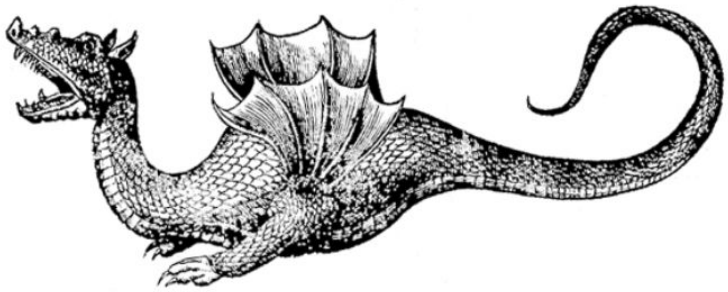
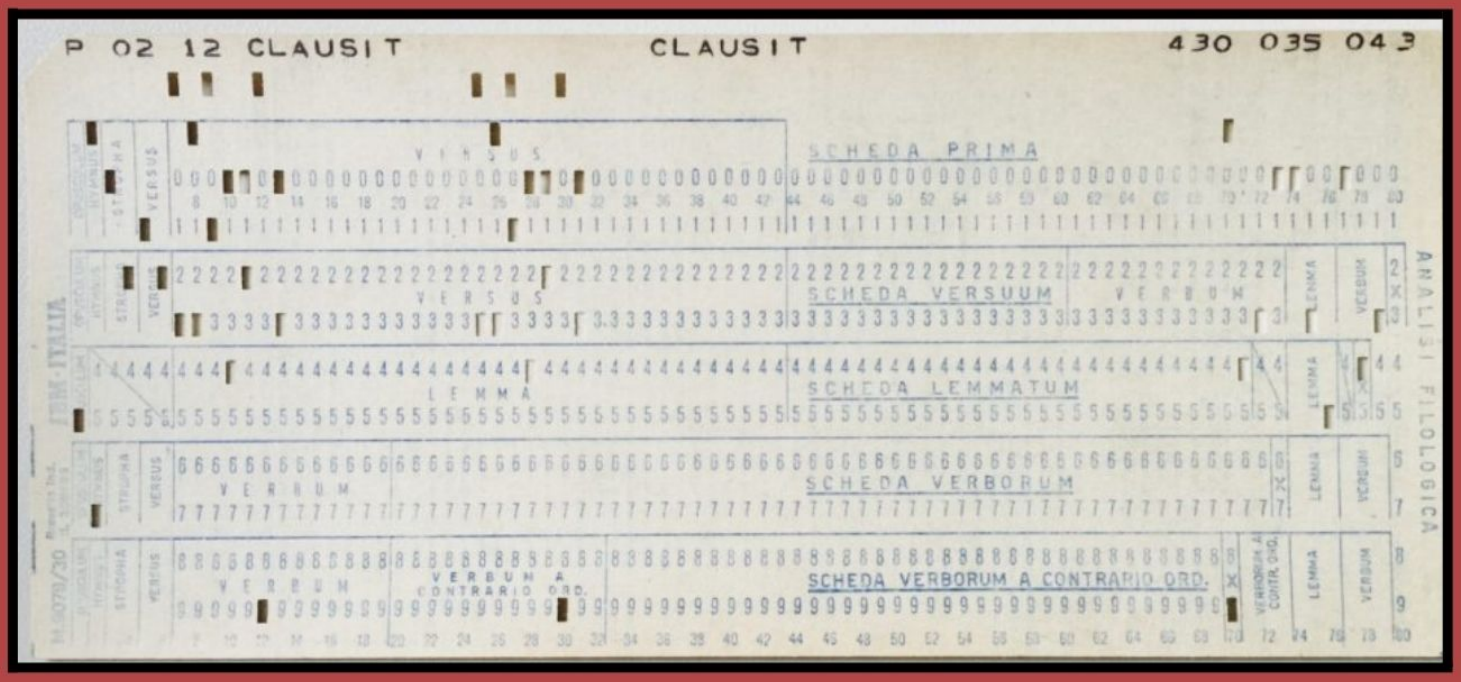
HISTORIAS FINGIDAS



Número Especial 1 (2022)



Humanidades Digitales y estudios literarios hispánicos

ed. Stefano Bazzaco

PROGETTO MAMBRINO

dir. Anna Bognolo e Stefano Neri
Università di Verona
Dipartimento di Lingue e Letterature Straniere

Imagen de portada: *Punched Card from Index Thomisticus project*
fuente: G. Rockwell, M. Passarotti «The Index Thomisticus as a Digital Humanities Big Data Project», *Umanistica Digitale*, 5 (2019), p. 22.

PROGETTO MAMBRINO
Dipartimento di Lingue e Letterature Straniere
Università degli Studi di Verona

HISTORIAS FINGIDAS
Il romanzo cavalleresco spagnolo come punto di osservazione per lo studio del romanzo europeo d'*ancien régime*.

NÚMERO ESPECIAL 1: HUMANIDADES DIGITALES Y ESTUDIOS LITERARIOS HISPÁNICOS

ed. Stefano Bazzaco

Tabla de contenidos

Editorial

Palabras recobradas

Monográfica

ISSN 2284-2667

Department of Foreign Languages and Literatures of the University of Verona | Declaración de privacidad

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Presentación

Stefano Bazzaco

(Università di Verona)

> Che rivelazioni di massa suscitino una problematica in qualche modo diversa da rivelazioni limitate, è fatto d'esperienza normale in ogni campo scientifico. Perciò l'accelerazione e l'incremento quantitativo sicuramente permessi al nostro lavoro dall'uso d'elaboratori elettronici, difficilmente potranno restare senza conseguenze anche qualitative, ossia d'ordine metodologico. Sulla portata di tali conseguenze sembra tuttavia prematuro avanzare previsioni che non siano del tutto generiche o al massimo tendenziali (Aurelio Roncaglia, «Le due culture», in *Almanacco Bompiani*: *Le applicazioni dei calcolatori elettronici alle scienze morali e alla letteratura*, 1962, 316).

En 1962, en el primer volumen de los Almanaques Bompiani expresadamente dedicado a la aplicación de técnicas computacionales a la investigación literaria, aparecía la encuesta «Las dos culturas» (143-144 y 313-317). La sección, cuyo título evidentemente aludía a la conferencia de Charles P. Snow de mayo de 1959 en Cambridge, volvía sobre el problema de las complejas relaciones entre ciencias duras y ciencias blandas, configurándose como experiencia fundacional de una reflexión que acompañaría el asentamiento de las Humanidades Digitales en Italia. Allí, en el espacio de pocas páginas, se les preguntaba a eminentes filólogos de la talla de Cesare Segre, Gianfranco Contini o Aurelio Roncaglia, entre otros, acerca del impacto epistemológico que los ordenadores posiblemente tendrían en los estudios humanísticos tradicionales. Los intelectuales entrevistados, que se dividían entre escépticos e ilusionados, trazaban los contornos de una disciplina híbrida, capaz de mezclar en una curiosa amalgama lo cuantitativo con lo cualitativo.

En la actualidad, 60 años después de la publicación de dicha encuesta, la idea de integrar el uso del ordenador a los estudios literarios que se debatía en esas páginas ha dado varios frutos. Dentro de la línea evolutiva

Stefano Bazzaco «Presentación», *Historias Fingidas*, Número Especial 1 (2022) Humanidades Digitales y estudios literarios hispánicos, pp. 1-4.
DOI: https://doi.org/10.13136/2284-2667/1229 - ISSN: 2284-2667.

de una disciplina que aseguraba combinar la automatización de los métodos con la indagación de los textos, las investigaciones de los últimos 30 años han representado la fase inaugural de un trayecto supuestamente capaz de aportar mayor cientificidad en el interior de estos estudios gracias a la formalización del lenguaje, un ingrediente crucial de la renovación, porque imponía repensar los fundamentos de la textualidad. Sin embargo, estas disciplinas siguen experimentando cierta dificultad frente al rápido desarrollo tecnológico y de los medios de comunicación, desarrollo que fue acompañado por el florecimiento de un enorme arsenal de herramientas computacionales que iban a respaldar ese cambio, con el resultado de que las Humanidades Digitales en los últimos diez años se han convertido en una disciplina «*tool*-céntrica», donde la marcha triunfal de las competencias informáticas a veces reemplaza la interpretación escrupulosa de los datos.

Con el intento de promover el diálogo entre distintos proyectos de literatura ibérica que veían en la incorporación del medio digital la posibilidad de ampliar sus propios horizontes de investigación sin renunciar a la coherencia y al rigor de los métodos filológicos tradicionales, a principios de 2021, varios jóvenes estudiosos del Progetto Mambrino (Univ. de Verona), del proyecto BIDISO (Biblioteca Digital Siglo de Oro, Univ. da Coruña) y del grupo de investigación COMEDIC (Catálogo de obras medievales impresas en castellano, Univ. de Zaragoza) abrieron un espacio de discusión que permitió engarzar entre ellas distintas experiencias surgidas en el contexto de las Humanidades Digitales. El resultado final de este proceso fue la organización del congreso *Humanidades Digitales y estudios literarios hispánicos. De los impresos de la Edad Moderna a las ediciones académicas digitales*, que tuvo lugar en Verona los días 22-23 de junio de 2021 en forma telemática, debido a las restricciones impuestas por la pandemia. En él participaron experimentados estudiosos de letras hispánicas y jóvenes investigadores que estaban guiados por el común deseo de indagar los posibles empleos del ordenador en el marco de proyectos de investigación filológica y literaria que contaban ya con una tradición larga y consolidada.

La presente publicación es fruto de los debates que surgieron durante el congreso. En estas páginas, de hecho, el lector encontrará unas posibles

claves para la aplicación rigurosa y apropiada de herramientas informáticas a los estudios humanísticos, en particular por lo que atañe a la edición de obras de la Edad Moderna y su remediación en el espacio digital.

Esta entrega, según la tradición de nuestra revista, empieza con la sección *Parole Ritrovate*, donde se propone un artículo del mismo Almanaque ya mencionado del 1962, escrito por el jesuita Padre Roberto Busa, precursor de la linguística computacional, que trata la automatización del lenguaje y, de forma premonitoria, apunta a unos posibles caminos para el estudio cuantitativo de la literatura. A pesar de la distancia que nos separa del trabajo de Busa, rescatar sus palabras permite volver a los fundamentos de las Humanidades Digitales, es decir la atención hacia la producción y explotación supervisada de los datos textuales. Con la idea de poner al alcance de los estudiosos españoles el texto del jesuita, Soledad Castaño Santos realizó una traducción del artículo al castellano, con una introducción que permite definir el contexto en que se generaron esas reflexiones.

La sección monográfica está compuesta por nueve artículos, dispuestos según un orden que idealmente recorre las distintas fases de trabajo editorial en un entorno digital. Inaugura este apartado el artículo de José Manuel Fradejas Rueda, quien señala de qué forma las herramientas computacionales supusieron para él un cambio determinante en la configuración de los procedimientos ecdóticos, partiendo de la digitalización de las fuentes y llegando a la colación de testimonios con métodos estadísticos.

El trabajo de Fradejas, que sigue de cerca las distintas fases evolutivas de las Humanidades Digitales en España, constituye el punto de acceso ideal para introducir un primer grupo de artículos que atañen a la transcripción automatizada y su integración en proyectos literarios de largo alcance. En el primero de ellos, Mónica Martín Molares, Ana Jiménez Ruiz, Ángela Torralba Ruberte y Stefano Bazzaco presentan unos modelos de reconocimiento de caracteres desarrollados con la plataforma Transkribus (READ Coop) y que permiten la transcripción automática de textos en gótica y redonda de los siglos XV-XVII. Siguen unas contribuciones de Nuria Aranda García, quien repasa las posibilidades ofrecidas por los sistemas de reconocimiento de textos en relación con la

base de datos COMEDIC, y de Manuel Ayuso García, quien indaga las distintas técnicas de explotación de la transcripción automática de impresos latinos dentro del proyecto BECLaR. Cierra la sección el trabajo de Giada Blasut, que presenta los resultados de una primera experimentación con el modelo de reconocimiento para la letra gótica en relación con su edición del *Silves de la Selva* (1546) de Pedro de Luján, libro doceno del ciclo de Amadís de Gaula.

El segundo conjunto de artículos atiende a la fase de modelización de ediciones digitales en formato XML TEI. Encabeza esta sección el ensayo de Antonio Rojas Castro sobre la realización de una edición digital académica de las *Soledades* gongorinas: el investigador retoma en esta ocasión el trabajo realizado durante los años del doctorado, ofreciendo una significativa muestra de las recientes posibilidades de personalización, visualización y uso de ediciones en la red. Sigue un trabajo colaborativo realizado por Carlota Fernández Travieso y Manuel Garrobo Peral centrado en la ampliación del portal de la Biblioteca Digital Siglo de Oro, gracias a la creación del repositorio BIDISO TEXTOS, que agrupará las ediciones académicas digitales de Relaciones de Sucesos y otras obras de interés del proyecto.

Concluyen este volumen dos incursiones en el campo del tratamiento computacional de los motivos literarios. En el primero de ellos, Federica Zoppi detalla los fundamentos teóricos para la creación de una base de datos de motivos caballerescos que se integrará en la Biblioteca Digital del Progetto Mambrino. En el artículo sucesivo, Giulia Tomasi relata los primeros pasos del proyecto MeMoRam en el campo de la minería de textos, apuntando al desarrollo de herramientas digitales para la detección semi-automática de motivos en el corpus de los libros de caballerías castellanos.

Mis agradecimientos van a todos los participantes al congreso, sobre todo a Anna Bognolo y Stefano Neri, por la oportunidad que me ofrecen de inaugurar una nueva sección de nuestra revista con la publicación de este primer número especial dedicado expresamente a las Humanidades Digitales.

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

L'analisi linguistica nell'evoluzione mondiale dei mezzi di comunicazione

Padre Roberto Busa s.j.*

§

I.

1. Il fenomeno linguistico è più grande di noi: uno degli ingredienti di quella strana formula di impasto che ciascuno di noi è. I valori infatti di cui noi siamo un così fragile e meraviglioso congegno sono in se stessi ben più diffusi e ben più grandi che noi stessi. Le mani per esempio servono a noi per tante semplici o complicate cose: ma sono, come i camerieri, per dir così, sempre alle nostre spalle: le adoperiamo senza farci molta attenzione. Se però ce le mettessimo sotto gli occhi e le scrutassimo e pensassimo un po' anche a loro, ci troveremmo di fronte a tutto un mondo da scoprire. Altro mistero è per esempio la nostra capacità di gusto estetico. In virtù di quale «programma», caricato in quel robot che siamo noi, noi sentiamo così prepotente il bisogno ad esempio della simmetria, la ripugnanza a ogni stonatura di colore, di linee, di suoni? Ma le vere mani nostre sono i nostri poteri espressivi: con i gesti, col volto, con le arti, con le

* Il presente articolo fu pubblicato in origine in *Almanacco Bompiani*: *Le applicazioni dei calcolatori elettronici alle scienze morali e alla letteratura*, 1962, pp. 103-108, 117. La trascrizione che si offre rispetta nel possibile l'articolo originale, con minimi interventi per quanto riguarda la formattazione del testo. I nostri più sentiti ringraziamenti vanno a Padre Francesco Pecori Giraldi, legale rappresentante della Provincia di Italia della Compagnia di Gesù; a Maria Macchi, responsabile dell'archivio della Provincia EUM della Compagnia di Gesù; a Marco Passarotti, docente dell'Università Cattolica di Milano ed erede del lavoro di Padre Busa. A loro va la nostra gratitudine per la generosità e l'entusiasmo dimostrati verso questa iniziativa.

Padre Roberto Busa s.j., «L'analisi linguistica nell'evoluzione mondiale dei mezzi di comunicazione», *Historias Fingidas*, Número Especial 1 (2022), pp. 5-17.
DOI: https://doi.org/10.13136/2284-2667/1227 - ISSN 2284-2667.

parole noi feriamo e medichiamo, rovesciamo ed eleviamo, miglioriamo o guastiamo tutto attorno a noi. Anche questi sono, dentro di noi, mondi da esplorare.

2. In quel nostro parlare, che abbiamo in bocca e che così poco conosciamo, vi sono tre strati: ciò che è presente al campo di coscienza, ciò che è subconscio e ciò che è affatto inconscio. E in quella stessa zona del nostro linguaggio che viene illuminata dalla nostra percezione e attenzione, una parte, ma non tutto, è passibile di controllo, nel senso inglese della parola: una parte può cioè essere governata e perciò anche educata da noi. È per lo meno teoricamente possibile a un milanese decidersi ad abituarsi a dire «vada», in luogo di quello scorretto «vadi», cui chissà per quali ataviche ereditarietà si mostra così affezionato! Altri settori vi sono che sfuggono sì a un controllo organizzativo, ma non del tutto a un rilevamento sistematico: non riusciamo a cambiarli, ma comunque arriviamo a rendercene conto, sia pure in qualche misura. Altre zone infine ubbidiscono soltanto al subconscio o addirittura all'inconscio. Per esempio solo con molta sottigliezza si arriverà a renderci conto che se noi preferiamo alcune parole ad altre lo facciamo perché comandati da una subconscia aspirazione al fare la cosiddetta bella figura, e questo in conseguenza di un maggior valore che noi aggiudichiamo a certe parole, così come fanno le signore con certe parole del tipo di «genare» o «flattare», mentre altri sceglierà le parole avendo per criterio la loro capacità definitoria, e altri ancora avendo per metro soltanto la loro esteticità, fonetica o semantica o di correlazione e ritmi. Ma su un piano più profondo, le strutture grammaticali e sintattiche paiono sgorgate dalle radici inconsce con le quali l'umanità succhia la propria evoluzione vitale da quell'universo in cui si agita, per così poco tempo!, come un'ameba nel suo brodo di cultura. Le basi del linguaggio si trovano tra le zone del comportamento umano che sono inaccessibili alla educazione e al self-control perché programmate e comandate esclusivamente da quanto sta alle radici della nostra fisiologia e della mescolanza, spiacevole e inevitabile, di fisiologia e patologia.

3. Non tutto dunque nei nostro parlare si lascia conoscere: di quel tanto o poco che si lascia conoscere, non tutto si lascia influenzare dalla nostra aggressiva ambizione di fare anche con noi e di noi «quello che vogliamo noi». Tuttavia perché non potremmo lasciare incolta anche quell'aiuola che pur riusciremmo, volendo, a vangare? Perché voler sottrarre al fiume delle nostre parole qualche filone di acqua per avviarlo entro condotte forzate? Esattamente per la stessa ragione per la quale cerchiamo di controllare l'acqua. Il parlare è infatti il principale potenziale di energia di cui l'uomo dispone e va quindi erogato economicamente. Le idee sono forza, solo quando si possono dire e scrivere. Né hanno altro tramite, per far presa fuori dell'individuo che le ha.

