Los problemas de identificación de caracteres OCR para la recuperación de texto en el libro antiguo: un análisis de caso en el Fondo Antiguo de la Biblioteca Central, UNAM

Identification problems for OCR characters for text retrieval in ancient books: A case study in the Ancient Collections of the Central Library at UNAM

Silvia Socorro Ballesteros Estrada,* Guillermo Morales Romero, Pavel Alfredo Cedillo Pérez*****

Resumen

El presente artículo describe de manera general los problemas enfrentados para lograr una correcta recuperación de texto por medio del reconocimiento óptico de caracteres (OCR) en el libro antiguo, tomando una muestra de las obras de los siglos XV al XVIII que resguarda el Fondo Antiguo de la Biblioteca Central de la Universidad Nacional Autónoma de México (UNAM), digitalizadas por la Dirección General de Bibliotecas (DGB). Se presenta, en primer lugar, la exposición teórica conceptual del OCR y su aplicación en la recuperación de texto para continuar con la ejemplificación de los factores que determinan la correcta o incorrecta identificación de los grafemas en estos libros mediante las pruebas aplicadas con el software Adobe Acrobat 8 Professional® y, por último, muestra algunos hallazgos obtenidos como producto del análisis e interpretación de los datos correspondientes a las variables.

Palabras Clave: Reconocimiento de texto, OCR, libro antiguo, digitalización de obras antiguas

Abstract

This article describes, in general terms, the problems faced for proper text retrieval through optical character recognition (OCR) in ancient books, by taking a sample of works from the fifteenth to the eighteenth centuries that are protected in the Ancient Collections of the Central Library at UNAM, and digitized by the General Directorate of Libraries. It first presents a conceptual theoretical exposition of OCR and its application in text retrieval to continue with the exemplification of the factors that determine the correct or incorrect identification of the graphemes in these books, by means of some tests applied with Adobe Acrobat 8 Professional and, last, it shows some findings obtained as a result of the analysis and interpretation of the data corresponding to the variables in question.

Keywords: Text recognition, OCR, ancient collections, digitization.

* Secretaria Técnica de Biblioteca Digital, Dirección General de Bibliotecas. Anexo de la DGB, Circuito de la Investigación Científica, UNAM-CU, C.P. 04510, México D.F., México. Correo electrónico: silviabe@dgb.unam.mx.

** Fondo Antiguo y Colecciones Especiales, Biblioteca Central. Décimo piso del Edificio de Biblioteca Central, Circuito Interior, UNAM-CU, C.P. 04510, México D.F., México. Correo electrónico: guillermoralesromero@gmail.com; guillermom@dgb.unam.mx.

*** Secretaría Técnica de Biblioteca Digital, Dirección General de Bibliotecas. Anexo de la DGB, Circuito de la Investigación Científica, UNAM-CU, C.P. 04510, México D.F., México. Correo electrónico: alfredoc@dgb.unam.mx.

Introducción

Los libros antiguos –entendidos como aquellos libros impresos cuyo común denominador es haber sido producidos manualmente antes de la aparición de la imprenta mecánica y la industria editorial– forman parte de nuestro patrimonio bibliográfico, su importancia y valor histórico han motivado a las instituciones que lo custodian a facilitar su preservación y acceso universal. Los esfuerzos que muchas bibliotecas alrededor del mundo están llevando a cabo para su rescate, estudio y divulgación como medidas de salvaguarda de los mismos son claros ejemplos de ello.

Los proyectos de digitalización de libros antiguos han permitido un enorme avance en su rescate y preservación. Con una buena digitalización se obtiene un documento digital que emula a su original impreso, facilita la consulta del contenido intelectual de las obras por varios lectores a la vez y proporciona una reproducción casi facsimilar de los documentos; además, se eliminan las restricciones de consulta por motivos conservativos y se evita el traslado de los usuarios al recinto donde son custodiados.

Pero la digitalización de estos libros no se reduce a la conversión de formatos –de análogo a digital– sino va más allá y permite la obtención de documentos digitales bien organizados que logren la recuperación y diseminación de la información vertida en esos libros para la generación de nuevo conocimiento. Todo lo anterior sin comprometer el estado conservativo de los documentos que, además de contener información textual, contienen gran información si se estudian como piezas arqueológicas a las que el paso del tiempo ha adherido o quitado elementos.

La Universidad Nacional Autónoma de México (UNAM), como muchas otras universidades alrededor del mundo, ha invertido recursos en digitalizar los documentos históricos que custodia para garantizar su conservación y consulta por parte de su comunidad. Durante el segundo trimestre de 2010, la Dirección General de Bibliotecas (DGB) de la UNAM inició un proyecto coordinado por la Subdirección de Biblioteca Central para digitalizar la totalidad de los libros antiguos que custodia su Fondo Antiguo. Se contemplaron en un principio todos los impresos entre los siglos XV y XVIII como prioridad, posteriormente se incluyeron aquellos libros decimonónicos cuya importancia histórica los convirtiera en candidatos ideales para su digitalización. Dicho proyecto procuró en todo momento la obtención de imágenes TIFF a 24 bits de color con una resolución superior a los 300 dpi y compresión LZW que asegurara una óptima calidad de imagen[1], capturando la mayor cantidad de información útil para la investigación de estas obras, es decir, información no sólo representada por el texto impreso sino también por el resto de los elementos que lo acompañan, como exlibris, marcas de fuego, anotaciones manuscritas en las portadas y al margen de las cajas tipográficas.

Tras un gran esfuerzo para ejecutar con éxito dicho proyecto debido a la delicadeza del trabajo, en este momento la DGB cuenta con más de 3 800 libros antiguos digitalizados y a disposición de la comunidad universitaria que desee consultarlos, descargarlos e imprimirlos.

Así pues, los avances tecnológicos en el campo de la digitalización nos han provisto de herramientas suficientemente poderosas para poder esquivar las limitaciones de recuperación de información a las que los modelos tradicionales nos ataban. Un ejemplo claro es el desarrollo de software para el reconocimiento óptico de caracteres que permite a las personas realizar búsquedas de texto completo en los documentos digitales que no nacieron originalmente como tales. ¿Cuáles son los problemas a los que nos enfrentamos si deseamos obtener un reconocimiento óptico de caracteres en documentos con más de trescientos años de antigüedad?

