.
Ben.Pues mira como ha de ser,
que no he de volver sin ella
yo al mar. Leon. De esta suerter
Riñen los dos.
Irisil. Cielos,
quien se vio en Lid tan opuesta,
que igualmente le esté mal
el vencido que el que venza?
Leon. Conmigo ven.
Sen. Ven conmigo.
Salen Deidamia, y las Damas. Y
Deid. Pues qué novedad es esta,
que la batalla campal,
en Civil batalla trueca?

Figura 7. Transcripción automática con modernización ortográfica de impreso

Para el cometido del proyecto solo nos interesa el segundo de los modelos, el que es capaz de modernizar la ortografía. Gracias a este, *Spanish Golden Age Prints (Spelling modernization) 1.0*, pude transcribir unas 950 obras (las cuales pueden consultarse en https://etso.es/cetso bajo el estado: Trans. aut. IMPR.) con resultados muy positivos. Este proceso hubiera llevado años a equipos enteros de filólogos dedicados a ello, pero, gracias a la Inteligencia Artificial, en apenas unas horas se puede completar el proceso. Afortunadamente, las pruebas estilométricas, las cuales tienen en cuenta el uso de las palabras más comunes, parecen no verse especialmente afectadas por un porcentaje irrelevante de errores en la trans-

cripción, que suelen producirse especialmente en palabras poco comunes[13]. Los resultados estilométricos, por tanto, son muy similares con obras procedentes de ediciones cuidadas y con obras procedentes de la transcripción automática. Esto quiere decir que los textos son perfectamente útiles para su análisis estilométrico. En el proyecto ETSO hemos procedido al estudio autorial de todos los impresos y hemos llegado a conclusiones novedosas de autoría que afectan a decenas de obras y que serán presentadas tanto en el portal web como en futuros trabajos.

6. Transcripción automática de manuscritos

El modelo creado para la transcripción y modernización automática de manuscritos, *Spanish Golden Age Manuscripts (Spelling modernization) 1.0*, ofrece resultados como el de la Figura 8 a partir del documento de ejemplo. Se trata de un resultado lejos de ser perfecto, pero que puede ayudar en la desafiante tarea de la transcripción de este tipo de textos. No en vano, se trata de una escritura difícil de entender y transcribir hasta para el ojo especialista. Encontraremos distintos fallos en los textos resultantes debido a un abanico de razones: una caligrafía especialmente engorrosa, la presencia de tachaduras, la suciedad del documento, la translucidez de la cara opuesta, la suerte de haber o no realizado el entrenamiento con documentos de caligrafía similar, y un largo etcétera. Esto no debe desalentarnos, porque también sorprende lo acertado de la transcripción en muchos casos, en los que incluso tildes que no existían previamente se colocan de forma adecuada, o pasajes muy intrincados para el ojo humano o incluso tachados son transcritos sin mayor problema por la máquina.

Empleé este modelo con unos 350 documentos (los cuales pueden consultarse en https://etso.es/cetso bajo el estado: Trans. aut. MSS.) con resultados prometedores. Es cierto, sin embargo, que existe una gran cantidad de errores en las transcripciones automáticas de este tipo de textos. Los manuscritos son mucho más desafiantes que los impresos y presentan más complejidades a la hora de ser abordados. De forma similar al caso de los impresos, estas transcripciones automáticas pueden ser empleadas para los análisis estilométricos, aunque, ahora sí, debemos ser muy precavidos con los resultados, debido al porcentaje de error que obtenemos. En nuestro ejemplo, la transcripción automática del documento manuscrito de *La francesa Laura* mostraba, ya desde el primer momento, una relación estilométrica muy fuerte con el repertorio de Lope de Vega. Utilizamos la transcripción generada por el modelo como texto base para corregirlo manualmente y repetir las pruebas. Estas confirmaron el resultado de partida y relacionaron la obra con el conjunto lopesco en variedad de análisis. A esto, se sumó, por supuesto, una intensa investigación filológica que concluyó en la atribución de la obra anónima al fénix de los ingenios[14]. Sin la transcripción automática hubiera sido prácticamente

13. Eder, 2013 y Camps *et al.*, 2020.
14. Cuéllar y Vega García-Luengos, 2023.

imposible dar con esta obra en la enorme madeja de documentos de teatro del Siglo de Oro que existen. Estos procesos, por tanto, nos ofrecen unas pistas valiosísimas que luego pasamos a investigar por los cauces tradicionales.