4. Aristotele dunque si è messo di buzzo buono a guardarci dentro e fra le pieghe del linguaggio ha scoperto la metafisica. E per quell'enciclopedico e positivo rilevatore di fatti che egli ha dimostrato di essere, questo è stato uno dei più clamorosi: sentirsi catapultato, dalla pista percorsa palmo a palmo da lui con l'indagine positiva, a cabrare verso l'alto: anche il placido e buon S. Tomaso d'Aquino lo stette a contemplare, ammirando col naso in aria la potenza con cui un pagano era riuscito da questa terra a penetrare il cielo. Ma Filone, conoscitore del Vecchio Testamento, e la Teologia Cristiana, partiti dall'esame della «parola», erano penetrati anche più in là dei cieli, superando di molte lunghezze la gran corsa di Aristotele e Platone. Tra le parole avevano intravisto il barbaglio del Logos, Verbum. E non c'è stato idealista assoluto che sia riuscito a osare tanto quanto ha fatto il filosofo cristiano nella enucleazione del valore d'espressione del «verbum mentis» e del conseguente riverberarsi di mutuo amore, all'interno di quel pensiero assoluto che è incendio di consistenza, vita e fantasia: regista il quale è insieme arco voltaico che proietta sullo schermo buio del nulla quel succedersi di immagini che siamo noi, il mondo e la storia.

5. Nella vita sociale, la grammatica e l'analisi logica hanno educato per tanti secoli quell'indefinibile non so che, che noi chiamiamo umanità e umanesimo: quel pizzico cioè di gusto del bello, di senso dell'armonia, quell'apprezzamento di valori formali, per cui anche al Politecnico sussiste la differenza fra chi viene dal liceo classico e chi viene da altre scuole. La

retorica aveva educato all'arte di esprimersi. Il vecchio Aristotele aveva ben detto che «signum scientis est posse docere»: le nostre conoscenze sono mature quando riusciamo a trasmetterle. Tutti abbiamo esperimentato che quando il professore impiega due ore a farci capire qualche cosa, è perché egli stesso non la possiede ancora perfettamente, e che per riuscire a possederla perfettamente non avrebbe dovuto far altro che prepararsi prima a dirla. Quante volte abbiamo visto nella vita l'inesauribile saggezza del detto che vi è una enorme distanza tra l'aver ragione e il saper farsela dare! Oggi ci si preoccupa spesso solo di far inghiottire nozioni, quasi l'uomo sia un magazzino generale, o quasi in lui non ci sia altro che memoria. Mentre l'uomo è, soprattutto e almeno per destinazione, capacità organizzativa e inventiva, e bisognerebbe non educarlo come uno zaino da rimpinzare secondo la lista di quanto occorrerà poi al campeggio, ma rifinirlo nei suoi congegni, lubrificarlo, rodarlo come una macchina utensile che sia in grado di lavorare poi a lungo su qualsiasi materiale. Sbaglio forse a pensare che pagherebbe la spesa di sapere la metà di quello che sappiamo, se con ciò si riuscisse a dire meglio quel poco che sappiamo? La cura dunque dei nostri mezzi espressivi esisteva una volta molto di più di oggi. Si educava a organizzarsi interiormente ad adeguare la combinazione di parole allo scopo voluto: a pensare cioè a come parlare prima di parlare. Ma già, anche il Manzoni diceva che questa sola cosa «pensare prima di parlare» è da sé sola così difficile, che anche noi siamo un tantino da scusare quelle tante volte che ci abbandoniamo a parlare così come capita.

6. Piano piano le universali leggi dell'invecchiamento, le quali intaccano le istituzioni come l'uomo e la natura, hanno usurato il mordente dell'analisi logica e della retorica. Il potere di decadenza ha esercitato la sua tirannia a tal misura che oggi è solo per il teatro che si va a scuola di recitazione, ma non per prepararci tutti a recitare il nostro copione nelle commedie e tragedie della vita. E se è stato scritto che la divinità del Vangelo è dimostrata se non altro dalla sua sopravvivenza alle spiegazioni domenicali, se cioè voi borghesi trovate che noi preti siamo spesso così sciatti nelle nostre prediche, è perché noi come voi siamo figli del nostro tempo. E il tramonto si presenta violaceo, anche perché in Italia si sta trattando il

latino come un vecchio nonno al quale si augurano altri cent'anni di vita, mentre il subconscio registra che, a pensare che tra poco ne saremo senza, non se ne sente in fondo un orrore per la verità infinito.

7. A questo punto è intervenuto il mostro della notte, il tecnicismo trionfante, con la sua ultima creatura: l'automazione. Qualcuno ha rabbrividito, pensandola come un crudo e duro bulldozer che procede ruggendo, schiacciando e stracciando i fiori. Tra questi, vittima delicata e gentile, l'umanesimo. Il domani è già qui. Il futuro è già cominciato: una colata di lava allaga e brucia i fianchi verdi della montagna. Nella torretta di comando del mostro, incapsulati tra manometri, cloches, luci-spia e quadranti, vi sono degli uomini. Forse all'inizio non si sono nemmeno accorti degli alti lai e lamenti ululati elegiacamente dagli «umanisti». Si accontentano infatti di… lavorare. Pretendono di prestare un servizio di pubblica utilità, poiché ritengono che senza di loro l'industria e il commercio non potrebbero più rispondere ai bisogni dell'uomo. Ma poi – non sono ancora passati dieci anni – gli uomini dell'automazione hanno cominciato a sporgere il capo dalla cabina della torre dell'elettronica, per rivolgere ai filologi e ai grammatici, occupati nei campi a scegliere fior da fiore, domande di questa natura: Di grazia, quanti sono in russo i verbi attivi transitivi e quanti quelli attivi intransitivi? Quanti sono in inglese? Qual è il maggior numero di lettere iniziali e finali in cui coincide il maggior numero di parole? Quali parole o situazioni linguistiche si trovano entro un raggio di n parole, solo quando e sempre quando «faccia» vuol dire volto, e quali altre solo e sempre quando «faccia» è voce del verbo fare? E ancora: Di grazia, la mi vuol raggruppare tutte le parole del vocabolario secondo le varie categorie morfologiche e grammaticali? Mi dica tutte le parole che si possono omettere, e quando, così da accorciare un testo senza scapito della sua espressività. Mi sa dire caso per caso l'ambientazione caratteristica di certe categorie semantiche che non sono né morfologiche né sintattiche né strutturali? È successo cioè un fatto clamoroso: la macchina che ci ha resi consapevoli che nessun umanista possiede la sua lingua così da saper dare una risposta a simili domande. La macchina, donna di servizio del banale commercio e della greve industria, ha documentato che di umanesimo, di quello serio e sistematico, ce n'è ancora troppo poco. I fatti economici

esigono oggi un incremento qualitativo delle scienze grammaticali e lessicali: come una delle necessità del loro sviluppo vitale. Ma ne offrono anche la possibilità. Il che non è stata piccola rivincita né piccola soddisfazione.

II.

8. Il Centro di Gallarate è ancora oggi il centro che nel mondo ha trasportato su schede la più grande quantità di parole: sono oramai quasi quattro milioni, e in continuo aumento. Si tratta di 7 lingue, (Aristotele, Antichi Italiani, Dante, Kant, Goethe, Testi Ebraici del Mar Morto, Fabbri, ecc) in tre alfabeti, latino greco ebraico. Ma quando nel 1946 cominciai a pensare sul serio agli indici verbali dei tredici milioni di parole di S. Tomaso d'Aquino, e quando più tardi nel 1949 iniziai i primi esperimenti con la IBM e ancora quando nel 1951 ne pubblicai i primi risultati, ero non soltanto il solo e il primo nel mondo che si avventurasse a insellare la lessicologia sull'ippogrifo, ma ero anche ignaro del momento storico in cui ciò mi capitava. L'aver avuto per primo un'idea non è un merito, ma un caso. Se non veniva a me, l'idea veniva certamente a qualche altro. E magari un giorno salterà fuori che prima di me era venuta in mente a qualche altro, al quale nessuno allora aveva fatto attenzione. E se si potesse parlare di merito, questo se mai consisterebbe nella lunga pazienza che ci vuole a risolvere passo passo tutte le difficoltà e gli imprevisti che si incontrano nel trasformare un'idea in una metodologia matura e pratica, applicabile per dir così a produzione in serie. Della celebre frase «genius is one per cent inspiration, ninety-nine per cent perspiration» (il genio è fatto per l'uno per cento d'ispirazione, per il novantanove per cento di sudore) l'unica parola che è certo che io non verifico è solo la prima. Ma chi andava allora a immaginare che le macchine a schede sarebbero oggi state considerate antiche, e che avremmo visto l'evoluzione o meglio metamorfosi dei calcolatori elettronici dalle memorie a superficie patinate di ossido di ferro, a quelle di reticolati dì anelli di ferrite e infine a quelle criogeniche (films sottilissimi sovrapposti a mo' di libro, utilizzabili a temperature vicino allo zero assoluto)? Non immaginavo certo che lo «stretch», costruito

per le ricerche nucleari, avrebbe posseduto una memoria di poco meno di due miliardi di posizioni, in cui tutta l'Enciclopedia Treccani potrebbe nuotare come un bambino nel lettone, e un'altra memoria di un milione e mezzo di posizioni che ha una velocità di accesso di qualche centesimo di milionesimo di secondo. Ma soprattutto ignoravo che venivo inserito nella successione dei passaggi, attraverso i quali l'automazione delle contabilità ha causato l'evoluzione mondiale dei mezzi d'informazione.

9. Posso condensare in quattro fasi il movimento che dopo il 1945 ha assunto l'accelerazione di una valanga. Primo stadio. – Lo sviluppo delle comunicazioni e delle tecniche organizzative ha permesso l'ingigantire di aziende che arrivano a coprire tutto il mondo. Altrettanto rapido è stato l'aumento della reciproca influenza dei mercati e tra politica e mercato. Con ciò è divenuto indispensabile per il dirigente di poter censire un gran numero di particolari, così da indurne velocemente delle sintesi: in tempo utile a controllare, e volendo modificare, l'andamento di grandi masse di piccoli ed estesi fenomeni periferici. I calcolatori risposero a questa necessità fornendo alla vita economica l'automazione della contabilità industriale e commerciale. Essi arrivano a svolgere fino a un milione di moltiplicazioni e divisioni al secondo. Giungono a stampare i risultati dei propri calcoli alla velocità di 60.000 righe all'ora per l'alfabeto e 300.000 per i soli numeri.

10. Secondo stadio. – L'industria, il cui sviluppo viene esasperato dalle esigenze della «difesa», e il parallelo infittirsi dei rapporti tra produzione industriale e ricerca scientifica, hanno imposto l'automazione del calcolo scientifico. L'Euratom, per esempio, si è sentito costretto ad acquistare per il proprio Centro di Ispra il calcolatore IBM 7090, che costa circa tre milioni di dollari, ossia quasi due miliardi di lire.

11. Terzo stadio. – Le attività di produzione, scambio e difesa, esigono dall'automazione l'«information retrieval», che io tradurrei come reperibilità tempestiva delle conoscenze utili. La quantità di pubblicazioni scientifiche, già enorme, è in continuo aumento. Gli Stati Uniti hanno una media oggi di 40.000 nuovi brevetti all'anno. D'altra parte la accelerazione

dell'evoluzione scientifica è tale che le pubblicazioni di fisica nucleare dopo due anni servono oramai solo alla storia della fisica. Ma per quanto concerne le tecniche degli elaboratori elettronici, probabilmente l'attualità utile delle notizie è una cresta d'onda di forse poco più che mezzo anno. Immaginate ora di avere bisogno per un'industria missilistica di conoscere il comportamento di determinati materiali in determinate nuove situazioni. Quanto tempo impiegherete a setacciare da tutto lo scibile delle scienze interessate quanto fa al vostro caso? Non vi serviranno gli indici analitici, perché voi per definizione ricercherete qualche cosa di non comunemente risaputo, né vi basteranno le indicazioni bibliografiche, poiché queste contengono solo i titoli, mentre voi, per la ragione ora detta, avete bisogno di frugare nel contenuto stesso di quanto viene stampato. Vi serviranno se mai gli abstracts. Ma provate a farli leggere tutti e mi direte se quando avrete finito non sarà troppo tardi. Come fare allora a tenersi al corrente con tutte le pubblicazioni di tutto il mondo quasi contemporaneamente al loro apparire? Mi pare che al DDT è successo di essere stato scoperto per la prima volta due o tre volte consecutive! Occorre perciò condensare un massimale di informazioni scientifiche in modo che si possa in un minimale di tempo individuare in esse tutto ciò che interessa la ricerca del nuovo. L'automazione ambisce di arrivarci.

12. I settori nei quali si è canalizzata sono: nuovi tipi di simbolizzazione delle conoscenze, ossia alfabeti a impressioni magnetiche; come trascrivere e ricopiare con questi nuovi alfabeti, che solo la macchina sa leggere, il contenuto di quanto è stampato con gli alfabeti a inchiostro su carta (si lavora accanitamente per riuscire a farlo a fotolettura, fonoscrizione, ecc.); come condensarlo (riassumerlo, ridurlo a stile telegrafico, abbreviare le parole); come classificarlo, come ricercarlo. Un capitolo di questo sforzo è rappresentato dalla traduzione automatica. Non dico la fantascienza di tradurre a macchina un testo letterario o filosofico, ma la tecnica di tradurre a macchina pubblicazioni contemporanee, sullo stesso argomento, in scienze unificate come lo sono oggi, pensate perciò ed espresse alla stessa maniera e con un vocabolario, le cui sole differenze consistano in quelle delle due lingue. Questa tecnica ai problemi di cui sopra aggiunge quello di come, sulla scorta di situazioni o fattori linguistici

presenti caratteristicamente nel contesto di una parola, se ne possano automaticamente individuare la funzione grammaticale e logica e, nei casi di polisemia, l'accezione in questo preciso luogo; e quello di come la macchina possa trasportare la sintassi di una lingua in quella di un'altra lingua. L'Università di Georgetown, Washington DC, ha aperto da un anno a Frankfurt/M un centro ove trenta persone perforano in continuità pubblicazioni scientifiche russe, che vengono poi tradotte in inglese dal calcolatore 704.

13. Quarto stadio. – L'automazione del trattamento dell'informazione esige l'automazione della compilazione di indici, di concordanze e di tutti i possibili tipi di statistica dei fatti linguistici. All'Euratom di Ispra visitate il gruppo Cetis. Andate a Washington al Georgetown Institute of Languages and Linguistics. Vi renderete conto come tra i ricercatori delle tecniche per il trattamento dell'informazione stiano sviluppandosi una lessicologia e una linguistica che sono più sistematiche, più esaurienti, più largamente utili, e oso dire, più umanistiche, dì quanto non lo siano state a tutt'oggi quelle tradizionali. E tra non molto, dagli orti dell'umanesimo le voci tenorili dei filologi orchestreranno le benemerenze dell'automazione, commentate baritonalmente dai matematici.

III.

14. Ma allora anche all'interno di quella massima espressione della nostra libertà, personalità, capricciosità, che è il nostro parlare, si trovano formule matematiche. Ed è proprio vero: non si riesce a parlare «come si vuole», senza ubbidire ad alcuna legge. Se vi abbandonaste alla voluttà di tirar fuori, al di là di certi confini, dal gran mare delle combinazioni che sono aritmeticamente possibili tra gli elementi del vostro vocabolario, certe sequenze di parole che sono inconsuete al di qua di quei confini, state sicuri che vi rinchiuderebbero da qualche parte per sottoporvi alla cura del sonno. Ma non è solo in questo senso che vi sono delle leggi nel parlare. Il numero – chissà che gioia ne proverebbe il buon Pitagora se fosse vivo

– è apparso struttura portante del linguaggio, così come proporzioni di misure e rapporti di rapporti sono lo scheletro delle formosità e del bello. E la statistica linguistica, di cui il nostro Davanzati si servì già secoli or sono, si sente tanto più incoraggiata in quanto il numero regna ancora tra i fondamenti delle idee e della logica, come dimostrano la logica simbolica e l'algebra delle proposizioni, così come appartiene alla sostanza del fondo e fonte dell'essere, come la teologia trinitaria cattolica mette in luce. Essendo poi il linguaggio traducibile in termini combinatori di una grande massa di piccoli elementi, essendo cioè esso un intrecciarsi di ripetizioni e di frequenze, la sua matematica non è solo quella deterministica, bensì e più ancora quella delle probabilità e del caso, matematica meravigliosa e maggiormente vicina al mistero di Dio, dello spirito e dell'arte. Giovanni Gioacchino Becher, morto nel 1682, poligrafo, coinvolto nelle gesta della teoria flogistica, ha meritato imperitura riconoscenza da parte della madre Germania per averle insegnato a cavar l'alcool fin dalle patate. Ebbene un uomo dagli interessi così vasti e così empirici, potrebbe essere nominato il precursore della codificazione numerica delle parole. Nel suo *Character pro notitia linguarum universali*, Francofurti 1660, egli ha proclamato che con una sola lingua si potranno comprendere tutte, alla condizione che ogni concetto venga espresso con una cifra o con un corrispondente geroglifico. Proprio tutto quello che occorre – e che in gran parte ancora manca e a cui si sta intensamente lavorando – affinché un qualunque elaboratore elettronico, digitale o analogico, possa servirci da traduttore fedele e riservato: quel calcolatore che i tedeschi chiamano Hochgeschwindigkeittrotel: cretino ad altissima velocità!

15. A che cosa possa servire la statistica dei fattori linguistici, estesa tanto largamente quanto lo permettono le incredibili possibilità dell'automazione, può essere illustrato dai pochi esempi che seguono.

A Gallarate, per conto dei Proff. Tagliavini e Croatto dell'Università di Padova, è stata compiuta automaticamente la trascrizione fonetica di un testo del Fabbri di circa 20.000 parole. Da lì si è partiti per un censimento dei fonemi e trifonemi della parlata italiana. La tesi con cui A. Zampolli ne presentò le conclusioni, fece molto chiasso, poiché si conobbero finalmente i trifonemi più frequenti, quelli cioè che concorrono a formare il

maggior numero di parole. Su questi si concentrerà d'ora in poi la rieducazione dei sordomuti, evitando a loro i dispiaceri che abbiamo avuto noi, quando da ragazzi ci hanno rimpinzati delle eccezioni francesi (ve le ricordate? hibou, genou, caillou... émail, épouvantail...) con il risultato che oggi noi possediamo correttamente parole che non usiamo mai e sbagliamo in quelle più comuni.