[1] Para la elaboración del proyecto se consideraron los estándares que la Universidad de Cornell adoptó y perfeccionó mediante una fórmula de Índice de Calidad (QI) para textos impresos, desarrollada por su Comité de normas C10 de AIIM y descrita en *Tutorial* [en línea]: *Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality. The Comission on Preservation and Access.* <http://www.clir.org/pubs/reports/pub53/pub53.pdf>

El OCR y su aplicación para libros antiguos

El software de reconocimiento óptico de caracteres u OCR (por sus siglas en inglés) es un proceso de escaneado y comparación de caracteres cuyo propósito es identificar letras o números impresos, evitando la necesidad de reescribirlos. El OCR intenta identificar las palabras por medido de la proximidad de caracteres y reconstruirlas según la disposición de la página original.[2] El software se basa en algoritmos que convierten las letras representadas en una imagen digital en caracteres ASCII, lenguaje que puede posteriormente ser leído y editado por algún procesador de textos común y corriente. De esta manera se puede ahorrar una gran cantidad de recursos humanos y se posibilita la búsqueda y recuperación de texto a través de una computadora.

En el procesamiento de OCR están presentes cuatro pasos básicos para lograr el reconocimiento de caracteres: *binarización, fragmentación, adelgazamiento* y *comparación de patrones.*[3]

La *binarización,* en términos llanos, consiste en convertir la imagen digital en bitonal, procurando que se conserven las *propiedades* esenciales de ésta. Una forma eficaz para realizar una óptima binarización es mediante su histograma[4], a partir del cual podemos identificar el número de pixeles en la escala de grises para dividirlos y convertirlos en negros o en blancos (imagen 1).

Posteriormente, la *fragmentación* localiza las zonas de interés (en este caso las letras) y las separara, basándose en la intensidad con la que están *dibujadas* o los *espacios blancos* entre ellas. Uno de los métodos para este proceso es segmentar la imagen digital en pequeños clusters o áreas que no contengan elementos unidos en algún punto.

El *adelgazamiento* de los componentes consiste en borrar de manera sucesiva los puntos del borde de cada letra, preservando su tipología. El borrado de puntos se realiza a partir de un esquema de barridos sucesivos para no deformar la imagen original y conservar su *figura* (imagen 2).

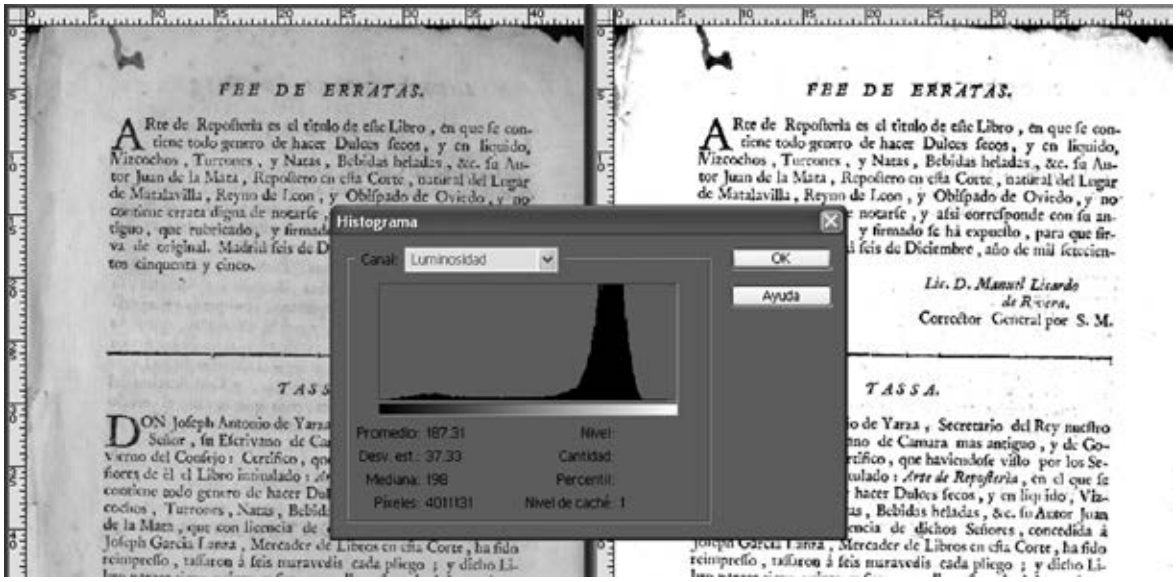


Imagen 1. Histograma de una imagen binarizada.

2 *OCR* [en línea]. Encyclopædia Britannica: Science & Technology. Britannica Academic Edition. <http://www.britannica.com/EBchecked/topic/430371/OCR>

3 Además, cabe señalar que se han desarrollado algoritmos que permiten identificar otros factores en el texto como: inclinación, ángulo y distancia, como *Optical Character Recongnition for Cursive Handwriting.*

4 Gráfico de barras que permite reconocer y analizar patrones de comportamiento en la información que no son aparentes a primera vista al calcular un porcentaje o la media.

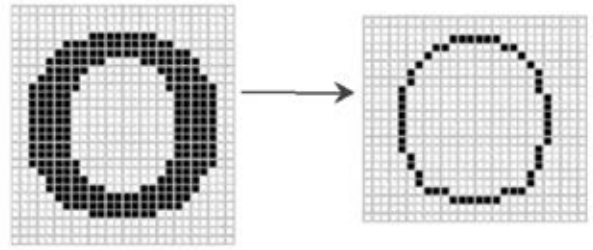


Imagen 2. Adelgazamiento de un componente.

Finalmente, el proceso de *comparación* identifica los caracteres resultantes del paso anterior y los coteja con una serie de plantillas almacenadas en una base de datos, permitiendo su identificación como letras o números. Esta etapa es definitiva para la recuperación del texto, ya que de su buen funcionamiento dependerá la obtención de la mayor cantidad de caracteres reconocidos correctamente.