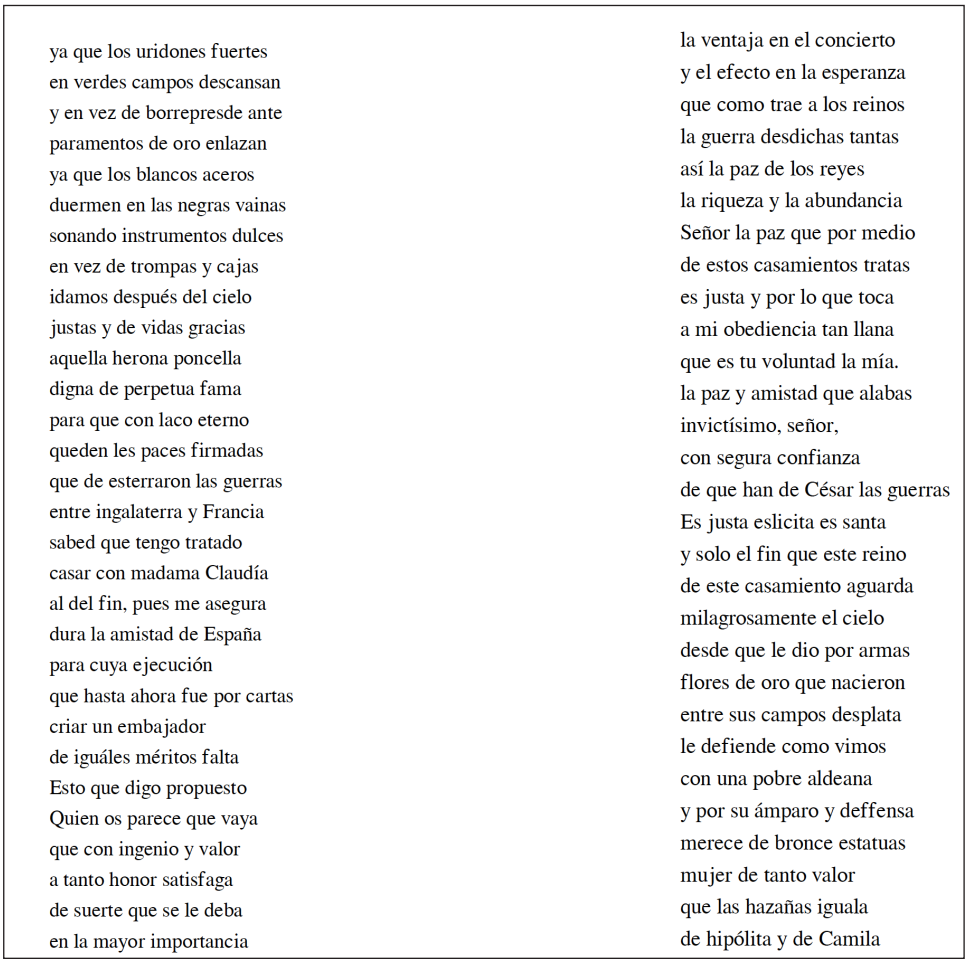

ya que los uridones fuertes
en verdes campos descansan
y en vez de borrepresde ante
paramentos de oro enlazan
ya que los blancos aceros
duermen en las negras vainas
sonando instrumentos dulces
en vez de trompas y cajas
idamos después del cielo
justas y de vidas gracias
aquella herona poncella
digna de perpetua fama
para que con laco eterno
queden les paces firmadas
que de esterraron las guerras
entre ingalaterra y Francia
sabed que tengo tratado
casar con madama Claudía
al del fin, pues me asegura
dura la amistad de España
para cuya ejecución
que hasta ahora fue por cartas
criar un embajador
de iguáles méritos falta
Esto que digo propuesto
Quien os parece que vaya
que con ingenio y valor
a tanto honor satisfaga
de suerte que se le deba
en la mayor importancia
la ventaja en el concierto
y el efecto en la esperanza
que como trae a los reinos
la guerra desdichas tantas
así la paz de los reyes
la riqueza y la abundancia
Señor la paz que por medio
de estos casamientos tratas
es justa y por lo que toca
a mi obediencia tan llana
que es tu voluntad la mía.
la paz y amistad que alabas
invictísimo, señor,
con segura confianza
de que han de César las guerras
Es justa eslicita es santa
y solo el fin que este reino
de este casamiento aguarda
milagrosamente el cielo
desde que le dio por armas
flores de oro que nacieron
entre sus campos desplata
le defiende como vimos
con una pobre aldeana
y por su ámparo y deffensa
merece de bronce estatuas
mujer de tanto valor
que las hazañas iguala
de hipólita y de Camila

Figura 8. Transcripción automática con modernización ortográfica de manuscrito

7. Conclusiones

A lo largo de cuatro siglos, cientos de impresos y manuscritos teatrales han permanecido olvidados y desatendidos, a la espera de poder ser trabajados desde la filología. He desarrollado tres modelos para la transcripción automática a través de la herramienta Transkribus y su sistema de Inteligencia Artificial. Estos modelos

están disponibles ya para el público investigador y pueden ser empleados en documentos con características similares a las del teatro aurisecular. Gracias a estos procesos he podido transcribir unos 1300 textos que ahora forman parte del corpus de ETSO (https://etso.es/) y del buscador TEXORO[15] (https://etso.es/texoro).