La proporzione dell'uso dei sostantivi, verbi, aggettivi, preposizioni, ecc. oscilla attorno a cardini fissi, variati però dall'età, sesso, temperamento, ecc. Un censimento di queste percentuali, esteso ai discorsi e componimenti di migliaia di alunni di ambienti diversi – estensione che solo l'automazione rende possibile – permetterebbe di individuare curve di normalità, le quali servirebbero di ulteriore sussidio diagnostico della psiche dell'uomo nell'età in cui è più plastico all'influsso educativo. Questa estate i quotidiani hanno dato risonanza mondiale alle conclusioni del censimento della metrica dell'Iliade compiuto a New York da James Mc Donough. L'Iliade ne è emersa di un solo autore. Avevo dato io, anni fa, a questo giovane studioso americano il primo avvio. Egli cominciò allora a perforare su schede la sola quantità delle sillabe di tutti i versi. Quando tutta l'Iliade fu così trascritta, un calcolatore mise in luce i ritmi e le proporzioni d'impiego dei vari metri. A farlo a mano, a parte il tempo che avrebbe richiesto, non ci sarebbe stata altra possibilità di controllo che quello di rifare tutto da capo alla stessa maniera. Ma oggi se voi non voleste credere alle conclusioni, potreste in pochi minuti ricontrollare tutti i calcoli ripartendo dalle schede iniziali.

La cronologia delle opere platoniche è stata a suo tempo ricostruita, e oggi resta fuori di discussione, appunto con la statistica degli stilemi, pur condotta senza sussidio di macchine automatiche.

Con analogo procedimento si potrebbe affrontare la controversia sull'autenticità di certi scritti, per esempio Shakespeare o Marlowe: appunto e sempre perché nello stile di chiunque esistono situazioni che sono sue caratteristiche personali e permanenti, non meno delle sue impronte digitali. Il che è del tutto ovvio, se si riflette che qualunque cosa noi esprimiamo, è sempre con noi stessi che la esprimiamo. Esistono studi sulla acuità degli accenti tonici. Parole con accento tonico sulla *i* e sulla *e* conseguono sentimenti alti e acuti; quelle con accento tonico sulla *o* o sulla *u*,

esprimono sentimenti deprimenti; quelle con l'accento sulla *a*, sentimenti neutri. Orbene si presero un brano dei Promessi Sposi e la sua traduzione francese. Il flusso dell'acuità dei rispettivi accenti tonici fu riportato con curve su carta millimetrata. Su altra carta millimetrata fu codificato in curve il succedersi dei vari livelli emotivi espressi dalle parole del testo. Ne risultò che l'andamento dell'acuità degli accenti nello scritto originale combacia con la curva descritta dal sentimento. Non così nella traduzione francese, ove il ritmo fonetico degli attenti né era sgorgato inconsapevolmente dall'ispirazione interiore, né era stato tenuto presente come un elemento da «tradurre».

IV.

16. Al tempo di Gutenberg, accanto ai manoscritti che rimasero sotto forma di quaderni e registri, si è collocato il libro stampato. Oggi accanto a quelli e a questo, che resteranno, si colloca il «libro magnetico». E per il deposito delle conoscenze umane ciò rappresenta un vero e proprio cambiamento di dimensione. Ma non è solo quantitativo né solo di velocità. È anche qualitativo. Se infatti è vero che il linguaggio dei calcolatori elettronici segnerà con tutta probabilità la cessazione dei tentativi di lingue universali artificiali, è anche vero che l'interpretazione induttiva del fenomeno linguistico mediante le formule della probabilità (evoluzione del linguaggio verso nuove specificazioni e insieme sua involuzione o entropia verso graduali perdite di semanticità: anche qui leggi di mescolanza di vita e morte) questa induzione, dico, nella misura in cui la rende possibile l'automazione, promette di far ricominciare il ciclo della consapevolezza linguistica e grammaticale con maggiori profondità, sistematicità e documentazione.

17. Uno dei segni che oggi ci si trova anche qui ad una svolta, è il fatto che ci sono nel mondo circa 200 centri occupati a questo rovesciamento della torre di Babele, a salvaguardare cioè nel linguaggio la fisiologia unificatrice di comunicazione e arginarne la patologia di barriera e separazione. Tra essi, una dozzina oramai dopo quello di Gallarate, si occupano

del settore dell'analisi lessicale pura. Gli altri la applicano alle tecniche dell'informazione e della traduzione meccanica. È infatti triangolare lo sviluppo dell'automazione linguistica. Altro segno sta nel fatto che istituzioni come ministeri del commercio e della difesa ed altre – USA, URSS, Nato, Euratom, ecc. – lo finanziano da qualche anno a questa parte. In Francia, Olanda, Israele, Cecoslovacchia sono in corso progetti di gigantesche elaborazioni elettroniche – si è arrivati a parlare di 120 milioni di schede per il Trésor de la langue française – onde avere i materiali per la compilazione dei dizionari storici della lingua nazionale. Del resto anche lo schedario dell'Index Thomisticus, in corso di produzione a Gallarate, potrebbe essere definito il primo Thesaurus della lingua scientifica del nostro Medio Evo.

18. Domenico De Domenichi, veneziano «de ordine plebejo», divenne vicario di Papa Sisto IV. Nella prefazione a un incunabulo stampato a Venezia nel 1480, egli così commentò la recentissima, allora, invenzione della stampa: «Placuit autem clementissimo Deo his nostris temporibus novam artem docere homines». Continua riportando la mirabolante notizia che tre uomini in soli tre mesi di lavoro sono riusciti a stampare ben 300 copie del volume: «ad quae tota eorum vita haud quaquam sufficeret si cum digitis et cum calamo aut penna scribenda forent» e poi conclude «si quid in me est auctoritatis etiam admoneo: ne tanta Dei beneficentia abutantur». Che cosa dovremmo dire oggi?

§

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

El análisis lingüístico en la evolución mundial de los medios de comunicación

Padre Roberto Busa s.j.

(introducción de Stefano Bazzaco)

(traducción de Soledad Castaño Santos)

Abstract

El artículo «L'analisi linguistica nell'evoluzione mondiale dei mezzi di comunicazione» de Padre Roberto Busa, pionero de las Humanidades Digitales, fue publicado en origen en el volumen *Almanacco Bompiani*: *Le applicazioni dei calcolatori elettronici alle scienze morali e alla letteratura* de 1962. A pesar de que el texto fue escrito en una fase preliminar de tales estudios, en que la aplicación de la informática al análisis de las lenguas y literaturas parecía todavía un sueño inalcanzable, la lúcida mirada del jesuita italiano constituye a nuestra manera de ver aún hoy en día un punto de interés para valorar la evolución y los retos futuros en el campo de las Humanidades Digitales. Presentamos aquí la traducción española del artículo con el fin de rescatar las palabras de Padre Busa y ponerlas nuevamente al alcance de un amplio número de lectores.
Palabras clave: Humanidades Digitales; Padre Roberto Busa; *Index Thomisticus*; procesamiento del lenguaje natural; automatización

The article «L'analisi linguistica nell'evoluzione mondiale dei mezzi di comunicazione» by Fr. Roberto Busa, a pioneer of the Digital Humanities, was originally published in the volume *Almanacco Bompiani*: *Le applicazioni dei calcolatori elettronici alle scienze morali e alla letteratura* of 1962. Despite the fact that the text was written in a preliminary phase of such studies, in which the application of informatics to the analysis of languages and literatures still seemed an unattainable dream, the lucid gaze of the Italian Jesuit in our opinion constitutes even today an observation point of interest to assess the evolution and future challenges in the Digital Humanities field. We present here the Spanish translation of the article with the intention of rescuing the words of Fr. Busa and making them available to a broader number of readers.
Keywords: Digital Humanities; Father Roberto Busa; *Index Thomisticus*; Natural Language Processing; automatization

Introducción: las Humanidades Digitales vistas desde la lente de Padre Roberto Busa

El jesuita Padre Roberto Busa, considerado el fundador de las Humanidades Digitales, redactó en 1962 el texto del que aquí proponemos la traducción. El artículo se titula en original «L'analisi linguistica nell'evoluzione mondiale dei mezzi di comunicazione» y fue publicado por primera vez en el *Almanacco Bompiani (1962). Le applicazioni dei calcolatori elettronici alle scienze morali e alla letteratura*, un volumen pionero dedicado a trazar un espacio de convivencia entre los estudios humanísticos y computacionales.

El concepto central que desarrolla Busa en estas páginas concierne el ingreso de la automatización en los ámbitos de las manifestaciones dialógicas humanas. Al respecto, el jesuita registra las nuevas vías que puede abrir esta innovación, señalando los pasos que ya se estaban dando en los campos del procesamiento del lenguaje natural y de la traducción no supervisada. El núcleo de su argumentación es por lo tanto la idea de que los ordenadores, imponiendo la «simbolización del conocimiento», suponen también la formalización del discurso crítico.

En esa época, Padre Roberto Busa estaba trabajando en la indexación de la *opera omnia* de San Tomás de Aquino por medio de fichas perforadas. Este proyecto lo ocupaba ya desde 1949, cuando supo vislumbrar los posibles avances que habría podido traer el tratamiento automatizado del lenguaje gracias al empleo de herramientas informáticas. La intuición, que se fundamentaba en el uso de la tecnología para el análisis del discurso no estructurado, tuvo repercusiones en distintas áreas del conocimiento, desde la lingüística computacional hasta el estudio de los documentos escritos, impulsando en las décadas posteriores la aplicación de técnicas estadísticas y de *machine learning* para la investigación literaria. Y si no sabemos hasta que punto el jesuita pudo tener conciencia del impacto que sus ideas habrían tenido en la mismísima manera de vivir y comunicar por parte de los seres humanos, lo cierto es que en la actualidad parece impensable imaginar nuestra sociedad, y con ella los espacios académicos, sin la presencia de medios informáticos que configuren, fomenten y sostengan nuestras modalidades de expresión y difusión del conocimiento

(Fiormonte, 2017, 16). En los años en que Alan Turing teorizaba el método formal y la noción de algoritmo y Vannevar Bush asentaba las bases del hipertexto[1], el lanzamiento del proyecto del *Index Thomisticus* por parte de Padre Busa constituyó un verdadero hito para las disciplinas que indagan la textualidad y sus múltiples manifestaciones, porque aseguraba que era posible formalizar, medir y clasificar el lenguaje a través de un ordenador.

Para determinar el impacto de tal intuición, sobre todo en relación con los estudios filológicos y literarios, hay que volver a la época en que el propósito de Busa se presentaba aún como un esbozo, es decir los años finales de la década de los 40, cuando la tecnología estaba ensayando el paso de la forma electro-mécanica a la computación digital *tout court* (Rockwell y Passarotti, 2019, 21). En ese momento liminal, el ordenador todavía no existía en la forma que conocemos hoy en día, sino que se presentaba como una enorme máquinaria de calculación, con la que se podía comunicar por medio de dispositivos mecánicos como las fichas perforadas (en inglés *punched cards*). Tales fichas, que aparecían como reducidas cédulas en papel rígido con una pequeña brecha en la parte superior derecha y unos agujeritos cuadrados en correspondencia de las palabras-token analizadas, eran las herramientas ideales para ingresar y extraer datos a través del ordenador de forma semi-automatizada. Su producción requiría un ingente trabajo manual y representaban una sólida manera de transmitar informaciones entre la máquina y los seres humanos, implicando la integración de dos distintos niveles informativos aptos para la comprensión y manipulación por parte de ambos.

Sin embargo, si la idea de emplear fichas perforadas para el manejo automático de los textos fue presentimiento de Busa, los medios para cumplir con ese sueño visionario fueron proveídos por la IBM (Passarotti, 2018). Thomas J. Watson, director de la empresa por esa época, tras la presentación por parte del jesuita de un boceto del proyecto que habría permitido indexar por medio del ordenador las concordancias en la obra completa de San Tomás, fue perspicaz en aceptar la propuesta. Por un lado Busa ponía las bases de la codificación de caracteres en formato

[1] Al respecto se vean Lucía Megías (2012, 43 y ss.) y Tomasi (2022, 20-21).

electrónico, yendo más allá de las diferencias lingüísticas; por otro Watson volcaba su compañía a la experimentación en ese campo, lo cual le habría asegurado notables éxitos y ganancias con la eclosión del *World Wide Web*.

La financiación concedida por la IBM a Busa en 1949, pactada durante un célebre encuentro en Nueva York, fue la primera de una larga serie que duró más de 30 años y que finalizó con la publicación en 1980 de los 56 volúmenes del *Index Thomisticus* (Busa, 1980). En ese largo lapso temporal, en principio a través de fichas perforadas y en un segundo momento gracias a cintas magnéticas oportunamente concebidas para la clasificación de palabras, Busa sentaba las bases de la lingüística computacional como disciplina científica y área de investigación. Sin embargo, su apuesta fue significativa no solamente desde el punto de vista pragmático (agilizando a nivel cuantitativo el procesamiento de los textos), sino en lo que supuso para el horizonte hermenéutico de las disciplinas humanísticas, es decir instituyendo procedimientos y espacios teóricos que consintieran la integración cualitativa de los medios informáticos a las prácticas discursivas de las Humanidades (Testori, 2017).

Las reflexiones de Busa sobre el fenómeno lingüístico, en otras palabras, guiaron los humanistas hacia una comprensión aún más profunda de los mecanismos del lenguaje y de sus niveles informativos, porque los ordenadores requerían un alto grado de formalización de algo que por su naturaleza se escapaba del análisis cuantitativo, es decir la semántica del discurso. Apuntando a nuevas vías para la interpretación de los artefactos textuales, Busa hacía hincapié en las propiedades performativas del lenguaje y arrojaba nueva luz sobre los patrones mentales que guían la producción textual, insistiendo en la posibilidad de transformar enormes moles de datos en conocimiento.

Si tomamos el *Index Thomisticus*, los números del proyecto son impresionantes: 11.000.000 de fichas –una para cada palabra del corpus procesada–, más de 20.000.000 líneas de texto, 70.000 páginas, 56 volúmenes. Estos datos fueron recogidos por más de 70 colaboradores que ocupaban la antigua fábrica textil de Gallarate en provincia de Milán, los cuales gestionaban una cantidad interminable de fichas perforadas de forma manual, señalando concordancias y lemas, tanto que el proyecto se ha descrito recientemente como una de las iniciativas de las Humanidades

Digitales más grande de todos los tiempos, precursora de la lectura distante y del estudio de la textualidad digital como *Big Data* (Rockwell y Passarotti, 2019; Tomasi, 2022, 20).

En los años de máximo esplendor de este proyecto, es decir en la década de los 60, cuando el laboratorio informático de Gallarate estaba en su punto más álgido, Busa redactó el artículo que aquí traducimos. En él introduce unos conceptos de gran interés para las Humanidades Digitales actuales: la transmisión al silicio de los contenidos de los documentos en papel, su sistematización y descripción por medio de metadatos, y su clasificación y explotación con técnicas de *information retrieval*. A la base de todas estas disposiciones subsiste la posibilidad de detectar unas constantes, unos patrones en el lenguaje que cada ser humano emplea para comunicar un determinado asunto. Presenta pues unos temas directamente relacionados con la individuación en el flujo del discurso científico y literario de unos *tics* lingüísticos que consientan ejecutar un estudio estadístico sobre las palabras con las que nos expresamos.

El análisis del discurso por medio de la automatización permitiría, en opinión de Busa, registrar las tensiones del discurso. Y esto porque demostrando ciertas costumbres de la lengua y cuantificando el empleo constante de ciertos términos se pueden registrar los elementos de continuidad entre distintos textos, detectar diferencias entre autores, disponer en un gráfico las palabras y atribuirles cierto grado de afinidad con las pasiones humanas. En suma, se pueden «operazionalizar» (Moretti, 2013) los contenidos informativos para tansformarlos en algo más sintético y conmensurable. Padre Busa, en otras palabras, repasando las posibles aplicaciones de la automatización al lenguaje natural, está abriendo nuevos caminos posibles para la investigación humanística, apuntando a los procedimientos de la estilometría, de la lingüística forense, de la atribución autorial y de la *sentiment analysis* soportadas por herramientas computacionales.

Concluye pues estableciendo un paralelo entre la automatización del análisis del lenguaje y la imprenta, lo cual le permite determinar lo que va a ser en el futuro el libro electrónico, es decir un pasaje de *medium* que implica no solamente un salto cuantitativo (con el rápido incremento y almacenamiento de datos), sino también la fijación de un nuevo modelo

cualitativo basado en la interpretación inductiva del lenguaje y en su atenta reproducción en el nuevo contexto. Volcar los textos al ordenador con el respeto de su tradición documental quiere decir actuar de forma sistemática, basarse en la observación empírica y la replicabilidad de las operaciones, demostrar una correcta atención a la calidad final de los datos que ingresamos, que tienen que ser fiables y certificados (Rockwell y Passarotti, 2019, 26-27).

Dispone y ordena lógicamente estos temas una voz espiritual y a la vez extremadamente lúcida en su argumentación, esa voz que en opinión de Passarotti y Nyhan se fundamenta en la necesidad de transmitir una entera visión del mundo frente al *digital turn* (2019, 3), que nos sigue hablando en la actualidad y a la cual pensamos que se debería oportunamente volver.

«L'analisi linguistica nell'evoluzione mondiale dei mezzi di comunicazione», en *Almanacco Bompiani. Le applicazioni dei calcolatori elettronici alle scienze morali e alla letteratura* (1962), pp. 103-108, 117.

I.

1. El fenómeno lingüístico es más grande que nosotros: uno de los ingredientes de esta extraña fórmula de masa que es cada uno de nosotros. Los valores, de hecho, de los que somos un dispositivo tan frágil y maravilloso son, en sí mismos, mucho más amplios y mucho más grandes que nosotros mismos. Las manos, por ejemplo, nos sirven para muchas cosas sencillas o complicadas, pero están, como los camareros, por así decirlo, siempre a nuestras espaldas, las utilizamos sin prestar mucha atención. Pero si las pusiéramos delante de nuestros ojos y las escrutáramos y pensáramos un poco también en ellas, nos encontraríamos frente a todo un mundo por descubrir. Otro misterio es, por ejemplo, nuestra capacidad de gusto estético. ¿En virtud de qué «programa», cargado en ese robot que somos, sentimos la necesidad tan imperiosa, por ejemplo, de la simetría, la repugnancia a cualquier disonancia de color, de línea, de sonido? Pero nuestras verdaderas manos son nuestras capacidades expresivas: con los gestos, con el rostro, con las artes, con las palabras nosotros dañamos y medicamos, derrocamos y elevamos, mejoramos o estropeamos todo a nuestro alrededor. También estas son mundos por explorar en nuestro interior.