Es importante anotar que el funcionamiento exitoso del proceso OCR radica esencialmente en la *clase* de imagen a la que se le apliquen estos procesos, es decir, una buena imagen supone dos cosas fundamentales: primera, que el texto en el documento original sea legible, exento de roturas o manchas, con letras uniformes y bien impresas (elementos comunes en una publicación moderna); segunda, que la representación digital que se obtenga de ella sea nítida, encuadrada, sin perspectiva o deformaciones por curvatura y a una resolución suficiente que permita la captura fiel del texto impreso. Con una buena ejecución en los procesos de digitalización pueden evitarse problemas futuros, logrando mejores resultados sólo con determinar los requisitos de conversión, tal como lo indica en su tercera sección el tutorial de digitalización de imágenes *Llevando la teoría a la práctica*[5] de la Universidad de Cornell. En resumen, mientras más clara y menos *ruido* contenga la imagen, mejor.

¿Y en qué se distingue un impreso moderno de un impreso antiguo? Como ya se dijo antes, su común denominador es haber sido producidos manualmente. Con esto no quiere decir que hayan sido *escritos a mano,* sino que su manufactura no se vio determinada por procesos mecánicos. Estos libros, como muchos saben, se elaboraban por medio de la composición de *tipos* móviles, los cuales eran letras de metal fundido producidas por una matriz que les daba forma; la sucesión de estos *tipos* formaban las palabras que se acomodaban dentro de lo que se denominó *caja tipográfica.* De esa manera el texto cobraba forma en la caja y se imprimía en papel mediante el golpe de una imprenta manual. Así pues una *buena* impresión estaba condicionada por muchos factores, por ejemplo: la viscosidad de tinta empleada, la fuerza con la que se *golpeaba* la prensa o el desgaste de los *tipos* de metal, por lo tanto los productos obtenidos no eran siempre homogéneos. Además, el libro antiguo, como documento histórico, es susceptible al deterioro que el tiempo mismo le confiere, como roturas, ataques de insectos, deformaciones por humedad, manchas, expurgos, desgaste del papel, etcétera.

Es claro entonces que los resultados de los procesos OCR se verán afectados por la naturaleza misma de los documentos y no sólo por la calidad de la imagen digital que se obtenga de éstos. Incluso con una resolución superior a los 600 dpi y una nitidez impecable, el OCR encontrará limitantes para las que no fue diseñado en el reconocimiento de caracteres del libro antiguo.

De la teoría a la práctica: pruebas de concepto OCR mediante Adobe Acrobat Professional®

Uno de los programas que tienen entre sus funciones el Reconocimiento Óptico de Caracteres es Adobe Acrobat Professional®, al ser uno de los software más completos disponibles en el mercado y con buenos resultados en el reconocimiento de patrones para textos modernos; se eligió para una prueba de OCR a algunos ejemplos de las imágenes obtenidas durante el proyecto de digitalización de la DGB-UNAM.

Para la aplicación de este proceso basta con abrir el documento, el cual puede tener una extensión .PDF o .JPEG, .TIFF, .GIF, .PNG, o cualquier otra de archivo de

[5] *Llevando la teoría a la práctica* [en línea]: *tutorial de digitalización de imágenes.* Biblioteca de la Universidad de Cornell, Departamento de Investigación. <http://www.library.cornell.edu/preservation/tutorial-spanish/conversion/conversion-04.html>

imagen, y posteriormente utilizar la opción *Document →OCR Text Recognition→Recognize Text Using OCR.*[6] El resultado del proceso es un documento digital en formato de texto, es decir, caracteres reconocibles por una computadora, que se *empalma* a los caracteres dibujados en la imagen y permite seleccionar, buscar, copiar y pegar el texto desde el archivo origen (imagen 3).

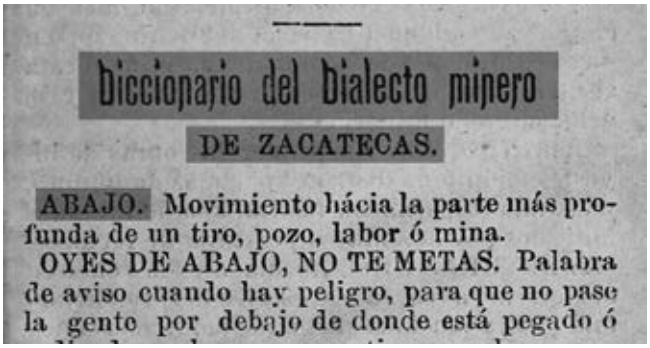

Biccionario del bialecto minero
DE ZACATECAS.
ABAJO. Movimiento hácia la parte más profunda de un tiro, pozo, labor ó mina.
OYES DE ABAJO, NO TE METAS. Palabra de aviso cuando hay peligro, para que no pase la gente por debajo de donde está pegado ó

Imagen 3. Texto seleccionable.

En este primer paso se deberá indicar en qué idioma está escrito el texto que el usuario desea reconocer, pues, como se mencionó, una vez que los procesos de *binarización, fragmentación, adelgazamiento* estén completos se intentará reconocer el texto a partir de la comparación de sus resultados con plantillas almacenadas en una base de datos. Es entonces donde surge el primer problema ¿Qué sucede si el texto que quiero reconocer está en latín o en griego clásico? Afortunadamente el software trabaja a nivel gráfico mas no semántico; dicho de otro modo, lo que busca es el reconocimiento de cada una de las letras que formen palabras, no de palabras completas con sentido lógico. En teoría, basta con que el texto que queramos procesar contenga los mismos caracteres alfabéticos que el del idioma seleccionado y el OCR se optimiza si seleccionamos correctamente el idioma a procesar. Para las pruebas se optó por textos en francés, español y alemán, ya que el software posee la opción de idioma para estos tres casos.

Debido a que todas las imágenes del proyecto de digitalización de la DGB fueron generadas con las mismas características técnicas, y considerando los elementos a partir de los cuales trabaja el OCR, los ejemplos seleccionados se basaron principalmente en la forma del texto impreso y no en el idioma en el que están escritos. Por lo tanto, se ejecutaron los procesos OCR en tres textos con tipografías diferentes, a saber: Romana, Cursiva y Gótica.