Las transcripciones automáticas cuentan, en muchos casos, con la calidad suficiente para ser tratadas con técnicas estilométricas y están arrojando resultados autoriales de interés, como la atribución a Lope de Vega de *La francesa Laura* y otras que verán la luz próximamente. Podemos utilizar textos que, si hubieran de ser transcritos por la vía tradicional, requerirían una cantidad ingente de esfuerzo. Gracias, por tanto, a la Inteligencia Artificial, podemos ahorrar mucho de este esfuerzo a los investigadores, sobre todo en trabajo mecánico con corpus de documentos de fácil lectura, pero de mucha extensión. En el futuro estos modelos deben perfeccionarse para disminuir la tasa de error y poder ofrecer textos finales menos deturpados. Esa es la razón por la que, en su título, los tres modelos desarrollados, denominados *Spanish Golden Age Prints 1.0*., *Spanish Golden Age Prints (Spelling modernization) 1.0* y *Spanish Golden Age Manuscripts (Spelling modernization) 1.0*, llevan el apéndice «1.0», porque es mi intención el seguir mejorándolos en el futuro. Por el momento, el ser humano entrenado sigue pudiendo transcribir con mayor precisión este tipo de documentos, pero, eso sí, con un ritmo mucho más lento. El trabajo aquí acometido habría supuesto varios años a un nutrido grupo de filólogos, sin embargo, gracias a la Inteligencia Artificial y al entrenamiento oportuno, cada obra queda transcrita en función de minutos.

Las Humanidades Digitales suponen una ayuda inconmensurable para el investigador del teatro del Siglo de Oro. Estos procesos y herramientas, lejos de reemplazarnos, facilitan y empujan nuestro trabajo hasta niveles nunca antes imaginados. Aliémonos con los nuevos recursos, en vez de enfrentarlos, y prosigamos nuestra labor.

Bibliografía

Aranda García, Nuria, «Humanidades Digitales y literatura medieval española: la integración de Transkribus en la base de datos COMEDIC», *Historias Fingidas*, Número Especial 1 Humanidades Digitales y estudios literarios hispánicos, 2022, pp. 127-149.

Ayuso García, Manuel, «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 Humanidades Digitales y estudios literarios hispánicos, 2022, pp. 151-173.

15. Cuéllar y Vega García-Luengos, 2022.

Bazzaco, Stefano, «El reconocimiento automático de textos en letra gótica del Siglo de Oro: creación de un modelo HTR basado en libros de caballerías del siglo XVI en la plataforma Transkribus», *Janus. Estudios sobre el Siglo de Oro*, 9, 2020, pp. 534-561.

Bazzaco, Stefano, *et al.*, «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 Humanidades Digitales y estudios literarios hispánicos, 2022, pp. 67-125.

Blasut, Giada, «Los modelos de HTR Silves1549_BNE y Spanish Gothic como herramientas de la labor ecdótica», *Historias Fingidas*, Número Especial 1 Humanidades Digitales y estudios literarios hispánicos, 2022, pp. 175-193.

Camps, Jean-Baptiste, Clérice, Thibault, y Pinche, Ariana, «Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis», *Digital Scholarship in the Humanities*, 36, 2, 2021, pp. ii49–ii71, https://doi.org/10.1093/llc/fqab033.

Cuéllar, Álvaro, «Spanish Golden Age Manuscripts (Spelling Modernization) 1.0», *Transkribus*, 2021a.

Cuéllar, Álvaro, «Spanish Golden Age Prints 1.0», *Transkribus*, 2021b.

Cuéllar, Álvaro, «Spanish Golden Age Prints (Spelling Modernization) 1.0», *Transkribus*, 2021c.

Cuéllar, Álvaro, y Vega García-Luengos, Germán, *ETSO: Estilometría aplicada al Teatro del Siglo de Oro*, 2017-2023, http://etso.es/.

Cuéllar, Álvaro, y Vega García-Luengos, Germán, *TEXORO: Textos del Siglo de Oro*. 2022, http://etso.es/texoro.

Cuéllar, Álvaro, y Vega García-Luengos, Germán, «*La francesa Laura.* El hallazgo de una nueva comedia del Lope de Vega último», *Anuario Lope de Vega. Texto, literatura, cultura*, XXIX, 2023, pp. 131-198.

Eder, Maciej, «Mind your Corpus: Systematic Errors in Authorship Attribution», *Literary and Linguistic Computing*, 28.4, 2013, pp. 603-614.

Fradejas Rueda, José Manuel, «De editor analógico a editor digital», *Historias Fingidas*, Número Especial 1 Humanidades Digitales y estudios literarios hispánicos, 2022, pp. 39-65.

Muehlberger, Guenter, *et al.*, «Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study», *Journal of Documentation*, 75.5, 2019, pp. 954-976.

Simón Palmer, Carmen (coord.), *Teatro Español del Siglo de Oro. TESO*, Chadwyck-Healey España, Madrid, 1998.