2. En nuestro hablar, el que tenemos en la boca y del que sabemos tan poco, hay tres estratos: lo que está presente en el campo de la conciencia, lo que es subconsciente y lo que es de ninguna manera inconsciente. Y en la misma zona de nuestro lenguaje que está iluminada por nuestra percepción y atención, una parte, pero no toda, es pasible de control en el sentido inglés de la palabra, es decir, una parte puede ser gobernada y, por tanto, también educada por nosotros. Es al menos teóricamente posible que un milanés decida acostumbrarse a decir «v*ada*» en lugar del incorrecto «*vadi*», ¡al que quién sabe por qué atávica herencia

es tan aficionado! Hay otros sectores que escapan a un control organizativo, pero no del todo a la detección sistemática: no llegamos a cambiarlos, pero aún así nos damos cuenta, aunque sea en cierta medida. Otras zonas, por último, solo obedecen al subconsciente o incluso al inconsciente. Por ejemplo, solo con mucha sutileza llegaremos a darnos cuenta que si nosotros preferimos ciertas palabras a otras, lo hacemos porque nos rige una aspiración subconsciente de causar una supuesta buena impresión, y esto como consecuencia del mayor valor que nosotros adjudicamos a ciertas palabras, así como hacen las señoras con ciertas palabras del tipo «*genare*» o «*flattare*», mientras que otras escogerán las palabras en función de su propia capacidad definitoria y otras, de nuevo, en función de su estética, fonética o semántica o de su correlación y ritmo. Pero en un nivel más profundo, las estructuras gramaticales y sintácticas parecen brotar de las raíces inconscientes con las que la humanidad succiona su evolución vital de ese universo en que se agita, ¡por tan poco tiempo! Como una ameba en su caldo de cultivo. Los fundamentos del lenguaje se encuentran entre las zonas del comportamiento humano, que son inaccesibles a la educación y al autocontrol, porque están programados y comandados exclusivamente por lo que está en las raíces de nuestra fisiología y la desagradable e inevitable mezcla de fisiología y patología.

3. Así que no todo en nuestro hablar se deja conocer: de lo poco o de lo mucho que se permite conocer, no todo está influenciado por nuestra agresiva ambición de hacer también con nosotros y de nosotros «lo que queramos». No obstante, ¿por qué no podríamos dejar también sin atender ese parterre que podríamos, si quisiéramos, desenterrar? ¿Por qué querríamos tomar unos cuantos hilos de agua del río de nuestras palabras y enviarlos por conductos forzados? Exactamente por la misma razón por la que intentamos controlar el agua. El hablar es, de hecho, el principal potencial energético del que el hombre dispone y, por tanto, debe proporcionarse de forma económica. Las ideas son fuerza, solo cuando se pueden decir y escribir. Tampoco tienen ningún otro medio para llegar al individuo que las tiene.

4. Por ello, Aristóteles echó un buen vistazo a su interior y descubrió la metafísica en los pliegues del lenguaje. Y para ese enciclopédico y positivo detector de hechos que ha demostrado ser, esta ha sido una de los más sensacionales: sentirse catapultado desde la pista que había recorrido, palmo a palmo, por la investigación positiva a un plano superior. Hasta el plácido y buen Santo Tomás de Aquino lo contempló, admirando con la nariz en el aire el puente, con el que un pagano, había logrado penetrar en el cielo desde esta tierra. Pero Filón, conocedor del Antiguo Testamento y la Teología Cristiana, partiendo del examen de la «palabra», había penetrado aún más los cielos, superando con creces la gran carrera de Aristóteles y Platón. Entre las palabras habían vislumbrado el destello del Logos, Verbum. Y no ha habido ningún idealista absoluto que haya logrado atreverse tanto como el filósofo cristiano a enuclear el valor de la expresión del «*verbum mentis*» y la consiguiente reverberación del mutuo amor, dentro de ese pensamiento absoluto que es el fuego de la consistencia, la vida y la fantasía: un director que es, al mismo tiempo, un arco voltaico que proyecta sobre la pantalla oscura de la nada esa sucesión de imágenes que somos nosotros, el mundo y la historia.

5. En la vida social, la gramática y el análisis lógico han educado durante muchos siglos ese algo indefinible que nosotros llamamos humanidad y humanismo, ese atisbo de gusto por la belleza, de sentido de la armonía, de apreciación de los valores formales, por el que incluso en el Politécnico subsiste la diferencia entre los que vienen de un liceo clásico y los que vienen de otras escuelas. La retórica había educado en el arte de la expresión. Bien decía el viejo Aristóteles que «*signum scientis est posse docere*», nuestros conocimientos están maduros cuando alcanzamos a transmitirlos. Todos hemos experimentado que cuando el profesor tarda dos horas en hacernos entender algo, es porque aún no lo posee perfectamente, y que para ser capaz de poseerlo perfectamente no debería haber hecho nada más que prepararse para decirlo primero. ¡Cuántas veces hemos visto en la vida la inagotable sabiduría del dicho de que entre el tener razón y el poder tenerla hay una enorme distancia! Hoy en día, a menudo nos preocupamos solo de hacer engullir nociones, como si el hombre fuera un almacén general, o como si no hubiera en él más que

memoria. En cambio, el hombre es, sobre todo, y al menos para su destino, una capacidad organizativa e inventiva, y no necesitaría educarlo como una mochila que se llena según la lista de lo que se necesitará para el campamento, sino que debe ser refinado en sus dispositivos, lubricado, rodado como una máquina-herramienta capaz de trabajar durante mucho tiempo en cualquier material. ¿Me equivoco al pensar que pagaríamos el precio de saber la mitad de lo que sabemos, si pudiéramos decir mejor lo poco que sabemos? Así que el cuidado de nuestros medios expresivos solía existir mucho más en el pasado que en la actualidad. Se enseñó a organizarse internamente para adecuar la combinación de palabras al propósito deseado: a pensar en cómo hablar antes de hacerlo. Pero el mismo Manzoni decía que esta única cosa, «pensar antes de hablar» es tan difícil en sí mismo que también a nosotros se nos debe disculpar un poco por esas tantas veces que nos abandonamos a hablar así como lo hacemos.

6. Poco a poco, las leyes universales del envejecimiento, las cuales afectan a las instituciones tanto como al hombre y a la naturaleza, han desgastado la mordacidad del análisis lógico y de la retórica. El poder de la decadencia ha ejercido su tiranía hasta tal punto que hoy en día solo se va a la escuela de recitación para el teatro, pero no para prepararnos todos para representar nuestro guion en las comedias y las tragedias de la vida. Y si ha escrito que la divinidad del Evangelio se ha demostrado, al menos, su pervivencia en las explicaciones dominicales, es decir, si a vosotros, burgueses, os parece que los curas somos frecuentemente tan descuidados en nuestras predicaciones, es porque nosotros, como vosotros, somos hijos de nuestro tiempo. Y el atardecer se presenta violáceo, también porque en Italia, se está tratando el latín como un viejo abuelo al cual se le desean otros cien años de vida, mientras que el subconsciente registra, piensa que dentro de pronto nos quedaremos sin él, no se siente en el fondo un horror por la verdad infinita.

7. En este punto, el monstruo de la noche, el tecnicismo triunfante, ha intervenido con su última criatura: la automatización. Algunos se estremecieron al pensar que era una excavadora cruda y dura que rugía, aplastando y arrancando las flores. Entre ellos, víctima delicada y gentil, el

humanismo. El mañana ya está aquí. El futuro ya ha comenzado: una colada de lava inunda y quema las verdes laderas de la montaña. En la torre de comando del monstruo, encapsulada entre manómetros, relojes, luces espías y diales, hay algunos hombres. Quizás al principio ni siquiera se dieron cuenta de los agudos gemidos y lamentos elegíacos de los «humanistas». Se contentan con… trabajar. Afirman que prestan un servicio de utilidad pública, porque creen que sin ellos su industria y el comercio no podrían satisfacer las necesidades humanas. Pero entonces– no han pasado todavía diez años– los hombres de la automatización han empezado a sacar la cabeza de la cabina de la torre electrónica, para preguntar a los filólogos y gramáticos, ocupados en los campos recogiendo la flor de las flores, cuestiones de esta naturaleza: ¿Si puede saberse, cuántos verbos transitivos activos hay en ruso, y cuántos intransitivos activos? ¿Cuántos hay en inglés? ¿Cuál es el mayor número de letras iniciales y finales en las que coincide el mayor número de palabras? ¿Qué palabras o situaciones lingüísticas se encuentran entre un radio de n palabras, solo cuando y siempre cuando «*faccia*» significa cara, cuáles otras solo y siempre cuando «*faccia*» proviene del verbo *fare*? Y de nuevo: ¿Si puede saberse, podrías agrupar todas las palabras del vocabulario según las distintas categorías morfológicas y gramaticales? Dígame todas las palabras que se pueden omitir, cuando, para acortar un texto sin perjuicio de su expresividad. ¿Pero sabe decirme en cada caso la configuración característica de ciertas categorías semánticas que no son ni morfológicas ni sintácticas ni estructurales? En otras palabras, ha ocurrido un hecho clamoroso, la máquina nos ha hecho conscientes de que ningún humanista posee un lenguaje propio de poder dar respuesta a tales preguntas. La máquina, criada del comercio banal y de la industria burda, ha documentado que todavía hay muy poco humanismo serio y sistemático. Los hechos económicos exigen, hoy en día, un incremento cualitativo de las ciencias gramaticales y léxicas: como una de las necesidades de su desarrollo vital. Pero también ofrecen una posibilidad. No es una pequeña venganza ni una pequeña satisfacción.

II.

8. El centro de Gallarate es aún hoy el centro que el mundo ha transportado la mayor cantidad de palabras en fichas del mundo: ya hay casi cuatro millones y en continuo aumento. Se trata de 7 lenguas, (Aristóteles, Antiguos Italianos, Dante, Kant, Goethe, Textos hebreos del Mar Muerto, Fabbri, etc.) en tres alfabetos, latino, griego y hebreo. Pero cuando en el 1946 comencé a pensar seriamente en los índices verbales de los trece millones de palabras de Santo Tomás de Aquino, y cuando más tarde, en 1939 inicié los primeros experimentos con el IBM y , de nuevo, cuando en 1951 publiqué los primeros resultados, no solo fui el único y el primero en el mundo que se aventuró a inscribir la lexicología en el hipogrifo, sino que además desconocía el momento histórico en el que esto me sucedía. El haber tenido primero una idea no es un mérito, sino casualidad. Si no me hubiese llegado a mí, seguro que la idea le habría llegado a alguien más. Y quizás un día resulte que antes que a mí se le habría ocurrido a cualquier otra persona, a la que nadie había prestado atención en su momento. Y si se pudiese hablar de mérito, eso consistiría en la larga paciencia requerida para resolver, paso a paso, todas las dificultades e imprevistos que se encuentran al transformar una idea en una metodología: madura y práctica, aplicable por así decirlo, a la producción en serie. De la célebre frase «*genius is one per cent inspiración, ninenty-nine per cent perspiration*» (el genio se compone de uno por ciento de inspiración, un noventa y nueve por ciento de sudor). La única palabra cierta que yo no verifico es la primera. Pero, ¿quién hubiera imaginado entonces que las máquinas de tarjetas se considerarían hoy en día antiguas, y que veríamos la evolución, o más bien la metamorfosis de las computadoras electrónicas desde las memorias de superficies recubiertas de óxido de hierro, a las de entramado de anillos de ferrita, y, finalmente, a las criogénicas (películas muy finas superpuestas como un libro, utilizables a temperaturas cercanas al cero absoluto)? Desde luego, no imaginaba que el «*strech*» construido para las investigaciones nucleares, tendría una memoria de algo menos de dos millones de posiciones, en la que toda la Enciclopedia Treccani podría nadar como un niño en la cuna, y otra memoria de un millón y medio de posiciones que tienen una

velocidad de conmutación de unas centésimas de milinésima de segundo. Pero, sobre todo, ignoraba que me estaba incluyendo en la sucesión de pasos, a través de los cuales la automatización de la contabilidad provocó la evolución mundial de los medios de comunicación.

9. Puedo condensar el movimiento que tomó la aceleración de una avalancha [en la evolución de los medios de comunicación] después de 1945 en cuatro fases. Primera etapa– El desarrollo de las comunicaciones y de las técnicas de organización permitió el crecimiento de las empresas hasta abarcar todo el mundo. Igualmente rápido fue el aumento de la influencia recíproca de los mercados y entre la política y el mercado. Por ello, se ha vuelto indispensable que el gestor pueda registrar un gran número de datos, para elaborar rápidamente resúmenes: a tiempo para controlar y si se desea modificar el curso de grandes masas de pequeños y extensos fenómenos periféricos. Los ordenadores respondieron a esta necesidad aportando a la vida económica la automatización de la contabilidad industrial y comercial. Pueden realizar hasta un millón de multiplicaciones y divisiones por segundo. Pueden imprimir los resultados propios a la velocidad de 60.000 líneas por hora para el alfabeto y 300.000 solo para los números.

10. Segunda etapa– La industria, cuyo desarrollo se ve exacerbado por las exigencias de la «defensa», y la intensificación paralela de las relaciones entre la producción industrial y la investigación científica, han impuesto la automatización del cálculo científico. El Euratom, por ejemplo, se ha visto obligado a comprar el ordenador IBM 7090 para su centro de Ispra, que cuesta aproximadamente tres millones de dólares, es decir, casi dos mil millones de liras.

11. Tercera etapa– Las actividades de producción, intercambio y defensa requieren la automatización para «*l'information retrieval*» que yo traduciría como la disponibilidad oportuna de conocimientos útiles. La cantidad de publicaciones científicas, ya enorme, está en continuo aumento. Estados Unidos tiene ahora una media de 40.000 patentes al año. Por otra parte, la aceleración de la evolución científica es tal que las

publicaciones en física nuclear después de dos años ya solo sirven para la historia de la física. Pero en lo que respecta a las técnicas informáticas, la actualidad útil de las noticias es probablemente de poco más de medio año. Ahora imaginen que una industria armamentística necesita conocer el comportamiento de ciertos materiales en determinadas nuevas situaciones. ¿Cuánto tiempo tardarán en examinar todas las ciencias relevantes para encontrar lo que le conviene? No necesitarán los índices analíticos, porque, por definición, encontrarán algo que no es comúnmente conocido, ni tampoco le bastarán las indicaciones bibliográficas, porque estas solo contienen los títulos, mientras que usted, por la razón ya expuesta, necesita hurgar en el contenido mismo de lo impreso. Necesitará alguna vez los resúmenes. Pero intenta que se lean todos y ya me dirá si es demasiado tarde cuando termine. ¿Cómo hacer entonces para seguir el ritmo de todas las publicaciones de todo el mundo casi simultáneamente a su aparición? ¡Me parece que el DDT se descubrió por primera vez dos o tres veces seguidas! Por lo tanto, es necesario condensar un máximo de información científica para poder encontrar en un mínimo de tiempo todo lo que interesa en la búsqueda de lo nuevo. El objetivo de la automatización es conseguirlo.

12. Los ámbitos en los que se ha canalizado son: los nuevos tipos de simbolización del conocimiento, es decir, los alfabetos con impresiones magnéticas, cómo transcribir y copiar con estos nuevos alfabetos, que solo la máquina puede leer, el contenido de lo que se imprime con los alfabetos de tinta sobre papel (se esfuerzan para hacerlo mediante la fotolectura, fonografía, etc.); cómo condensarlo (resumirlo, reducirlo al estilo telegráfico, acortar las palabras); cómo clasificarlo; cómo investigarlo. Un capítulo de este esfuerzo se representa en la traducción automática. No me refiero a la ciencia ficción de traducir un texto literario o filosófico a máquina, sino a la técnica de traducir a máquina publicaciones contemporáneas sobre el mismo tema, en las ciencias unificadas tal y como son hoy, por tanto, concebidas y expresadas de la misma manera y con un vocabulario, cuyas únicas diferencias son las de las dos lenguas. Esta técnica añade a los problemas anteriores el de cómo, a partir de situaciones o factores lingüísticos característicos del contexto de una palabra, se puede

identificar automáticamente su función gramatical y lógica y, en los casos de polisemia, su significado en ese preciso lugar; y el de cómo la máquina puede traducir la sintaxis de una lengua a la de otra lengua. La Universidad de Georgetown, en Washington DC, ha abierto hace un año un centro en Fráncfort del Meno donde treinta personas capturan continuamente publicaciones científicas rusas, que son después traducidas al inglés por el ordenador 704.

13. Cuarta etapa– La automatización del tratamiento de la información requiere la automatización en la elaboración de índices, concordancias y todo tipo de estadísticas posibles de hechos lingüísticos. En el Euratom en Ispra visite el grupo Tetis. Id a Washington al Instituto de Lenguas y Lingüísticas de Georgetown. Se darán cuenta de que entre los investigadores de las técnicas de tratamiento de la información se está desarrollando una lexicología y una lingüística más sistemáticas, más exhaustivas, más ampliamente útiles, y me atrevo a decir, más humanísticas que las tradicionales hasta hora. Y dentro de poco, desde los huertos del humanismo, las voces de tenor de los filólogos orquestarán los méritos de la automatización, comentados baritonalmente por los matemáticos.

III.

14. Pero entonces, incluso dentro de esa máxima expresión de nuestra libertad, personalidad, capricho, que es nuestro hablar, hay formulas matemáticas. Y es cierto: no se puede hablar «como uno quiera», sin obedecer ninguna ley. Si os abandonaseis a la voluntad de sacar, más allá de ciertos límites, del gran mar de combinaciones que son aritméticamente posibles entre los elementos de vuestro vocabulario, ciertas secuencias de palabras que son inusuales a este lado de esos límites, estad seguros que os encerrarían en algún lugar para someteos a la cura de sueño. Pero no solo en este sentido hay leyes en el habla. El número –quién sabe qué alegría sentiría el bueno de Pitágoras si estuviese vivo– ha aparecido como la estructura de soporte del lenguaje, al igual que las

proporciones de las medidas y las relaciones de las proporciones son el esqueleto de la forma y de la belleza. Y la estadística lingüística, que nuestro Davanzati utilizó hace siglos, es tanto más alentadora cuanto que el numero sigue reinando entre los fundamentos de las ideas y de la lógica, como lo demuestran la lógica simbólica y la álgebra de las preposiciones, al igual que pertenece a la sustancia del suelo y fuente del ser, como destaca la teología trinitaria católica. Dado que el lenguaje puede traducirse en términos que combinan una gran masa de pequeños elementos, ya que es un entretejido de repeticiones y frecuencias, su matemática no es solo una matemática determinista, sino aún más la matemática de la probabilidad y el azar, una matemática maravillosa que está más cerca del misterio de Dios, del espíritu y del arte. Juan Joaquín Becher, fallecido en 1682, polígrafo, implicado en las hazañas de la teoría flogística, mereció la gratitud imperecedera de su madre Alemania por haberle enseñado a extraer el alcohol de las patatas. Pues bien, un hombre con tan vastos y empíricos intereses podría ser llamado el precursor de la codificación numérica de las palabras.