El idioma español –al igual que el italiano, el francés y el resto de lenguas romances– emplea letras del alfabeto latino. Del mismo modo que cualquier lengua, ha sufrido procesos evolutivos hasta el establecimiento de sus reglas ortográficas; asimismo, la representación gráfica de las mismas también puede variar de acuerdo a la época en la que fueron escritas o impresas. Por ejemplo, un texto en el español más correcto puede estar representado con grafemas que en la actualidad han caído en desuso, tal es el caso de la *"ç"* o la *"s"* larga[7] (imagen 4). Uno de los principales problemas hallados en el reconocimiento por OCR fue la incapacidad de identificar este grafema. En todas las pruebas de las diferentes tipografías el software la reconoce como *"f".* Del mismo modo fácilmente confunde *"t"* por *"r"* y *"l"* por *"j".*

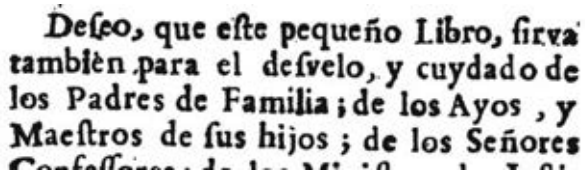

Deſeo, que eſte pequeño Libro, ſirva
también para el deſvelo, y cuydado de
los Padres de Familia; de los Ayos, y
Maeſtros de ſus hijos; de los Señores

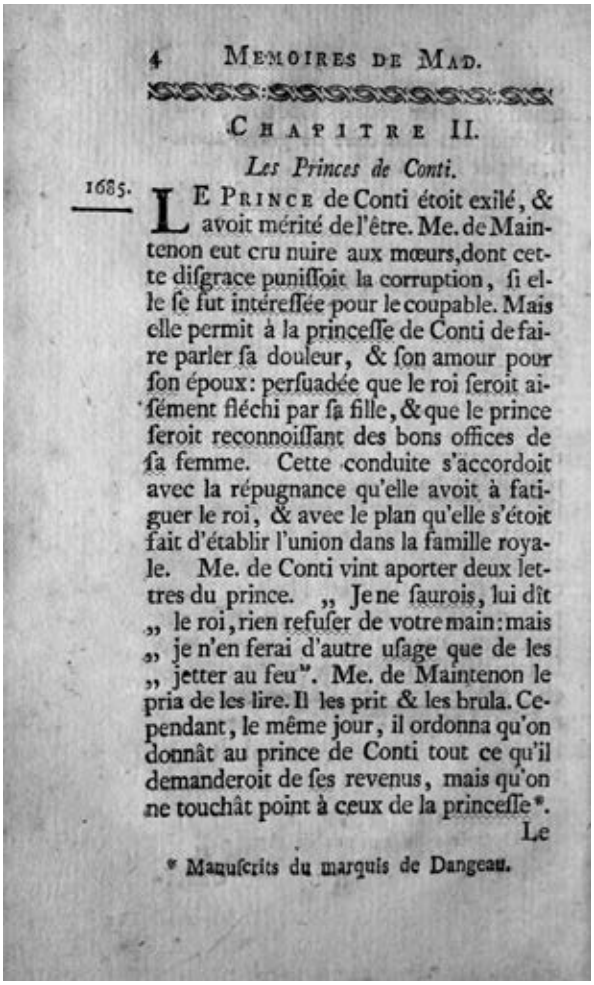
Imagen 4: Ejemplo del grafema "s" larga en un impreso español del siglo XVIII.

El siguiente ejemplo (imagen 5) es un texto francés del siglo XVIII impreso en letras romanas, al cual se le aplicó el OCR; dentro de la imagen aparecen marcadas todas las palabras que el software no reconoció correctamente. Como se puede apreciar, los caracteres son claros, legibles, bien impresos y con interlineados suficientemente espaciosos. El documento está en buenas condiciones y no presenta rasgaduras ni manchas que puedan entorpecer el proceso OCR; sin embargo, las pa-

[6] *Acrobat* OCR [en línea]: *make your scanned documents searchable. Acrobat blog insights, trends, news and highlights on all things Acrobat.* <http://blogs.adobe.com/acrobat/acrobat_ocr_make_your_scanned/>

[7] Grafema empleado en muchos impresos entre los siglos XV y XVIII no sólo en español sino en francés, italiano, latín y alemán.

labras no reconocidas representan el 19% del total del texto. De este porcentaje de errores el 10% son palabras con el grafema *"s"* larga (gráfica 1).



4 MEMOIRES DE MAD.

CHAPITRE II.

Les Princes de Conti.

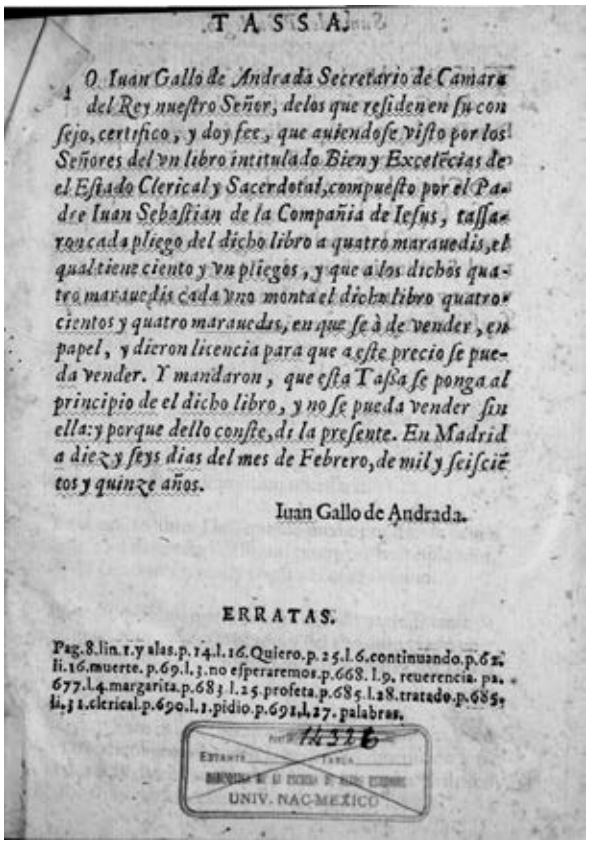
1685. LE PRINCE de Conti étoit exilé, & avoit mérité de l'être. Me. de Maintenon eut cru nuire aux mœurs, dont cette disgrace punissoit la corruption, si elle se fut intéressée pour le coupable. Mais elle permit à la princesse de Conti de faire parler sa douleur, & son amour pour son époux: persuadée que le roi seroit aisément fléchi par sa fille, & que le prince seroit reconnoissant des bons offices de sa femme. Cette conduite s'accordoit avec la répugnance qu'elle avoit à fatiguer le roi, & avec le plan qu'elle s'étoit fait d'établir l'union dans la famille royale. Me. de Conti vint aporter deux lettres du prince. „ Je ne saurois, lui dît „ le roi, rien refuser de votre main: mais „ je n'en ferai d'autre usage que de les „ jetter au feu". Me. de Maintenon le pria de les lire. Il les prit & les brula. Cependant, le même jour, il ordonna qu'on donnât au prince de Conti tout ce qu'il demanderoit de ses revenus, mais qu'on ne touchât point à ceux de la princesse*.