En su *Character pro notitia linguarum universali*, Francofurti, 1661, proclamó que no hay una sola lengua que pueda ser entendida por todos, salvo que cada concepto se exprese con un numeral o un jeroglífico correspondiente. Esto es todo lo que hace falta –y aún falta mucho y se está trabajando intensamente en ello– para que cualquier ordenador, digital o analógico, pueda servir de traductor fiel y confidencial: ese ordenador que los alemanes llaman Hochgeschwindigkeittrottel: ¡un idiota de altísima velocidad!

15. Para lo que puede servir la estadística de los factores lingüísticos, ampliada tan largamente como lo permitan las increíbles posibilidades de la automatización, pueden ilustrarse los siguientes ejemplos.

En Gallarate, por encargo de los profesores Tagliavini y Croatto de la Universidad de Padua, se realizó la transcripción fonética de un texto de Fabbri de aproximadamente 20.000 palabras de forma automática. Este fue el punto de partida de un censo de los fonemas y los trifonemas del habla italiana. La tesis con la que A. Zampolli presentó las conclusiones, tuvo mucha repercusión, porque finalmente se conocieron los trifonemas

más frecuentes, es decir, los que se combinan para formar el mayor número de palabras. A partir de ahora, la reeducación de los sordomudos se concentrará en estos, para evitar las desgracias que tuvimos nosotros, cuando como de niños nos atiborraron de expresiones francesas (¿Las recordáis? *Hibou*, *genou*, *caillou*, *émail*, *épouvantail*…) con el resultado de que hoy nosotros poseemos palabras que nunca utilizamos correctamente y cometemos errores en las más comunes.

La proporción del uso de los sustantivos, verbos, adjetivos, preposiciones, etc. oscila en torno a unas bisagras fijas, qua varían, sin embargo, según la edad, el sexo, el temperamento, etc. Un censo de estos porcentajes, ampliado en los discursos y a la composición de miles de alumnos de diferentes entornos –extensión que solo la automatización hace posible– permitiría identificar curvas de normalidad, que servirían como ayuda para el diagnóstico de la psique del hombre en la edad en que es más susceptible al influjo educativo.

Este verano los periódicos dieron cobertura mundial a las conclusiones del inventario de la métrica de la *Ilíada* realizado en Nueva York por James Mc Donough. La *Ilíada* surgió como obra de un solo autor. Fui yo quien, hace años, le dio a este joven estudioso americano su primera oportunidad. Él comenzó entonces a marcar en fichas solo el número de sílabas de todos los versos. Cuando se transcribió así toda la *Ilíada*, un ordenador reveló los ritmos y las proporciones del uso de los distintos metros. Para hacerlo a mano, a parte del tiempo que habría llevado, no habría habido más posibilidad de control que volver a hacerlo de la misma manera. Pero hoy en día, si no queréis creer en las conclusiones, podréis volver a comprobar todos los cálculos en unos minutos desde las fichas iniciales.

La cronología de las obras platónicas ha sido reconstruida en el pasado, y hoy en día sigue fuera de discusión, precisamente mediante la estadística de los estilemas, aunque realizada sin ayuda de máquinas automáticas.

Un procedimiento similar podría utilizarse para abordar la controversia sobre la autenticidad de ciertos escritos, por ejemplo, Shakespeare o Marlowe: precisamente y siempre porque en el estilo de cualquiera hay situaciones que son sus características personales y

permanentes, no menos de sus huellas dactilares. Lo cual es bastante obvio, si se reflexiona que todo lo que expresamos, está siempre con nosotros mismos que lo expresamos. Existen estudios sobre la agudeza de los acentos tónicos. Las palabras con acento tónico en la *i* y la *e* dan lugar a sentimientos elevados y agudos; las que tienen acento tónico en la *o* o la *u*, expresan sentimientos depresivos; y las que tienen el acento en la *a*, sentimientos neutros. Tomamos un paisaje de *Los novios* y su traducción al francés. El flujo de la agudeza de los respectivos acentos tónicos se informó con curvas en papel milimetrado. La sucesión de los distintos niveles emocionales expresados por las palabras del texto se codificó en curvas en otro papel milimetrado. El resultado fue que el curso de la agudeza de los acentos en el texto original coincide con la curva descrita por el sentimiento. No es el caso de la traducción francesa, en la que el ritmo fonético de los acentos no fluyó inconscientemente de la inspiración interior ni se tuvo en cuenta como elemento a «traducir».

IV.

16. En la época de Gutenberg, junto a los manuscritos que permanecían en forma de cuadernos y registros, surgió el libro impreso. Hoy en día, junto a estos y a este, que permanecerá, se encuentra el «libro magnético». Y para el depósito de los conocimientos humanos, esto representa un verdadero cambio de dimensión. Pero no solo es cuantitativo, ni solo en términos de velocidad. También es cualitativo. Si bien es cierto que el lenguaje de las calculadoras electrónicas marcará con toda probabilidad el fin de los intentos de lenguas universales artificiales, también es cierto que la interpretación inductiva del fenómeno lingüístico a través de las fórmulas de la probabilidad (evolución del lenguaje hacia nuevas especificaciones y al mismo tiempo su involución o entropía hacia pérdidas graduales de semanticidad: aquí también leyes de mezcla de la vida y la muerte), esta inducción, digo, en la medida en la que la automatización lo hace posible, promete reiniciar el ciclo de la conciencia

lingüística y gramatical con mayor profundidad sistemática y documentación.

17. Uno de los signos que hoy se encuentra en un punto de inflexión es el hecho de que hay en el mundo aproximadamente 200 centros ocupados en este derrocamiento de la torre de Babel, esto es, en salvaguardar en el lenguaje la fisiología unificadora de la comunicación y en frenar la patología de las barreras y la separación. Entre ellos, una docena, después del de Gallarate, se ocupan en estos momentos del campo del análisis léxico puro. Los demás lo aplican a las técnicas de información y traducción automática. De hecho, el desarrollo de la automatización lingüística es triangular. Otra señal es que instituciones como los ministerios de comercio o de defensa y otros –Estados Unidos, URSS, Nato, Euratom, etc.– lo financian desde hace algunos años. En Francia, Holanda, Israel, Checoslovaquia, están en marcha proyectos de gigantescas elaboraciones electrónicas –se habla de 120 millones de fichas para el *Trésor de la langue Française*– con el fin de disponer de materiales para la compilación de diccionarios históricos de la lengua nacional. Por otra parte, incluso el expediente del *Index Thomisticus,* que se está elaborando en Gallarate, podría definirse como el primer Tesauro del lenguaje científico de nuestra Edad Media.

18. Domenico De Domenichi, un veneciano, «*de ordine plebeio*», se convirtió en el vicario del Papa Sixto IV. En el prefacio de un incunable estampado en Venecia en 1480, comenta a la muy reciente invención de la imprenta: «*Placuit autem clementissimo Deo his nostris temporibus novam artem docere homines*». A continuación, informa de la asombrosa noticia de que tres hombres, en solo tres meses de trabajo, consiguieron imprimir 300 ejemplares del volumen: «*ad quae nota eorum vita hand quaquam sufficeret si cum digitis et cum calamo aut penna scribenda forent*» y luego concluye: «*si quid in me auctoritatis etiam admoneo: ne tanta Dei beneficencia abutantur*» ¿Qué debemos decir hoy?

Bibliografía citada

Busa, Roberto, «The annals of humanities computing: The Index Thomisticus», *Computers and the Humanities*, 14/2 (1980), 83-90.

Fiormonte, Domenico, *Per una critica del testo digitale. Letteratura, filologia e rete*, Roma, Bulzoni, 2017.

Lucía Megías, José Manuel, *Elogio del texto digital: Claves para interpretar el nuevo paradigma*, Madrid, Fórcola, 2012.

Moretti, Franco, «Operationalizing: Or, the Function of Measurement in Literary Theory», *New Left Review*, II, 84 (2013), 103-19.

Passarotti, Marco, «Padre Busa, il gesuita che inventò l'ipertesto grazie ai computer IBM», *IBM thinkMagazine*, 2017 <https://www.ibm.com/easytools/runtime/hspx/prod/public/X0027/PortalX/page/pageTemplate?s=78c374df5c884363b46454a5ffefb5d9&c=6623351d59604a11b2c845760f87280f> (cons. 03/06/2022).

Passarotti, Marco y Julianne Nyhan, «Introduction, or Why Busa still matters», en *One origin of Digital Humanities. Fr Roberto Busa in his own words*, Springer, 2019, 1-17. DOI: <https://doi.org/10.1007/978-3-030-18313-4> (cons. 03/06/2022).

Rockwell, Geoffrey y Marco Passarotti, «The Index Thomisticus as a Digital Humanities Big Data Project», *Umanistica Digitale*, 5 (2019), 13-34.

Testori, Marinella, «Methods of quality, quality of methods. What does Roberto Busa have to communicate to digital humanists in the 21st century? From hermeneutics to performativity», *Digital Humanities Quartely*, 11/3 (2017) <http://www.digitalhumanities.org/dhq/vol/11/3/000329/000329.html> (cons. 03/06/2022).

Tomasi, Francesca, *Organizzare la conoscenza: Digital Humanities e Web semantico*, Editrice Bibliografica, Milano, 2022.

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

De editor analógico a editor digital

José Manuel Fradejas Rueda

(Universidad de Valladolid)*

Abstract

En el artículo se cuentan los recientes desarrollos del proyecto colaborativo 7PartidasDigital. En concreto, se señalan los principales rasgos del trabajo filológico tradicional y se apuntan los posibles avances que pueden ofrecer los medios informáticos a la hora de realizar una edición crítica digital. Entre los aspectos más destacados, insistimos en la transcripción automatizada de testimonios impresos y en el uso de sistemas de colación automática, dos recursos que de manera concreta agilizan y semplifican el trabajo filológico.
Palabras clave: Humanidades Digitales; *Siete Partidas*; Alfonso X el Sabio; filología digital

The present article is centered on the recent developments of the collaborative project 7PartidasDigital. Specifically, the main aspects of traditional philological work and the possible advances that computational tools can offer in the making of a digital critical edition are pointed out. Among the most outstanding aspects, we insist on the automated transcription of printed testimonies and the use of automatic collation systems, two resources that concretely speed up and simplify philological work.
Keywords: Digital Humanities; *Siete Partidas*; Alfonso X the Sage; digital philology

§

* Este trabajo forma parte de los resultados del proyecto *7PartidasDigital* (referencias FFI-2016-75014 y PID2020-112621GB-I00) cuyo objetivo es una edición crítica digital de las *Siete Partidas*. Este proyecto <https://7partidas.hypotheses.org/> (cons. 03/05/2022) se desarrolla desde la Universidad de Valladolid y cuenta con la financiación de la Agencia Estatal de Investigación del Reino de España y se integra dentro de la Red de Excelencia 'Cultura escrita medieval hispánica: del manuscrito al soporte digital (CEMH)' (RED2018-102330-T), Agencia Estatal de Investigación - Ministerio de Ciencia e Innovación.

En los comienzos del proyecto 7PartidasDigital, la Biblioteca Nacional de España me invitó a hablar en una jornada sobre la edición digital[1]. En esa ocasión verbalicé un sueño: poder tomar las imágenes digitales de un texto que hubiera en un repositorio y que gracias a un flujo de trabajo adecuado con las herramientas informáticas pertinentes pudiera obtener al final una edición crítica digital.

En 2021 estamos un poco más cerca de cumplir ese sueño, porque durante el último año hemos añadido al equipamiento filológico digital dos nuevas herramientas que facilitan el objeto final de la investigación: ofrecer la edición crítica digital de las *Siete Partidas*, aunque aún estamos lejos de llegar al final. Las herramientas incorporadas son la transcripción automatizada de testimonios impresos y un sistema de colación automática.

Antes de exponer cómo trabajan los editores digitales, quiero exponer cómo trabajaba como editor analógico. No es que un día me planteara dejar el mundo analógico y entregarme al digital. *Primum philologia, deinde computatrum scientia*; no se puede ser editor digital sin una sólida base de filología y crítica textual.

En 1979, me inicié en la crítica textual con un texto de principios del siglo XIV muy breve, apenas 28 folios a dos columnas, y con un total de casi 21000 palabras-token[2]: el ms. 9 de la Real Academia Española que contiene un *Tratado de cetrería*; un manuscrito encuadernado junto con una copia del XVIII que hizo Antonio de Santiago y Palomares, y del que localicé otra copia dieciochesca en la Biblioteca de la Fundación Universitaria Española (Archivo Campomanes 52-2) que había realizado el mismo Palomares. Por último, hallé el antecedente de la copia del trescientos, un grueso manuscrito de la Real Biblioteca de San Lorenzo de El Escorial, el ms. V-II-19 y el antecedente de una parte de este: el hoy RES/270 de la BNE, por aquel entonces en manos de unos anticuarios que estaban vendiendo algunos de los códices de la colección de Sir Thomas Phillipps (ms. 11719). El ms. de la RAE no era una copia del

[1] No existe publicación, pero en la red se puede visionar la comunicación <https://www.youtube.com/watch?v=QWlfSqtvGt4&t=1100s> (cons. 28/10/2021).

[2] Se entiende por palabra-token cualquier secuencia de caracteres alfanuméricos entre dos espacios en blanco o signos de puntuación; es decir, las *palabras ortográficas*.

escurialense, sino que, a partir de las seis obras recopiladas por una misma mano y con una misma letra, se había creado el texto antológico que se conserva en el manuscrito académico.

Por aquel entonces la filología española, a pesar de los grandes éxitos de la escuela pidaliana, no había producido un corpus teórico de referencia sobre crítica textual en el que los neófitos pudieran aprender algo que no se enseñaba en las aulas ni en los seminarios universitarios. Tan solo se contaba con un artículo publicado en 1917 por Américo Castro[3] y, casi tres lustros más tarde, otro artículo de Marín Ocete (1932), en el que exponía el estado actual de la crítica de textos. Para comprender la metodología empleada por la escuela de filología española había que recurrir a las introducciones de algunas ediciones preparadas por algunos de los más egregios miembros del Centro de Estudios Históricos como las que el mismo Castro menciona en su artículo, o la de Onís (1912) y, especialmente, la de García Solalinde (1930). Aunque de estos trabajos se infiere que lo que les preocupaba a sus autores era establecer «textos fidedignos» o «ediciones fidedignas» (Santiago, 2020, 228, n.15) o «ediciones dignas» (Santiago, 2020, 231), a pesar de que, como indica Fernández-Ordoñez (1988, 231), «don Ramón […] creía en el carácter científico y objetivo del método» de la crítica textual.

Así, pues, el único medio que había para aprender cómo hacer una edición de un texto (no me atrevo a decir que crítica) era leer y examinar las ediciones de textos medievales. De esas lecturas inferí un procedimiento que, con el tiempo y la aparición en 1983 del *Manual de crítica textual* de Alberto Blecua, fui perfeccionado y enriqueciendo desde el punto de vista teórico. Este manual llegó tarde para mí, pues defendí la tesis (Fradejas Rueda, 1983) el mismo año en el que el *Manual* salió de las prensas[4].

Años más tarde se estableció que en el *Diccionario de términos filológicos* de Fernando Lázaro Carreter (Santiago, 2005) se recogía la primera

[3] Reeditado siete años después (Castro, 1924).

[4] Wehrli (1951), López Estrada (1979) y Jauralde Pou (1981) dedicaron un breve capítulo informativo en sus manuales de introducción a los estudios literarios, como lo haría, años más tarde Ruiz (1985), esta a la vista ya del manual de Blecua (1983). Había mucho más movimiento en este sentido en la Universidad de Buenos Aires, en el Seminario de Crítica Textual fundado por Germán Orduna y en el que se comenzó en 1981 a editar la revista *Incipit*.

descripción del neolachmanianismo. Pero un diccionario no es el mejor lugar para aprender técnicas y métodos de investigación porque lo primero, quien lo hubiera pretendido, tendría que haber conocido de antemano cuáles eran esos términos técnicos. Además, la primera edición (Lázaro Carreter, 1953) solo incluía seis: *aparato crítico*, *colacionar*, *edición*, *fijación de un texto*, *filología* y *variante*. En la segunda (Lázaro Carreter, 1961) aumentó la nómina de términos hasta la cuarentena (Santiago, 2005). Tampoco se contaba con formación ni fuentes de información codicológica. El magnífico *Manual de codicología* de Elisa Ruiz no apareció hasta 1988.

La lectura y estudio del manual de Blecua (1983) confirmó que el barrunto metodológico no había sido muy equivocado. La verdad es que, a grandes rasgos los principios básicos para la edición crítica de un texto romance –no puedo hablar de otros ámbitos lingüísticos– son los mismos y son prácticamente inamovibles.

Lo primero es establecer el catálogo de las copias de la obra que se pretende editar. Por entonces, la *Bibliografía de la literatura hispánica* de Simón Díaz (1950-1993)[5] era la obra de referencia básica donde se iniciaban las búsquedas, aunque, como veremos, para el ámbito castellano medieval, en 1975, aparece una bibliografía que cambiará totalmente el panorama (Cárdenas *et al.*, 1975).

El siguiente paso era obtener las reproducciones de los testimonios. Hasta la primera década del siglo XXI lo que se obtenían eran micropelículas, en blanco y negro y en formato de 35 mm, aunque a veces podían ser de 16 mm. La mayor dificultad era que solían ser películas positivas, no negativas, por lo que no se podía ir a un laboratorio fotográfico del barrio y encargar que las ampliaran sobre papel fotográfico. Había que utilizar unas voluminosas máquinas que permitían leerlas. No obstante, poco a poco fueron apareciendo las que podían, con tan solo pulsar un botón, reproducirlas sobre papel, como si de una fotocopia se tratara. El último gran avance fue cuando se pudieron escanear los microfilms, obtener imágenes JPEG y manejarlas posteriormente con un ordenador.

[5] Los volúmenes dedicados a la Edad Media se publicaron en 1963 y 1965. En 1972 se inició la publicación de los tomos dedicados a los Siglos de Oro

Ya fueran ampliadas e impresas en papel, ya fueran leídas en la pantalla, había que transcribirlas a mano. Es decir, se realizaba una nueva copia del manuscrito y sobre esta copia, con muchísimo cuidado, se plasmaban todas las anotaciones necesarias. Un excelente ejemplo de este sistema son los cuadernos en los que Rafael Lapesa esbozó en los años 1950 su edición, jamás publicada, del *Rimado de Palacio* de López de Ayala (Lapesa, 2010).

Una vez hecha la transcripción y la labor crítica, había que pasarlo a máquina, mecanografiarlo. Pero las máquinas de escribir tenían unas posibilidades gráficas muy limitadas. Las máquinas de escribir españolas, por lo general, carecían de <ç> y no digamos de los imprescindibles corchetes y, por supuesto, de la ese alta <ſ> y la nota tironiana (el *ampersand* <&>, un excelente sustituto, no estaba en teclado español usual), por lo que había que dibujarlos a mano (la <ç> se obtenía al escribir una <c>, retroceder el carro una posición y sobrescribir una coma)[6].

Es decir, seguíamos inmersos en un mundo de transmisión manuscrita, sujeta a los mismos errores mecánicos de copia que sufrieron los copistas que escribieron los códices objeto de nuestro estudio. Pero no hacíamos una única copia, sino dos, al menos, la manuscrita y la mecanografiada, con lo que la tipología y las posibilidades de error de copia se ampliaban.

En agosto de 1981, IBM lanzó al mercado el IBM Personal Computer. En este momento es cuando realmente se inició la popularización del ordenador. Es cierto que ya existía el Apple II (1977) y otras máquinas como los Sinclair ZX Spectrum[7] o el Commodore Vic-20 (1980). Estos dos últimos, casi juguetes, fueron la plataforma de lanzamiento, pero los IBM y los clónicos que le siguieron fueron los que lograron que los ordenadores fueran accesibles a los especialistas de humanidades.

Al principio, no pasaron de ser potentísimas máquinas de escribir y con funciones de almacenaje. Los procesadores de texto, ya fueran Wordstar, WordPerfect o Word, cambiaron el modo de editar los textos,

[6] Tampoco tenía los números 1 y 0 para los que se usaba la <l> minúscula y la <O>.
[7] El Sinclair ZX Spectrum fue una evolución del Sinclair ZX80, lanzado en 1980, y del Sinclair ZX81, de 1981.

puesto que había un sistema de almacenaje y modificación de lo copiado que permitía una fácil corrección, sin tener que volver a copiar todo el folio. Los bloques de texto podían moverse de un lado a otro de los documentos, o guardarlos en ficheros separados para reutilizarlos cuando hiciera falta. No todo era perfecto, al principio el juego de caracteres fue muy limitado, pero pronto se amplió para incorporar la mayoría de las letras con diacríticos de la mayoría de las lenguas europeas occidentales.

Téngase en cuenta que estoy hablando de las parcas posibilidades de un profesor de universidad sin fondos para investigación y sin acceso a los centros de computación. Solo contaba con una grandísima curiosidad. Sin embargo, en otras latitudes estan empleando los ordenadores para la edición de textos y otras fruslerías filológicas (Nitti, 1978). Basta leer los capítulos de Hockey (1980, 149-167) acerca de los procedimientos de edición de textos a principios de los años 1980 para darse cuenta de por dónde iba el mundo.

Aunque ya había acceso a los ordenadores, como he dicho, los percibíamos potentes máquinas de escribir y la forma de trabajar en crítica textual se ha mantenido, salvo contadas excepciones, casi invariable hasta la segunda década del 2000. Quizá mejorara un poco debido a las muchas facilidades que los ordenadores e internet fueron ofreciendo con el paso del tiempo.

A la hora de editar la versión castellana medieval de la *Epitome rei militaris* (Fradejas Rueda, 2011) seguí empleando los viejos sistemas de colación[8]. Una vez decidido cuál era el texto que iba a usar de base, procedí a su transcripción y corrección en pantalla de todos los errores que se hubieran deslizado en este proceso. A continuación, lo imprimí en hojas A3 y, con un juego de rotuladores ultrafinos, con una amplia paleta de colores, un bloc de notas autoadhesivas para apuntar las notas que surgían a lo largo del proceso colación y una caja de hojas de etiquetas adhesivas –de las que se suelen usar para las etiquetas de dirección postal– para corregir los errores que pudiera cometer al anotar variantes, procedí a la *collatio* de los testimonios conservados. Este mismo procedimiento lo

[8] Varios autores han descrito los sistemas de colación manual que utilizaron (Manley y Rickert, 1940; West, 1973) e incluso algunos de si eran sistemas sólidos (Moorman, 1975, 47) o no (Foulet y Speer, 1979, 48).

utilicé poco después para la edición del *Libro de la caza de las aves* de Pero López de Ayala (Fradejas Lebrero y Fradejas Rueda, 2016). Este caso fue un poco más complejo porque se trataba de obra con una cuarentena de testimonios.

Sin embargo, desde mis orígenes como editor filológico, me interesé por los métodos informáticos. Al principio, se limitaban a un sistema de transcripción semipaleográfica codificada, que el Hispanic Seminary of Medieval Studies (HSMS) creó para el que quizá sea el proyecto pionero de las humanidades digitales hispánicas, aunque ideado por hispanomedievalistas norteamericanos de la Universidad de Wisconsin - Madison (Buelow y Mackenzie, 1986): el *Dictionary of Old Spanish Language* (DOSL), del que una de las derivadas más conocidas y utilizadas es la Bibliography of Old Spanish Texts (BOOST) que se convirtió en la Bibliografía Española de Textos Antiguos (BETA) dentro del base de datos bibliográfica *Philobiblon*, un catálogo de fuentes primarias para la literatura medieval peninsular[9].

Este sistema lo apliqué al texto del ms. 9 de la RAE y a todos los textos cetreros castellanos que fui localizando y transcribiendo a lo largo del tiempo. Además, desarrollé pequeños programas en BASIC para hacer las concordancias y los cálculos de frecuencia.

Años más tarde, los textos se publicaron dentro del corpus textual de Madison, en un sistema esotérico e inimaginable hoy día: en microfichas que requerían localizar una máquina que pudiera leerlas. Con el tiempo, se publicaron en CD-ROM, como sucedió con el corpus del *scriptorium alfonsí* (Kasten, Nitti, Jonxis-Henkemans, 1997) y el más amplio de manuscritos e impresos antiguos del HSMS (O'Neill, 1999).

La lección que extraje de esta época inicial, algo que siempre ha sido mi mayor preocupación, fue la absoluta necesidad de reutilizar los materiales y su perdurabilidad. Esto no siempre ha sido fácil porque, hasta

[9] La primera edición de este catálogo estuvo a cargo de Cárdena, Nitti y Mackenzie (1975). La última versión impresa, la tercera, fue cosa del equipo encabezado por Faulhaber *et al.* (1984). Se diseñaron las normas para una cuarta edición (Faulhaber y Gómez Moreno, 1986) pero se abandonó el proyecto analógico para convertirse en otro informático. Al principio como parte del proyecto *Admyte* (1991), del que se independizaría y adquiriría vida como un gran catálogo en línea que integra tres bibliografías: gallego-portuguesa (BITAGAP), catalana (BITECA) y española (BETA) bajo el nombre general de *Philobiblon* (Faulhaber, 1997).

la estandarización del sistema MS-DOS, cada fabricante utilizaba su propio sistema operativo. El sistema operativo CP/M (Control Program for Microcomputers) era casi específico para cada fabricante o grupo de fabricantes. Bajo este sistema operativo se desarrolló el procesador de textos Wordstar y la base de datos dBase, por lo que pasar de una máquina a otra podía ser problemático.

En mi caso, empecé con un ordenador Toshiba T-100 con CP/M, y cuando quise cambiar de ordenador, a un IBM PC, tuve que buscar una pasarela que me permitiera pasar los datos desde discos grabados bajo CP/M al sistema de archivos de IBM-DOS. El traspaso de los datos fue a través de dos máquinas diferentes: del Toshiba T-100 a un Osborne 1, pues este tenía ordenador una rutina que permitía leer los discos CP/M del Toshiba y, una vez en formato Osborne, mediante un Decision Mate V de NCR, que podía leer el CP/M de Osborne, los pude transformar en MS-DOS (IBM-DOS) sin pérdida de datos. Desde ahí han llegado a ordenadores Apple y desde discos de 5.25" de 160 Kb, a discos de 3.5" de 1.4 Mb, a memorias USB y, finalmente a *drives* en la nube. Es un problema que hay que tener en cuenta. Hoy casi lo tenemos solucionado, pero muchas cosas de gran interés como Admyte (1992-1998) o la magnífica edición del *Poema del Cid* de la BNE (Manuscrito 1998) han quedado obsoletas puesto que no se actualizaron.

Personalmente, a lo largo de estos años de cambios, he participado en varios proyectos de informatización, digitalización y edición de textos medievales españoles. Algunas de las transcripciones que preparé para el HSMS (Fradejas Rueda, 1992a; 1992b) se incorporaron a Admyte (1992-1998). Participé en la Colección Clásicos Tavera (Fradejas Rueda, 1999), que consistía en la digitalización de los objetos físicos (manuscritos e impresos) a base de convertir los microfilm de la BNE en imágenes jpeg e incluirlas en CD-ROM. Por su parte, el Centro Virtual Cervantes creó una colección de ediciones digitales en línea y me encargaron la de la *Historia de Enrique fi de Oliva* (Fradejas Rueda, 1997). En todos estos casos, actué como editor filológico; no estaba entre mis cometidos la parte tecnológica, pero no quiere decir que no estuviera al tanto de las posibilidades.

En la segunda mitad de los años 90 se abrió un nuevo mundo:

cualquiera podía crear páginas web. En la UNED, que era la universidad en la que era profesor en aquellos momentos, comencé a desarrollar, más como *hobby* que como proyecto de investigación, una web en la que editar y publicar todos los textos de cetrería que había ido trascribiendo y editando a lo largo de los años.

Cuando en el año 2000 me trasladé a la Universidad de Valladolid, reubiqué en sus servidores de la Faculta de Filosofía y Letras aquella página web. Sorprendió bastante al personal del centro de cálculo (STIC) que uno «de letras» quisiera hacer una página web para un proyecto. Al principio hubo algo de reticencia, pero lo conseguí[10].

A raíz de los trabajos para la declaración de la cetrería como Patrimonio Intangible de la Humanidad diseñé un proyecto que presenté durante el congreso *Falconry: A World Heritage*, celebrado en Abu Dhabi en septiembre de 2005, para el que solicité financiación al MEC / MINECO al año siguiente (referencias HUM2006-0932/FILO y FFI20210-15128). Lo titulé *Archivo Iberoamericano de Cetrería* (www.aic.uva.es). Hoy sigue vivo, aunque en fase letárgica porque no hay materiales que añadir.

[10] No quedan rastros de esas viejas versiones, salvo lo que se puede localizar en la Wayback Machine, un sistema que almacena instantáneas de sitios web y hoy se pueden recuperar con cierta facilidad, como en este caso <https://web.archive.org/web/20010520162153/http://gramola.fyl.uva.es/~cetreria/> (cons. 29/10/21), que ofrece el acceso a cómo estaba la página el 20 de mayo de 2001.

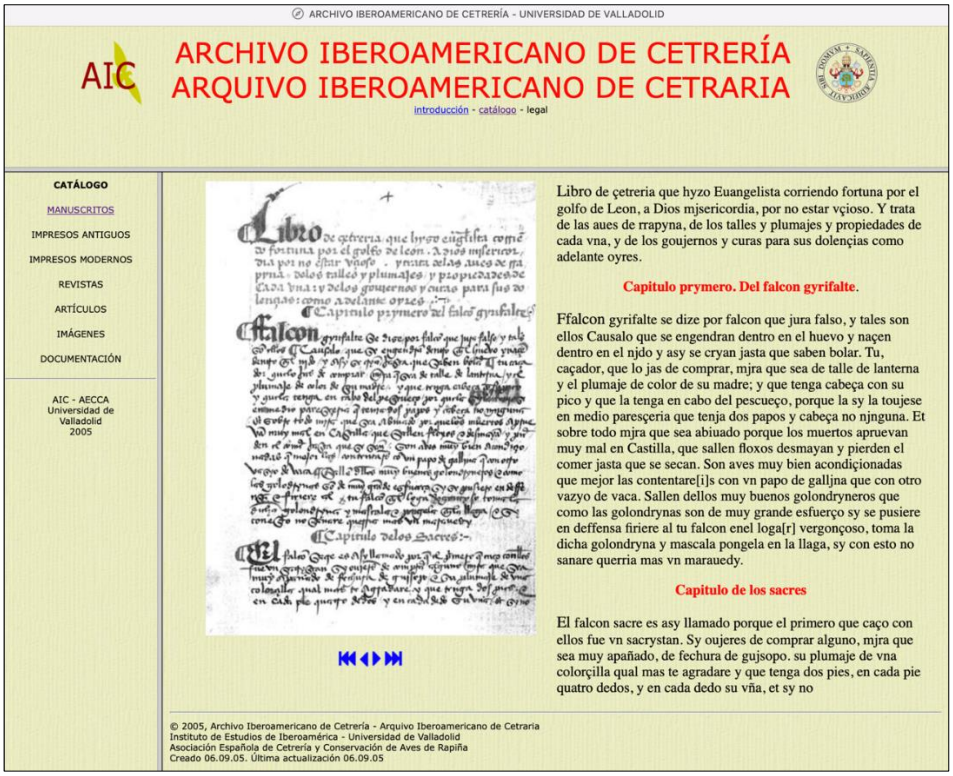


Fig. 1. Edición del primer folio del *Libro de cetrería* de Evangelista según el Mss/21549 de la BNE en la primera versión del *Archivo Iberoamericano de Cetrería* (2005)

En el diseño original, aspiraba a crear una edición que se asemejara a lo que se había desarrollado en el disco 1 de Admyte (1992-1998): editar los textos junto con las imágenes. Esto se desarrolló en HTML puro, un lenguaje presentacional en el que los datos y los aspectos visuales van unidos. Además, se diseñó con un sistema de marcos, por lo que el mantenimiento era complicado y la adición de nuevos materiales un difícil (Fig. 1).

Esta tarea se simplificó con el uso de hojas de estilos en cascada (CSS), sistema que separa el contenido (los textos) del aspecto visual. Esta fue otra lección muy importante en mi evolución como editor digital: mantener separados el contenido y el aspecto visual. A esto último se le suele dar mucha importancia hoy en día; algo que no ha preocupado mucho a un buen número de proyectos de edición digital, los cuales nacen

con una visión de lo que quieren obtener como producto final y no tanto de las fases intermedias, que son las básicas.

Charles Faulhaber, entonces director de la Bancroft Library (Universidad de California, Berkeley) en respuesta a uno de mis correos, dirigió mi atención hacia un sitio web de la Universidad de Berkeley: *Digital Scriptorium*[11]. Ahí fue donde descubrí la Text Encoding Initiative y sus posibilidades[12]. Esto fue especialmente revelador porque uno de los ejemplos con los que los desarrolladores ilustraban su metodología era un texto medieval castellano, con lo que pude ver las posibilidades que ese sistema tenía. *Digital Scriptorium* y los motores de búsqueda me llevaron al proyecto Medieval Nordic Text Archive[13] (MENOTA) y sus tres capas de transcripción: 1) facsimilar, 2) paleográfica y 3) normalizada. Este proyecto me llevó, a su vez, al Medieval Unicode Font Initiative (MUFI)[14], algo imprescindible para algunas de las posibilidades que abordaba MENOTA.

La TEI era la respuesta a los problemas con los que me había topado hasta entonces en la edición de textos medievales e inicié la redacción, tras un curso del TEI@Oxford Summer Schools en 2010, de una especie de manual con ejemplos castellanos de todas aquellas etiquetas que creí que eran interesantes para mi idea de una edición digital de un texto medieval castellano (Fradejas Rueda, 2010).

Por esos años, daba por agotado mis potenciales contribuciones sobre el vastísimo campo de los libros de cetrería españoles y comencé a diseñar un nuevo proyecto que rompiera textualmente con el anterior. La chispa iniciadora de este cambio de rumbo fue la presentación de *Alma littera* (Herrero de la Fuente, 2014). Uno de los artículos de este volumen

---

[11] <https://digital-scriptorium.org/> (cons. 28/10/21). Los contenidos actuales poco tienen que ver con lo que había durante la primera década del siglo XXI. Se puede explorar por medio de Wayback Machine: <https://web.archive.org/web/20070609225805/; https://www1.columbia.edu/sec/cu/libraries/bts/digital_scriptorium/technical/ds-xml/transcription_dtd/documentation/toc.html> (cons. 28/10/21).

[12] Faulhaber ya había hablado sobre la TEI en España, pero sus ideas quedaron soterradas en las actas del congreso de la lengua española de Sevilla de 1992 (Faulhaber, 1994).

[13] <https://www.menota.org/EN_forside.xhtml> (cons. 28/10/21). Este proyecto se inició en 2001 y en él se estableció que «Det var enighet om at arkivets primære målgruppe er forskere og studenter innenfor middelalderfilologi», pero que no descartaban que en el futuro fuera «for et bredere publikum». <https://www.menota.org/DOK_RaadsReferat2001-09-10.xml> (cons. 29/10/21).

[14] <https://mufi.info/m.php?p=mufi> (cons. 28/10/21).

era la edición y transcripción de un fragmento de los *Bocados de Oro* (Ruiz Albi, 2014) que se encontraba en la sección de pergaminos del Archivo de la Real Chancillería de Valladolid. Allí descubrí cuatro fragmentos de las *Siete Partidas* (Fradejas Rueda, 2015), aunque después supe de dos más.

De ahí surgió el proyecto de una edición crítica digital de las *Siete Partidas*, un proyecto considerado descabellado por algunos colegas porque había cientos de testimonios (véase más adelante). Es obvio que se trataba de una tarea compleja, sin embargo, mi interés no disminuyó porque es la única gran obra de la edad media castellana que aún carece de una edición crítica; el último intento es de hace dos siglos: la edición de la Real Academia de la Historia.

Estimé que la mejor manera de conseguir la edición crítica digital de las *Siete Partidas* era por medio de la transcripción de todos los testimonios conservados, por lo que mi idea inicial, actual y posiblemente final sea una edición sinóptica digital integral alineada al estilo de la que ofrece Enrique-Arias en el proyecto *Biblia Medieval*[15].

El gran problema residía en cómo, algo que había sido escrito sobre pergamino y papel entre los reinados de Alfonso X y el de los Reyes Católicos y que conoció una primera etapa trepidante y no muy bien aclarada aún, y una segunda como obra en letra de molde, entre 1491 y 1555, y que tenía letras que eran relativamente sencillas de leer para un humano medianamente entrenado, podía ser convertido en texto comprensible y manejable por los ordenadores en un tiempo razonable.

Decidido a que se tratara de una edición sinóptica integral alineada (paralela) digital, la siguiente decisión fue el uso del sistema de codificación TEI. No había otra posibilidad. Para tener un arranque abarcable y relativamente sencillo, se comenzó por los impresos. De ahí que la primera fase del proyecto (2016-2020) se centrara en las llamadas ediciones históricas, la *princeps* incunable de 1491, la renacentista de 1555 y la académica de 1807. Se iniciaron los trabajos por la edición de 1555, el *textus receptus*, la *editio vulgata*, por la sencilla razón de que se trata de la versión con validez legal y que los especialistas del derecho utilizan.