Le

* Manuscrits du marquis de Dangeau.

Imagen 5. Memoires pour servir a l'Hisroire de Madame de Maintenon. Amsterdam, 1756.

En un segundo ejercicio (imagen 6), se tomó un texto del siglo XVII en español, impreso en letras itálicas, cuya característica es la inclinación de las palabras y la concatenación de varios grafemas. Los resultados arrojaron sólo un 25% de palabras reconocidas correctamente, mientas que de las no reconocidas el 10% son originadas por la aparición de la *"s"* larga y el 65% es por otros casos. (Gráfica 2).



TASSA.

Yo Iuan Gallo de Andrada Secretario de Cámara del Rey nuestro Señor, delos que residen en su consejo, certifico, y doy fee, que auiendose visto por los Señores del vn libro intitulado Bien y Excelencias de el Estado Clerical y Sacerdotal, compuesto por el Padre Iuan Sebastian de la Compañia de Iesus, tassaron cada pliego del dicho libro a quatro marauedis, el qual tiene ciento y vn pliegos, y que a dos dichos quatro marauedis cada vno monta el dicho libro quatrocientos y quatro marauedis, en que se à de vender, en papel, y dieron licencia para que a este precio se pueda vender. Y mandaron, que esta Tassa se ponga al principio de el dicho libro, y no se pueda vender sin ella: y porque dello conste, de la presente. En Madrid a diez y seys dias del mes de Febrero, de mil y seiscientos y quinze años.

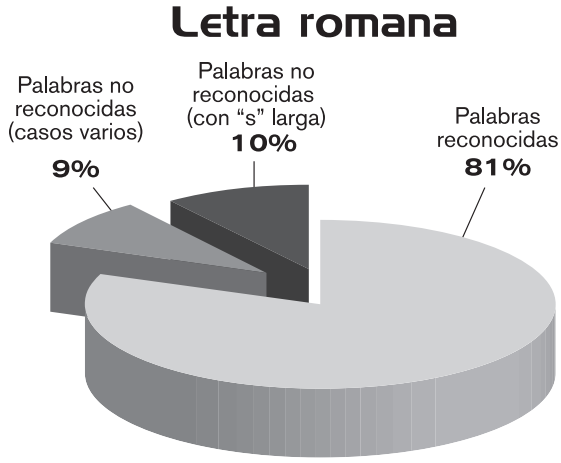
Iuan Gallo de Andrada.

ERRATAS.

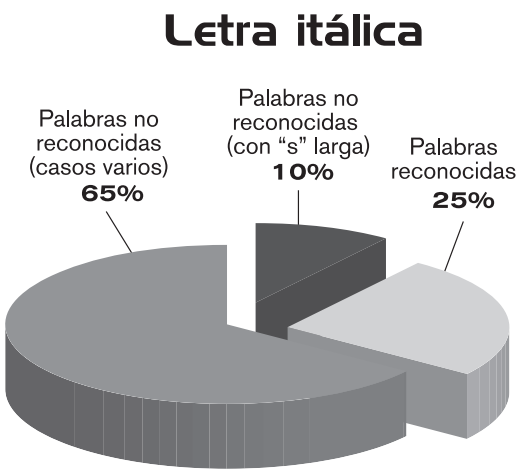
Pag. 8. lin. 1. y alas. p. 14. l. 16. Quiero. p. 25. l. 6. continuando. p. 62. li. 16. muerte. p. 69. l. 3. no esperaremos. p. 668. l. 9. reuerencia. pa. 677. l. 4. margarita. p. 683. l. 25. profeta. p. 685. l. 28. tratado. p. 685. li. 31. clerical. p. 690. l. 1. pidio. p. 691. l. 27. palabras.

UNIV. NAC. MEXICO

Imagen 6. De el bien: excelencias y obligaciones de el estado clerical y sacerdotal. Sevilla, 1615.

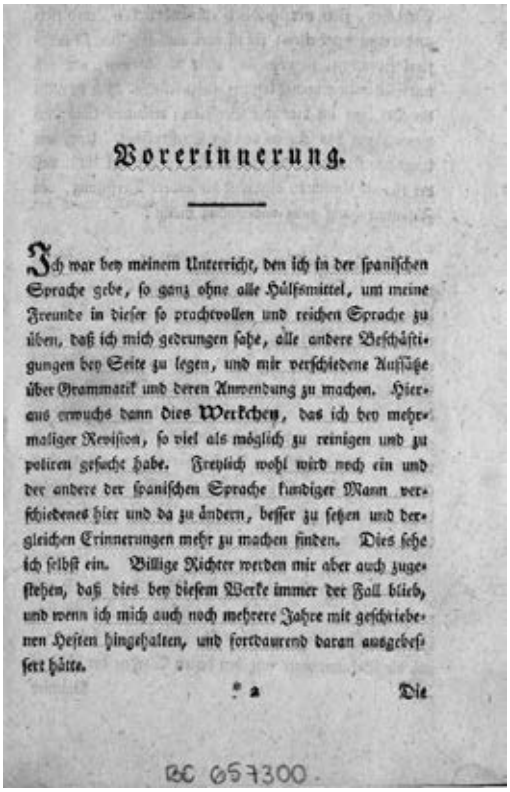


Gráfica 1. Porcentaje de texto reconocido con letra romana.



Gráfica 2. Porcentaje de texto reconocido con letra itálica.

En el último ejercicio, se eligió un texto en alemán del siglo XVIII con letra gótica. Se trata de grafemas angulosos y fracturados, es decir, una sola letra puede aparecer con trazos separados (imagen 7). Las letras góticas son de difícil lectura, incluso para una persona, si no se está habituado a ellas. Los resultados del OCR arrojaron sólo un 12% de palabras reconocidas correctamente, los caracteres mejor identificados son *"e", "f", "g", "i", "m",* mientras que entre los errores más frecuentes se hallan *"s", "d", "z"* y prácticamente todas las mayúsculas, además se confunde la *"d"* por *"b"* y la *"z"* por *"3"*. (Gráfica 3).

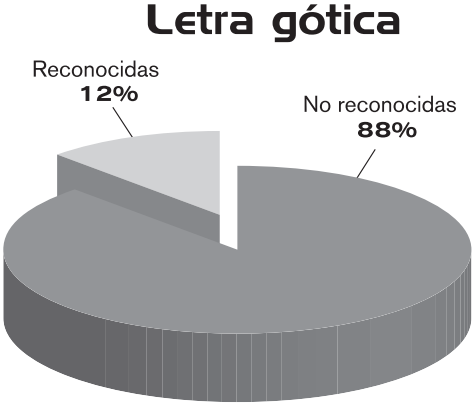

Vorerinnerung.

Ich war bey meinem Unterricht, den ich in der spanischen Sprache gebe, so ganz ohne alle Hülfsmittel, um meine Freunde in dieser so prachtvollen und reichen Sprache zu üben, daß ich mich gedrungen sahe, alle andere Beschäftigungen bey Seite zu legen, und mir verschiedene Aufsätze über Grammatik und deren Anwendung zu machen. Hieraus erwuchs dann dies Werkchen, das ich bey mehrmaliger Revision, so viel als möglich zu reinigen und zu poliren gesucht habe. Freylich wohl wird noch ein und der andere der spanischen Sprache kundiger Mann verschiedenes hier und da zu ändern, besser zu setzen und dergleichen Erinnerungen mehr zu machen finden. Dies sehe ich selbst ein. Billige Richter werden mir aber auch zugestehen, daß dies bey diesem Werke immer der Fall blieb, und wenn ich mich auch noch mehrere Jahre mit geschriebenen Heften hingehalten, und fortdaurend daran ausgebessert hätte.

* 2 Die

BC 057300

Imagen 7. Spanische Sprachlehre (La enseñanza del español). Leipzing, 1795.



Gráfica 3. Porcentaje de texto no reconocido en letra gótica.

Estos son, en cifras, los resultados de las pruebas OCR aplicadas a nuestras imágenes. Aunque debemos reconocer que –concordando con Hildelies[8]– además de las letras y los elementos ajenos al texto como manchas o expurgos, existen factores que limitan el buen funcionamiento del OCR, como la baja calidad de impresión, el espaciado irregular entre columnas y letras, las tipografías muy pequeñas, el gramaje del papel, etcétera. En los tres casos presentados se determinó como una variante la clase de tipografía impresa en el documento. Sin embargo, los porcentajes de palabras reconocidas pudieron estar o no determinados también por la calidad de la impresión, pues durante el proceso de *adelgazamiento,* como se mencionó, es necesario que los caracteres estén bien dibujados para que el software logre mantener la *figura* de la letra e identificarla como tal. Esto quiere decir que si nuestro texto presenta de origen letras muy delgadas o desgastadas (como es el caso del ejemplo con letra itálica) el proceso de *adelgazamiento* no proyectará valores reconocibles que empaten con las plantillas almacenadas en la base de datos.

Este problema no es exclusivo de los textos en letra itálica o gótica o romana, sino de muchos libros antiguos. Los *tipos* empleados por algunos impresores eran usados hasta que se desdibujaban o se rompían debido a la presión ejercida por la prensa, en ocasiones ese desgaste puede provocar problemas en los procesos OCR e identificar erróneamente una *"i"* por una *"t"* o una *"c"* por una *"e",* y reconocer *"G-c-n-t-c-s"* por *"G-e-n-t-e-s"* (imagen 8).

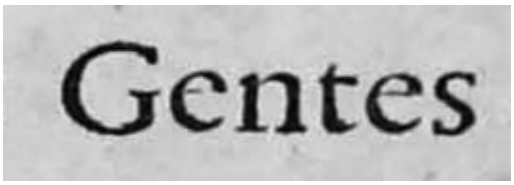
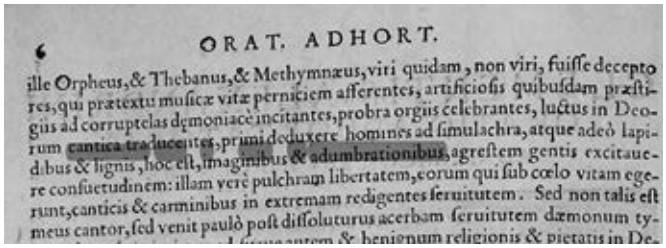

Gentes

Imagen 8. Tipos desgastados.

Otro problema que merece especial atención es el efecto de *curvatura* que se presenta al abrir un libro. Este efecto obedece a la naturaleza misma del documento,

[8] Hildelies Balk, Lieke Ploeger, aborda este tema de manera más extensa en su texto IMPACT: working together to address the challenges involving mass digitization of historical printed text.

debido a que algunas encuadernaciones conservan costuras muy tensas que dificultan la apertura del libro para la captura de las imágenes. El resultado de este fenómeno es una imagen con letras curveadas (imagen 9), lo que imposibilita más el reconocimiento de los caracteres y su reconstrucción de acuerdo con su posición original.