No podemos olvidar que las *Siete Partidas* es un texto jurídico que

[15] <https://www.bibliamedieval.es/> (cons. 28/10/21).

sigue teniendo validez en los tribunales de España y América, incluido, y en especial, Estados Unidos de América, aunque desde el punto de vista filológico tuviéramos que huir de ella porque es una edición contaminada, no por una mera *consultatio* o *contaminatio* sino porque es un caso flagrante de *conflatio*[16]. Pero poseía un elemento de sencillez adicional: la tipografía era redonda y no gótica, como en el caso de los incunables.

Se comenzó por la edición de Gregorio López de 1555 para ahorrar el esfuerzo extra que supone el aprendizaje del sistema de codificación TEI (Fradejas Rueda, 2021b). Interesaba que los transcriptores se preocuparan por los problemas del texto, no por las posibles complejidades de las etiquetas TEI. Para ello se diseñaron unas protomarcas muy elementales, lo que permitía a cada colaborador utilizar el procesador de textos con el que se sintiera más cómodo. La única restricción era seguir al pie de la letra las instrucciones.

Cada fichero solo debe contener un título; se transcribe respetando la longitud de línea; se desarrollan y marcan las abreviaturas con cursivas; las rúbricas con negritas; se unen y separan las palabras con criterio actual; se marca comienzo de folio y de columna entre corchetes; los títulos corrientes con una llave de apertura y las *littera nobilior* con una llave seguida de las letras IN y un dígito que declara la altura de la inicial expresada en número de líneas. En la eventualidad de que algo no se entendiera, o hubiera erratas tipográficas, se añade a continuación de esa palabra [sic] porque no se corrige en ningún caso lo que dice el *original*.

Estos ficheros posteriormente se convertían en otros etiquetados con TEI muy básico por medio del filtro TEIOOP5 para Open / Libre Office. Después ese fichero se procesaba con un *script* en el lenguaje de

[16] La edición de López (1555) toma como base un ejemplar la de Díaz Montalvo (1491, 1501, 1528, 1542 y 1550) y la corrige con ayuda de varios manuscritos que nos son totalmente desconocidos. Se ha podido demostrar que López usa como texto base la edición de 1550 (Fradejas Rueda, en prensa), la cual deriva de la edición de 1528 (por el momento no se puede establecer si la de 1550 es copia directa de la de 1528 o de la de 1542, lo más probable). La edición de 1528 es una edición corregida a la vista de otros manuscritos, corrección que llevó a cabo Francisco de Velasco. En el caso del texto de la *Primera Partida*, tuvo como modelo para la corrección un ejemplar de la llamada redacción primitiva (Craddock, 1974) y esto se demuestra por que presenta veinticinco títulos debido al desdoblamiento del título 1.19 y la incorporación del correspondiente proemio para el nuevo título 1.20. Por lo tanto, de Velasco tuvo acceso a otros códices de otras *Partidas*, pero no puede determinarse cuáles (Fradejas Rueda, en prensa).

programación R que se ocupaba de establecer las particularidades del modelo de transcripción diseñado para el proyecto e incluía la generación del xml:id que individualiza cada una de las casi 2700 leyes que constituyen las *Partidas*.

En el caso del incunable de 1491 se partió de la transcripción del HSMS, se analizaron las etiquetas y marcas utilizadas y por medio de otro *script* en R se obtuvo la codificación TEI que había establecido para el proyecto 7PartidasDigital.

En el caso de la edición de 1807 se partió de la versión OCR que ofrece la Biblioteca Digital Hispánica de la Biblioteca Nacional de España. Esta se ha corregido (aún estamos ello) y codificado en TEI. Este testimonio ha dejado de tener interés en nuestro proyecto porque se trata de un *descriptus* de un grupo de códices de la BNE y no supone nada en la transmisión de las *Siete Partidas*, aunque puede tener implicaciones serias en los estudios lingüísticos que se basan en esta edición porque se trata, como en el caso de la edición de Montalvo (1491) y de López (1555), de un caso de *conflatio* en el que solo ocasionalmente se advierte el cambio de modelo o la procedencia de la lección seleccionada.

Mientras se llevaba a cabo este proceso de transcripción y codificación semiautomático, se desarrollaron las protomarcas e instrucciones para la transcripción de los testimonios manuscritos. Se han probado con MN0, MN6 y ZAB y funcionan, aunque siempre es necesario afinarlas, pues no hay manuscrito que no depare alguna sorpresa. Gracias a esto, en la nueva andadura de 7PartidasDigital (2021-2024, ref. PID2020-112621GB-I00) se aplicará el mismo procedimiento para la transcripción y codificación de todos los testimonios manuscritos de la *Primera* y *Cuarta Partida*.

El siguiente problema con el que nos enfrentamos es el de la colación de los testimonios. El corpus de la 7PartidasDigital lo constituyen unos 100 testimonios con desigual número de copias para cada una de las *Partidas* (Fradejas Rueda, 2021d). Además, un problema básico de las *Siete Partidas* es que no se conoce ningún testimonio manuscrito completo. Esto se consiguió en 1491 con la edición de Díaz de Montalvo, revisada a la vista de varios manuscritos, que desconocemos, en 1528 por Francisco Velasco y cuya edición fue tomada como testimonio de base por Gregorio

López, que la corrigió a la luz de una serie ignota de manuscritos.

Partida	Copias	Palabras	Palabras totales
1	15 - 1	143.000	2.000.000
2	25 - 1	142.000	3.500.000
3	17 - 3	184.000	2.000.000
4	12 - 3	62.000	560.000
5	13 - 3	85.000	850.000
6	15 - 1	64.000	900.000
7	17 - 2	83.000	1.240.000
Totales	100	763.000	11.000.000

Tabla 1. Cálculo estimado de palabras gráficas (tokens) en los testimonios de las *Siete Partidas*

Estos cien testimonios suponen unos once millones de palabras, como puede comprobarse en la tabla 1, por lo que colacionar manualmente, como se ha expuesto con anterioridad, sería imposible.

A principios de la década de 1960, Vinton A. Dearing (1962, 20) describió un programa creado para un ordenador *mainframe*, un IBM 7090, que le permitía comparar hasta noventa y nueve textos y aventuraba que «reasonably flexible machines will in time appear on the market at prices that will make them attractive to universities […]. It will them be possible to do textual comparisons in minutes instead of months».

La evolución de los ordenadores, tanto en potencia de procesamiento como en sencillez de manejo y abaratamiento, ha sido enorme desde entonces. Así, Hockey (1980) ofreció una reseña de las técnicas de computación utilizadas en la crítica textual en la primera época, revisión que reelaboró posteriormente y consideró que «of the many collation system that have been written, two stand out because of their sophistication and functionality»: TUSTEP y Collate (Hockey, 2000, 126). Ambos programas fueron diseñados para ser ejecutados en ordenadores personales, TUSTEP - TÜbinger System von TExtverarbeitungs-Programmen (Castrillo Benito, 1992) sobre MS-DOS (Windows) y Collate en los Apple Macintosh (Robinson, 1989a y 1989b). Así mismo, ambos

programas siguen existiendo y se han actualizado constantemente; sin embargo, no son programas sencillos de manejar puesto que están diseñados para llevar a cabo todas las fases necesarias para una edición crítica de un texto, incluso la preparación final para la imprenta (TUSTEP) o Internet (Collate). Sin embargo, diversos centros de investigación han ofrecido otros programas menos complicados, que pueden realizar la colación de manera sencilla y rápida, a la par que ofrecer otras informaciones y funcionalidades básicas.

En España, Marcos Marín desarrolló un programa llamado UNITE que era «capaz de comparar hasta 30 versiones de un mismo texto para obtener una versión única»[17] (Admyte, 1991: 5). Este programa que se incluyó en el disco 0 de Admyte (1992-1998), tenía una fuerte limitación: estaba diseñado para trabajar con textos en verso. El único ejemplo de la aplicación de este programa es la edición unificada del *Libro de Alexandre* que él mismo publicó (1987)[18].

Alonso Rioja (1996) desarrolló una aplicación para colación de textos que llamó AFTL, acrónimo de Análisis Filológico de Textos según el método de Lachmann. Lo diseñó para comparar «tres copias o traducciones de un mismo original» y «entresacar las diferencias existentes línea a línea y palabra a palabra entre los tres textos» (1996, 8). Con esta aplicación pretendía «automatizar el proceso de comparación siendo capaz de reconocer en un máximo de tres líneas consecutivas, todos los tipos de diferencias (omisión, interpolación, difracción o inversión)» (Alonso Rioja, 1996, 8). Aunque existen los discos, ha sido imposible instalar la aplicación de acuerdo con las instrucciones que se ofrecen en el manual de usuario de la aplicación AFTL (Alonso Rioja, 1996, 81-91). Sin embargo, se ofrece una serie de impresiones de las diversas ventanas que muestran qué era capaz de hacer la aplicación.

Un programa de colación básico, pero potente y flexible es Juxta. Se

[17] La parte inglesa de este folleto es menos ambiciosa, pues indica que es «[a] program designed to collate and compare up to ten versions of a given poetic text in order to obtain their presumed archetype, serving thereby as an aid in the production of a critical edition of that text» (Admyte, s. d., 5).

[18] Esta edición es hoy accesible a través de la Biblioteca Virtual Miguel de Cervantes <http://www.cervantesvirtual.com/obra/libro-de-alexandre--0/> (cons. 28/10/21) y la información técnica es, por tanto, accesible.

trata de una aplicación desarrollada en la University of Virginia[19] escrita en Java que funciona en ordenadores Windows, Linux y Apple. Sin embargo, ha dejado de funcionar en las máquinas Apple cuyo sistema operativo sea 11.4 o superior, aunque sigue funcionando en el entorno Windows 10 y Linux. Es un software muy valioso dado que puede leer ficheros con etiquetado TEI.

El flujo de trabajo es que tras cargar los textos y seleccionar uno de ellos como término de comparación, tras unos segundos se oscurecen todas aquellas palabras en las que no hay igualdad. En una de las posibles visualizaciones, al posar el cursor sobre una palabra con variación se despliegan todas las variantes que presentan los diferentes testimonios (Fig. 2). Juxta tiene, incluso, la capacidad de generar el aparato crítico, que no es nada más que una página HTML que puede transformarse, con relativa sencillez, en un aparato crítico de acuerdo con el sistema TEI.

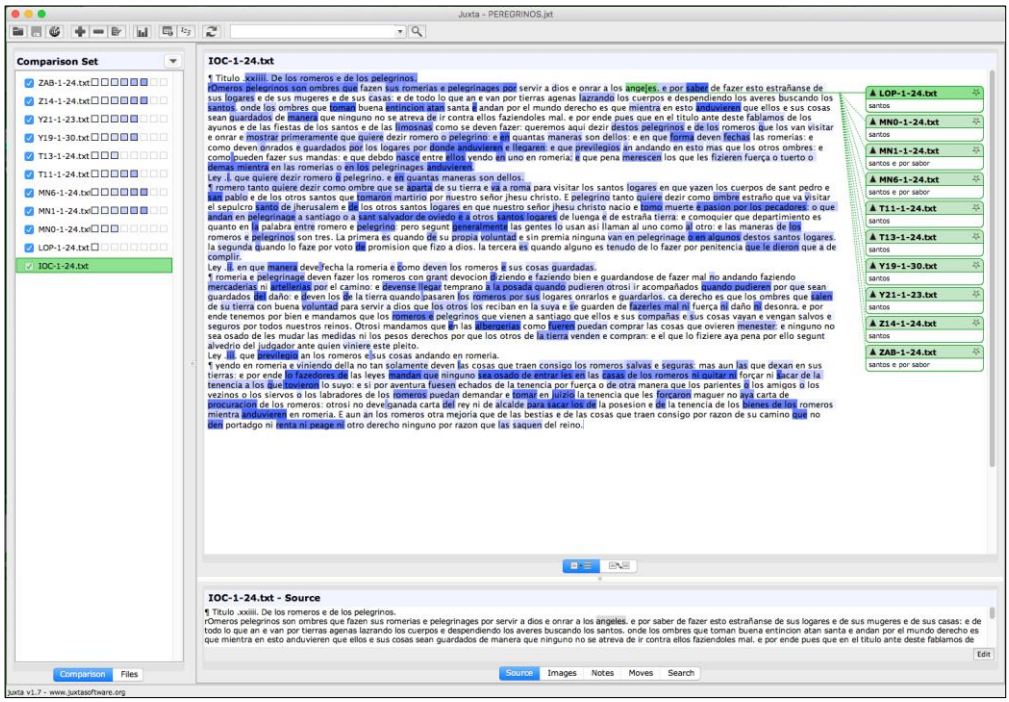


Fig. 2. Resultado en Juxta de la colación de todos los testimonios del título 1.24 de las *Siete Partidas*

[19] <https://www.juxtasoftware.org/> (cons. 28/10/21).

Es un programa magnífico, pero tiene un problema: los ordenadores son obstinadamente literales y la variación ortográfica que presentan los manuscritos medievales provoca gran ruido. En una primera instancia se puede resolver el problema por medio de una regularización textual. En el caso de los textos castellanos medievales apliqué, aunque no comparto los criterios, los principios de presentación crítica desarrollados por Sánchez-Prieto Borja (1998; 2011). Para conseguirlo procesé el texto con la aplicación desarrollada para el etiquetado morfológico automático del Old Spanish Textual Archive (Gago Jover y Pueyo, 2018a; 2018b; 2020).

Puesto que Juxta no funciona en las máquinas Apple con sistemas operativos más recientes, se buscó otra solución para llevar a cabo la colación automática. La respuesta ha sido el programa CollateX[20]. Aunque puede funcionar como una aplicación Java, es posible usarlo desde Python, lo cual me permite mecanizar ciertos aspectos del pre- y postprocesado de los textos.

```
from collatex import *
collation = Collation()
testimonio_IDI = open( "PEREGRINOS/IDI-1-24.txt", encoding='utf-8' ).read()
testimonio_IOC = open( "PEREGRINOS/IOC-1-24.txt", encoding='utf-8' ).read()
testimonio_LOP = open( "PEREGRINOS/LOP-1-24.txt", encoding='utf-8' ).read()
testimonio_MN0 = open( "PEREGRINOS/MN0-1-24.txt", encoding='utf-8' ).read()
testimonio_MN1 = open( "PEREGRINOS/MN1-1-24.txt", encoding='utf-8' ).read()
testimonio_MN6 = open( "PEREGRINOS/MN6-1-24.txt", encoding='utf-8' ).read()
testimonio_T11 = open( "PEREGRINOS/T11-1-24.txt", encoding='utf-8' ).read()
testimonio_T13 = open( "PEREGRINOS/T13-1-24.txt", encoding='utf-8' ).read()
testimonio_Y19 = open( "PEREGRINOS/Y19-1-30.txt", encoding='utf-8' ).read()
testimonio_Y21 = open( "PEREGRINOS/Y21-1-23.txt", encoding='utf-8' ).read()
testimonio_Z14 = open( "PEREGRINOS/Z14-1-24.txt", encoding='utf-8' ).read()
testimonio_ZAB = open( "PEREGRINOS/ZAB-1-24.txt", encoding='utf-8' ).read()
salida = open( 'pregrinos.txt', 'w', encoding='utf-8')
collation.add_plain_witness( "IOC", testimonio_IOC )
collation.add_plain_witness( "IDI", testimonio_IDI )
collation.add_plain_witness( "LOP", testimonio_LOP )
collation.add_plain_witness( "MN0", testimonio_MN0 )
collation.add_plain_witness( "MN1", testimonio_MN1 )
collation.add_plain_witness( "MN6", testimonio_MN6 )
collation.add_plain_witness( "T11", testimonio_T11 )
collation.add_plain_witness( "T13", testimonio_T13 )
collation.add_plain_witness( "Y19", testimonio_Y19 )
collation.add_plain_witness( "Y21", testimonio_Y21 )
collation.add_plain_witness( "Z14", testimonio_Z14 )
collation.add_plain_witness( "ZAB", testimonio_ZAB )
tabla = collate(collation, segmentation=False, near_match=True, layout='vertical')
print(tabla, file=salida)
```

Fig. 3. *Script* en Python para la colación de los testimonios

[20] <https://collatex.net/> (cons. 28/10/21).

Es muy sencillo de usar. La figura 3 muestra el *script* que he utilizado para colacionar el título 1.24, el último de la *Primera Partida*, en todos los testimonios a los que he tenido acceso. Las dos primeras líneas invocan la librería. Las líneas que comienzan con `testimonio_` leen el texto del título elegido en todos los testimonios que lo tienen. A continuación, se declara cómo se llamará el fichero de salida y qué codificación tendrá. En el siguiente paso, todas las líneas que comienzan con `collation.`, alimentan cada uno de los textos, pero identificados por la sigla, a la función `collation`, la cual tendrá en cuenta que no se quiere segmentar el texto, que se ha de aplicar el parámetro `near_match` y que el resultado ha de ofrecerse verticalmente, pues no sucede como en Juxta, en la que el resultado es visual.

CollateX puede ofrecer grafos (Figura 4) que permiten seguir la trayectoria y conexión entre las diferentes variantes y, aunque también puede ofrecer los resultados en TEI (según las normas del módulo *Critical Apparatus*), considero que la mejor opción es en forma de una tabla (Figura 5), pues en esta fase del proyecto porque permite detectar errores de transcripción que de otra manera pasarían desapercibidos. Reprocesar los ficheros tras las correcciones manuales no toma demasiado tiempo.