ORAT. ADHORT.

ille Orpheus, & Thebanus, & Methymnæus, viri quidam, non viri, fuisse decepto
res, qui prætextu musicæ vitæ perniciem afferentes, artificiosis quibusdam præsti-
giis ad corruptelas dæmoniacè incitantes, probra orgiis celebrantes, luctus in Deo-
rum cantica traducentes, primi deduxere homines ad simulachra, atque adeò lapi-
dibus & lignis, hoc est, imaginibus & adumbrationibus, agrestem gentis excitauc-
re consuetudinem: illam verè pulchram libertatem, eorum qui sub cœlo vitam ege-
runt, canticis & carminibus in extremam redigentes seruitutem. Sed non talis est
meus cantor, sed venit paulò post dissoluturus acerbam seruitutem dæmonum ty-

Imagen 9. Texto curveado.

Sumados a estos problemas se encuentran aquellos donde nuestro software no posea la opción para reconocer, por ejemplo, caracteres griegos. Así sucede en libros donde frecuentemente se presentan confrontados texto latino con texto griego, como en muchas obras de la tradición clásica impresas a lo largo del Renacimiento. En estos casos las letras griegas se confunden con letras del alfabeto latino: la *"η"* (eta) por *"n"*, la *"ρ"* (rho) por *"p"*, la *"γ"* (gamma) por *"y"*, etcétera.

El desempeño de otros proyectos: la herramienta de búsqueda en Google Books®

Como parte del desarrollo del presente trabajo se decidió realizar un breve comparativo de nuestros resultados con la herramienta de búsqueda y recuperación de texto implementada por Google Books® que, entre su vasto universo de documentos digitalizados, contiene libros similares a los que residen dentro del Fondo Antiguo de la Biblioteca Central, UNAM. Como resultado obtuvimos que la herramienta de Google® sí logra identificar correctamente nuestro más común problema: la *"s"* larga[9] (imagen 10).

¿Cómo logra Google Books® esquivar éste y otros problemas de OCR para los libros antiguos? La respuesta la tiene Luis Von Ahn, científico y profesor de ciencias de la computación en Carnegie Mellon University y fundador de la compañía Recaptcha[10], quien desarrolló la prueba desafío-respuesta *captcha,* ideada para demostrar la interacción de humanos en los protocolos de uso de Internet y proteger a las páginas web de fraude o *spam*. Esta prueba nos pide interpretar una imagen distorsionada de un conjunto de letras y números para teclearlos junto con los registros de formularios en línea o al realizar una compra por medio de Internet (imagen 11). Tras la compra de la tecnología *captcha* por parte de Google® en el 2009, se convirtió en la herramienta

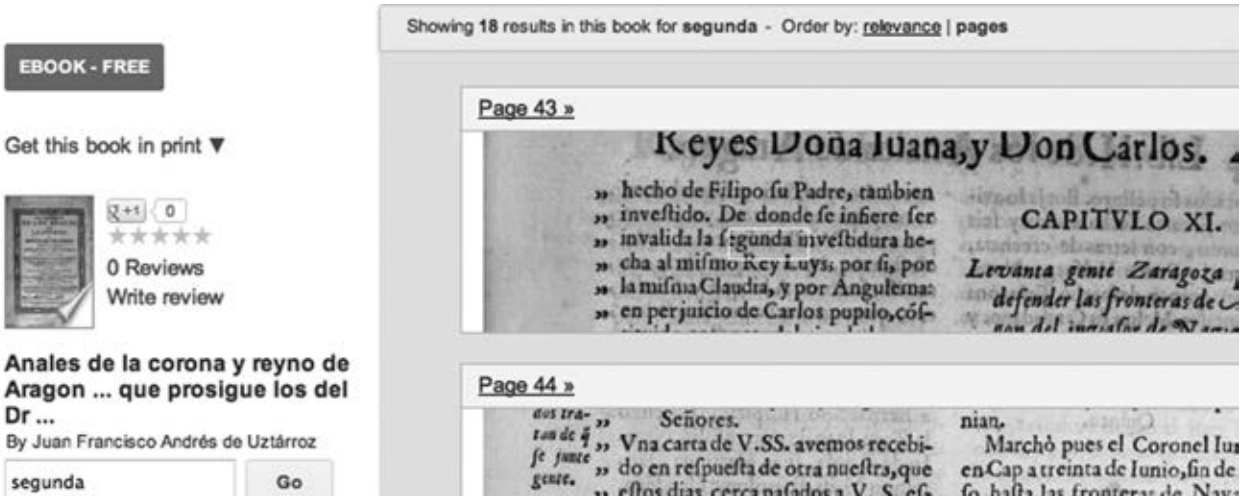


Imagen 10. Búsqueda de "segunda" dentro del texto Anales de la corona y reyno de Aragon. Zaragoza, 1663.

[9] Captura de pantalla en: http://books.google.com.mx/books?id=XeFFAAAAcAAJ&q=segunda

[10] *ReCaptcha* [en línea]. < http://www.google.com/recaptcha>

de reconocimiento de caracteres para su proyecto de digitalización que muestra al usuario, mediante la interfaz de captura, todas aquellas palabras que las computadoras son incapaces de reconocer. De este modo, cada vez que una persona teclea en el Recaptcha® un conjunto de caracteres para acceder a una página web inconscientemente alimenta su base de datos, digitalizando alrededor de 100 millones de palabras diarias, gracias a la participación involuntaria de 900 millones de usuarios por día alrededor del mundo.[11]

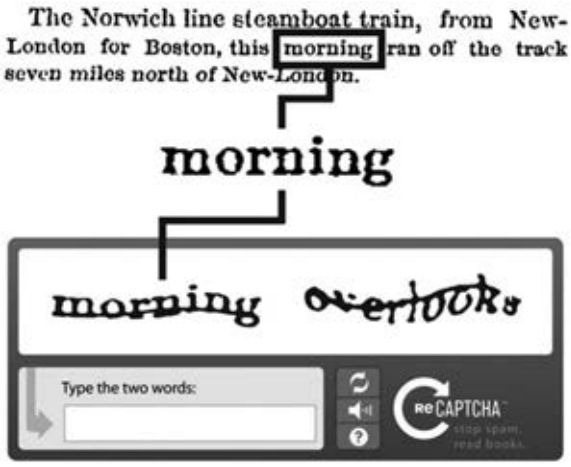


Imagen 11. Interfaz de captura de la prueba Recaptcha.

Conclusiones

En este trabajo se presentaron los factores generales que permiten determinar el eje de problemas comunes en la recuperación de texto por medio del OCR del software Acrobat Reader 8 Professional®. Aunque se aplicó un ejercicio sólo a tres casos, de aproximadamente dos millones de imágenes generadas en la digitalización del Fondo Antiguo de la Biblioteca Central de la UNAM, ello permite hacernos una idea sobre los retos por vencer si es que se planea su aplicación a la totalidad del conjunto de obras digitalizadas.

Por un lado, entender los procesos del software empleado, y comprender paso a paso la manera como transforma las imágenes en caracteres legibles por computadora, es de gran ayuda para identificar tanto los alcances como los límites de la herramienta y conseguir los mejores resultados de ésta; por el otro, comprender el tipo de documentos que se someterán a este proceso sirve para especificar los requerimientos técnicos de las imágenes generadas en cualquier otro proyecto de digitalización que busque la aplicación del OCR a sus imágenes.

Los objetivos de la digitalización de los libros antiguos de la DGB fueron la preservación y difusión de estas piezas, por lo tanto, exigió imágenes a color que capturaran la mayor cantidad de detalles inscritos en los documentos. Pero para los procesos OCR esos detalles, como anotaciones manuscritas, pequeñas manchas, incluso la transparencia del texto impreso en el anverso de las páginas, provocaron *ruido* que dificultó la recuperación de un porcentaje elevado de palabras.

Es preciso aclarar que el objetivo que se persigue con la implementación del proceso OCR en estos libros es lograr la recuperación de términos o fragmentos de palabras dentro de un vasto conjunto de documentos digitalizados mediante la búsqueda libre. Teniendo en cuenta esto, podemos apuntar que, a pesar de que exista un porcentaje determinado de caracteres bien reconocidos –sea elevado o no–, las búsquedas no se realizan por letras aisladas, sino por palabras completas; de tal manera que una letra mal identificada vuelve irrecuperable una palabra completa. Ahora bien, cabe preguntarse qué tanto afectan a la recuperación de términos los errores de identificación en palabras de menor valor semántico como artículos, conjunciones, preposiciones o adverbios. Además, la correcta identificación de grafemas por los procesadores OCR no significa necesariamente una recuperación del texto que cubra el 100% de los casos de búsqueda. Por ejemplo, la falta de normalización ortográfica hasta la segunda mitad del siglo XVIII[12] para diferenciar entre el

[11] VON AHN, Luis. *Utilizando el poder de millones de mentes humanas* [en línea]. TEDX Rio de la Plata, 1 noviembre 2011. <http://www.tedxriodelaplata.org/videos/utilizando-poder-millones-mentes-humanas>

[12] MARTÍNEZ, MARÍN, Juan. La ortografía española [en línea]: perspectiva historiográfica. *CAUCE*, 1992, no. 14-15, p. 125-134. <http://cvc.cervantes.es/literatura/cauce/pdf/cauce14-15/cauce14-15_11.pdf>

grafema "v" con valor vocálico y consonántico produjo textos donde aparece tanto *"Tesaurus"* como *"Tesavrus", "Segvunda"* como *"Segunda", "Vltima"* como *"Ultima";* incluso, si nos adentramos a estudiar los textos latinos nos encontraremos con el empleo de los grafemas *"i"* y *"j"* con el mismo problema, o el empleo de grafemas especiales para diptongos como "æ" (a-e), "œ" (o-e). Los casos se amplían si consideramos también que por economía del espacio dentro de una página un impresor podía elidir algunas letras de ciertas palabras conocidas por contexto, como *"homes"* por *"homines"* o *"Dmūs"* por *"Dominus".*

Las comparaciones nos ayudan a tener una mejor perspectiva de los resultados obtenidos. La ventaja de recuperación con Google Books® radica en el empleo de una herramienta que no delega toda la responsabilidad a una computadora, sino involucra la participación de seres humanos en la conversión de textos análogos a digitales. Para nuestro caso es preciso comprender la naturaleza de los documentos históricos, los procesos de su digitalización y el desarrollo de software cada vez más innovador para la recuperación de texto, esto nos permitirá implementar las herramientas correctas que permitan rescatar el legado cultural vertido en estos libros para que sea incluido como una fuente de información accesible en la generación de nuevo conocimiento.

Obras consultadas

Acrobat OCR [en línea]: *make your scanned documents searchable. Acrobat blog insights, trends, news and highlights on all things Acrobat.* <http://blogs.adobe.com/acrobat/acrobat_ocr_make_your_scanned/> [Consulta: mayo 2012].

Hildelies, Balk, Lieke Ploeger. IMPACT: working together to address the challenges involving mass digitization of historical printed text. *OCLC Systems&Services,* 2009, vol. 25, no: 4, p.233–248.

Llevando la teoría a la práctica [en línea]: *tutorial de digitalización de imágenes.* Biblioteca de la Universidad de Cornell, Departamento de Investigación. <http://www.library.cornell.edu/preservation/tutorial-spanish/conversion/conversion-04.html> [Consulta: mayo 2012].

Martínez, Marín, Juan. La ortografía española [en línea]: perspectiva historiográfica. *CAUCE*, 1992, no. 14-15, p. 125-134. <http://cvc.cervantes.es/literatura/cauce/pdf/cauce14-15/cauce14-15_11.pdf> [Consulta: 7 mayo 2012].

Nafiz Arica, Fatos T. Yarman-Vural. Optical Character Recognition for Cursive Handwriting. *IEEE,* 2002, vol. 24, no. 6, p.801-813.

OCR [en línea]. Encyclopædia Britannica: Science & Technology. Britannica Academic Edition. <http://www.britannica.com/EBchecked/topic/430371/OCR> [Consulta: 7 mayo 2012].

ReCaptcha [en línea]. <http://www.google.com/recaptcha> [Consulta: mayo 2012].

Tutorial [en línea]: *Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality.* The Comission on Preservation and Access. <http://www.clir.org/pubs/reports/pub53/pub53.pdf> [Consulta: mayo 2012].

Von Ahn, Luis. *Utilizando el poder de millones de mentes humanas* [en línea]. TEDx Rio de la Plata, 1 noviembre 2011. <http://www.tedxriodelaplata.org/videos/utilizando-poder-millones-mentes-humanas> [Consulta: mayo 2012].