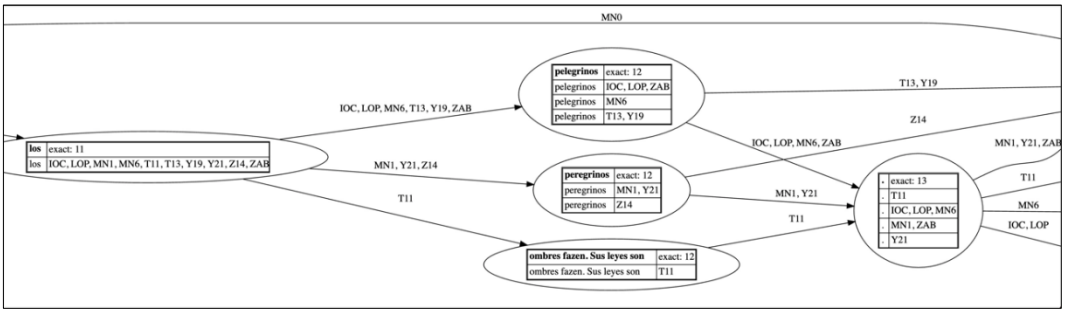


Fig. 4. Grafo (parcial) del resultado de la colación del título 1.24 de la *Primera Partida*

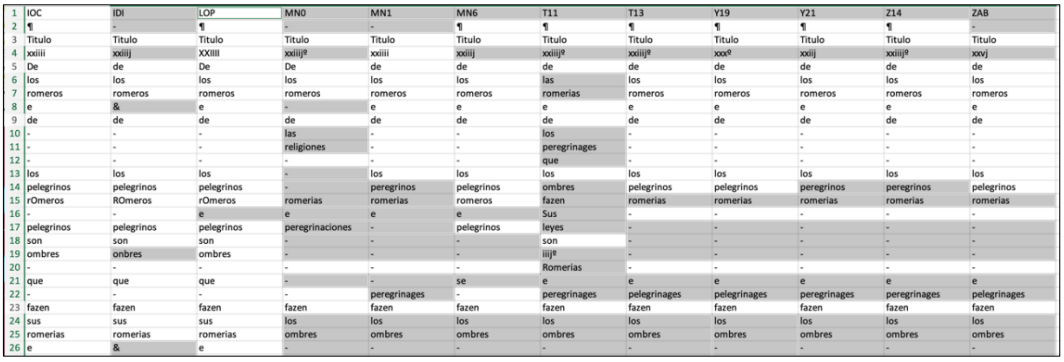
Antes he dicho que los ordenadores son recalcitrantemente literales y consideran como formas diferentes palabras como *rOmeros*, *Romeros*, *romeros*, lo que arroja un elevado número de variantes de nulo valor. Se puede reducir este ruido, como en el caso de Juxta, por medio de una transformación de los ficheros al modelo de presentación crítica

desarrollado por Sánchez-Prieto Borja (1998; 2011). Sin embargo, CollateX tiene integrado en su algoritmo la posibilidad de `near_match`, que, declarándolo como verdadero, obvia el problema y considera todas esas variantes como indiferentes, con lo que se reduce el ruido y las manipulaciones del texto.

IOC	IDI	LOP	MN0	MN1	MN6	T11	T13	Y19	Y21	Z14	ZAB
¶	-	¶	-	-	¶	¶	¶	¶	¶	¶	-
Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo
xxiiii	xxiiij	XXIIII	xxiiij⁰	xxiiii	xxiiij	xxiiij⁰	xxiiij⁰	xxx⁰	xxiij	xxiiij⁰	xxvj
De	de	De	De	de	de	de	de	de	de	de	de
los	los	los	los	los	los	las	los	los	los	los	los
romeros	romeros	romeros	romeros	romeros	romeros	romerias	romeros	romeros	romeros	romeros	romeros
e	&	e	-	e	e	e	e	e	e	e	e
de	de	de	de	de	de	de	de	de	de	de	de
-	-	-	las	-	-	los	-	-	-	-	-
-	-	-	religiones	-	-	peregrinages	-	-	-	-	-
-	-	-	-	-	-	que	-	-	-	-	-
los	los	los	-	los	los	los	los	los	los	los	los
pelegrinos	pelegrinos	pelegrinos	-	peregrinos	pelegrinos	ombres	pelegrinos	pelegrinos	peregrinos	peregrinos	pelegrinos
rOmeros	ROmeros	rOmeros	romerias	romerias	romeros	fazen	romerias	romerias	romerias	romerias	romerias
-	-	e	e	e	e	Sus	-	-	-	-	-
pelegrinos	pelegrinos	pelegrinos	peregrinaciones	-	pelegrinos	leyes	-	-	-	-	-
son	son	son	-	-	-	son	-	-	-	-	-
ombres	onbres	ombres	-	-	-	iiij⁰	-	-	-	-	-
-	-	-	-	-	-	Romerias	-	-	-	-	-
que	que	que	-	-	se	e	e	e	e	e	e
-	-	-	-	peregrinages	-	peregrinages	pelegrinages	pelegrinages	peregrinages	peregrinages	pelegrinages
fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen
sus	sus	sus	los	los	los	los	los	los	los	los	los
romerias	romerias	romerias	ombres	ombres	ombres	ombres	ombres	ombres	ombres	ombres	ombres
e	&	e	-	-	-	-	-	-	-	-	-

Fig. 5. Tabla que ofrece como resultado el *script* de la figura 3

La tabla (Fig. 5) que ofrece el *script* es un tanto muda. Tras editarla para convertirla en una hoja de Excel (Fig. 6) es mucho más elocuente, dado que en Excel tiene la posibilidad de marcar las filas (casillas) en las que hay una diferencia. Aquí volvemos a la literalidad de los ordenadores, aunque no es tan desesperante como en el caso de Juxta porque Excel no considera como diferentes *rOmeros*, *Romeros*, *romeros* (línea 15) pero sí la ausencia de un término (línea 22), o el cambio de un término o forma por otro: *las* por *los* (línea 6), *romeros* frente a *romerias* (línea 7), o *romerias* frente a *ombres* (línea 25).



	IOC	IDI	LOP	MN0	MN1	MN6	T11	T13	Y19	Y21	Z14	ZAB
1	¶	-	¶	-	-	¶	¶	¶	¶	¶	¶	-
2	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo	Titulo
3	xxiiii	xxiiij	XXIIII	xxiiij⁹	xxiiii	xxiiij	xxiiij⁹	xxiiij⁹	xxx⁹	xxiij	xxiiij⁹	xxvj
4	De	de	De	De	de	de	de	de	de	de	de	de
5	los	los	los	los	los	los	las	los	los	los	los	los
6	romeros	romeros	romeros	romeros	romeros	romeros	romerias	romeros	romeros	romeros	romeros	romeros
7	e	&	e	-	e	e	e	e	e	e	e	e
8	de	de	de	de	de	de	de	de	de	de	de	de
9	-	-	-	las	-	-	los	-	-	-	-	-
10	-	-	-	religiones	-	-	peregrinages	-	-	-	-	-
11	-	-	-	-	-	-	que	-	-	-	-	-
12	los	los	los	-	los	los	los	los	los	los	los	los
13	pelegrinos	pelegrinos	pelegrinos	-	peregrinos	pelegrinos	ombres	pelegrinos	pelegrinos	peregrinos	peregrinos	pelegrinos
14	rOmeros	ROmeros	rOmeros	romerias	romerias	romeros	fazen	romerias	romerias	romerias	romerias	romerias
15	-	-	e	e	e	e	Sus	-	-	-	-	-
16	pelegrinos	pelegrinos	pelegrinos	peregrinaciones	-	pelegrinos	leyes	-	-	-	-	-
17	son	son	son	-	-	-	son	-	-	-	-	-
18	ombres	onbres	ombres	-	-	-	iiij⁸	-	-	-	-	-
19	-	-	-	-	-	-	Romerias	-	-	-	-	-
20	que	que	que	-	-	se	e	e	e	e	e	e
21	-	-	-	-	peregrinages	-	peregrinages	pelegrinages	pelegrinages	peregrinages	peregrinages	pelegrinages
22	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen	fazen
23	sus	sus	sus	los	los	los	los	los	los	los	los	los
24	romerias	romerias	romerias	ombres	ombres	ombres	ombres	ombres	ombres	ombres	ombres	ombres
25	e	&	e	-	-	-	-	-	-	-	-	-

Fig. 6. La misma tabla de la figura 5 pero editada para manejarla en Excel

El uso de estos métodos informáticos me llevó a detectar, por ejemplo, problemas lingüísticos que de otra manera habrían pasado desapercibidos. En un recuento léxico por lemas de las ediciones *princeps* y de 1555 de las *Siete Partidas* constaté una interesante diferencia: la edición del quinientos tiene, en la *Primera Partida*, un tinte lingüístico medieval que la de 1491 no presenta. Simplificando la cuestión, en el texto de López hay preferencia por *maguer*, *guisar*, *toller*, *vegadas* y *ca* con una diferencia de uso que oscila entre el 100 % (caso de *toller*) y el 50 % (*guisa*), frente a las formas del texto de 1491. Esto llevó a preguntarnos cómo un texto que, sabemos, es copia de otro anterior (Fradejas Rueda, 2021c), pudo dar un paso atrás lingüísticamente.

He llegado a la respuesta por medio de la bibliografía material (Fradejas Rueda, 2021a), pero la respuesta incontestable la ofrecerán los métodos digitales. Ahí se ha incorporado una nueva herramienta al flujo de trabajo de 7PartidasDigital: la transcripción automática por medio de Transkribus[21]. Ya la he probado con la edición de diciembre de 1491 de las *Siete Partidas*, gracias al modelo desarrollado para las góticas del siglo XV y XVI castellanas (Bazzaco, 2020). Con este modelo he obtenido un porcentaje de acierto del 99.3 %, o lo que es lo mismo, solo el 0.7 % de las formas son erróneas, por lo que se han transcrito las casi mil páginas que constituyen esta edición en poco menos de una semana[22].

21 <https://readcoop.eu/transkribus/> (cons. 28.10.21).

22 Eduardo Camero Santos, estudiante de doctorado en la Universidad de Valladolid, va a poner el sistema a prueba con la transcripción y codificación de la edición de 1528 de las *Siete Partidas*, pues,

Transkribus puede entregar un fichero docx, un fichero pdf, un fichero en texto plano, pero también un fichero etiquetado en TEI; sin embargo, es un etiquetado que solo se preocupa de los aspectos estructurales básicos: encabezados, columnas, líneas, además de los referentes a las imágenes que en 7PartidasDigital no son de interés. Por eso se ha desarrollado un *script* en R[23] para transformar el fichero TEI de Transkribus y adecuarlo al modelo editorial establecido en el proyecto.

Lo importante, y con esto finalizo, es que los viejos y probadísimos métodos de la filología clásica se pueden y tienen que complementar con los métodos y herramientas digitales que tenemos al alcance de la mano. Además, no podemos contentarnos con las herramientas cerradas que nos ofrecen la mayoría de los programas, tenemos que complementarlos con el manejo de un lenguaje de programación que ayude a resolver, en segundos, problemas mecánicos y repetitivos que manualmente llevarían muchísimas horas de trabajo y con el riesgo de errores. También hay que recordar, y tener en cuenta, que debemos mantener los ficheros de las transcripciones lo más puros y limpios posible: GitHub[24] y su control de versiones es la mejor opción en la actualidad.

§

como se ha demostrado (Fradejas Rueda, en prensa), en esta edición cuidada por Francisco de Velasco, se introduce la medievalización del texto que muestra Gregorio López. Su objetivo final es determinar qué corrigió Gregorio López, qué corrigió Francisco de Velasco y hasta que punto cambia la lengua de una edición a otra.

[23] Véase: <https://github.com/7PartidasDigital/XML-TEI/blob/master/scripts/Transkribus-7PD.R> (cons. 27/10/2021).

[24] 7PartidasDigital guarda sus ficheros TEI en <https://github.com/7PartidasDigital/XML-TEI> (Fradejas Rueda, 2018).

Bibliografía citada

Admyte, *Admyte*. Madrid: Micronet, Quinto Centenario, Biblioteca Nacional, 1991.

Admyte, *Archivo digital de manuscritos y textos españoles* [ADMYTE] (1992-1998), eds. Francisco Marcos Marín, Gerardo Meiro, Charles B. Faulhaber, Ángel Gómez Moreno, Aurora Martín de Santa Olalla, Julián Martín Abad y John Nitti, Madrid, Micronet, 1992-1998, vols. 0, 1 y 2.

Alonso Rioja, Valvanera, «Herramienta para análisis filológico según el método de Lachman (AFTL)», Valladolid, Universidad, Proyecto Fin de Carrera, 1996.

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*, 9 (2020), 534-561 <https://www.janusdigital.es/articulo.htm?id=160> (cons. 29/10/2021).

Blecua, Alberto, *Manual de crítica textual*, Madrid, Castalia, 1983.

Buelow, Kenneth y David Mackenzie, *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language*, Madison, HSMS, 1977.

Cárdenas, Anthony J., John J. Nitti y Jean Gilkison Mackenzie, *Bibliography of Old Spanish Texts (literary texts)*, Madison, Hispanic Seminary of Medieval Studies, 1975.

Castrillo Benito, Nicolás, *Programas Tustep (TUebinger System von TExtverarbeitungsProgrammen): aplicación del tratamiento de textos a la investigación*, Valladolid, Universidad de Valladolid, 1992.

Castro, Américo, «La crítica filológica de los textos», *Boletín de la Institución Libre de Enseñanza*, 41, n. 682 (1917), 26-31.

—, «La crítica filológica de los textos», *Lengua, enseñanaza y literatura (esbozos)*, Madrid, Victoriano Suárez, 1924, 171-197.

Craddock, Jerry R., «La nota cronológica inserta en el prólogo de las *Siete Partidas*», *Al-Andalus*, 39 (1924), 363-390.

Dearing, Vinton A., *Methods of Textual Editing. A Paper Delivered at a Seminar on Bibliography Held at the Clark Library, 12 May, 1962*, Los Angeles, William Andrews Clark Memorial Library-University of California, 1962.

Faulhaber, Charles B., «La *Text Encoding Initiative* y su aplicación a la codificación textual y explotación», en *Actas del Congreso de la Lengua Española: Sevilla, 7 al 10 octubre, 1992*, Madrid, Instituto Cervantes, 1994, 331-340.

—, *PhiloBiblon*, Bancroft Library, University of California, Berkeley, 1997, <http://vm136.lib.berkeley.edu/BANC/philobiblon/index.html> (cons. 28.10.2021).

Faulhaber, Charles B. y Ángel Gómez Moreno, *Normas para BOOST4 (Bibliography of Old Spanish Text 4th Edition)*, Madison, HSMS, 1986.

Faulhaber, Charles, Ángel Gómez Moreno, Anthony J. Cárdenas, John J. Nitti, y Jean Gilkison Mackenzie, *Bibliography of Old Spanish Texts* (3.ª edición), Madison, Hispanic Seminary of Medieval Studies, 1984.

Fernández-Ordóñez, Inés, «Reseña a "Alberto Blecua: Manual de crítica textual, Madrid: Castalia, 1983"», *Edad de Oro*, 7 (1988), 231-240.

Foulet, Alfred y Mary Blakely Speer, *On Editing Old French Texts*, Lawrence, The Regents Press of Kansas, 1979.

Fradejas Lebrero, José y José Manuel Fradejas Rueda, *Pero López de Ayala, Libro de la caza de las aves*, Barcelona, Castalia, 2016.

Fradejas Rueda, José Manuel, «*Tratado de cetrería.* Texto, gramática y vocabulario (según el Ms. 9 de la R.A.E.)», Madrid, Universidad Complutense, 2 vols., 1983.

—, *Texto y concordancias de los textos menores del MS. V.II.19 de El Escorial: «Gerardus falconarius», «Dancus Rex», «Guillelmus falconarius», «Libro de los açores»*, Madison, HSMS, 1992a.

—, *Texto y concordancias del MS Additional 16392 de la British Library: «Libro de la caça de las aues» de Pero López de Ayala*, Madison, HSMS, 1992b.

—, *Historia de Enrique, fi de Oliva*, Madrid, Centro Virtual Cervantes, 1997 <https://cvc.cervantes.es/literatura/clasicos/fi/default.htm> (cons. 28/10/21).

—, *Textos clásicos de cetrería, montería y caza*, Madrid, Fundación Histórica Tavera, Digibis, 1999.

—, «La Codificación XML/TEI de Textos Medievales», *Memorabilia*, 12 (2010), 219-247 <http://parnaseo.uv.es/Memorabilia/Memorabilia12/PDFs/Codificacion.pdf> (cons. 28/10/21).
—, *La versión castellana medieval de la «Epitome rei militaris»*, San Millán de la Cogolla, Cilengua, 2011.
—, «Cuatro nuevos testimonios manuscritos de las *Siete Partidas*», *Revista de literatura medieval*, 27 (2015), 13-52.
—, 7PartidasDigital/XML-TEI: Primera versión (Version v.0.1), Zenodo, 2018 <http://doi.org/10.5281/zenodo.1195642> (cons. 28/10/21).
—, «Incunables de las *Siete Partidas* en Hispanoamérica» , en *Las "Siete Partidas" del Rey Sabio una aproximación desde la filología digital y material*, eds. J. M. Fradejas, E. Jerez y R. Pichel Madrid, Iberoamericana, 2021a, 175-189.
—, «La codificación TEI de las ediciones de 1491 y 1555 de las *Siete Partidas*», en *Las "Siete Partidas" del Rey Sabio: Una aproximación desde la filología digital y materia*l, eds. J. M. Fradejas Rueda, E. Jerez Cabrero y R. Pichel, Madrid, Iberoamericana, 2021b, 253-265.
—, «Las *Siete Partidas*: del pergamino a la red», en *Alfonso el Sabio y la conceptualización jurídica de la monarquía en las "Siete Partidas"*, eds. Mechthild Albert, Ulrike Becker y Elmat Schmidt, Bonn, University Press, 2021c, 223-264.
—, «Los testimonios castellanos de las *Siete Partidas*», en *Las "Siete Partidas" del Rey Sabio: Una aproximación desde la filología digital y materia*l, eds. J. M. Fradejas Rueda, E. Jerez Cabrero y R. Pichel, Madrid, Iberoamericana, 2021d, 21-35.
—, «Las *Siete Partidas*: el texto base de la edición de Gregorio López (1555)», Actas del Congreso *Historiador y Poder, el Historiador en el Poder. VIII Centenario del nacimiento de Alfonso X el Sabio*, Moscú, en prensa.
Gago Jover, Francisco y F. Javier Pueyo Mena, «El *Old Spanish Textual Archive*, diseño y desarrollo de un corpus de textos medievales: lematización y etiquetado gramatical», *Scriptum Digital*, 7 (2018a), 25-35.

—, «El *Old Spanish Textual Archive*, diseño y desarrollo de un corpus de textos medievales: el corpus textual», *Cuadernos del Instituto Historia de la Lengua*, 11 (2018b), 165-209.

—,*Old Spanish Textual Archive*. Hispanic Seminary of Medieval Studies 2020 <http://osta.oldspanishtextualarchive.org.> (cons. 29/10/21).

García Solalinde, Alfonso, *Alfonso X el Sabio, General Estoria. Primera Parte*. Madrid, Centro de Estudios Históricos, 1930.

Herrero de la Fuente, Marta, *«Alma littera»: Estudios dedicados al profesor José Manuel Ruiz Asencio*, Valladolid, Universidad, 2014.

Hockey, Susan, *A Guide to Computer Applications in the Humanities*, London, Duckworth, 1980.

—, *Electronic Texts in the Humanities: Principles and Practice*, Oxford, OUP, 2000.

Jauralde Pou, Tablo, *Manual de investigación literaria*, Madrid, Gredos, 1981.

Kasten, Lloyd A., John J. Nitti y Wilhelmina Jonxis-Henkemans, *The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio*, Madison, HSMS, 1997.

Lapesa, Rafael, *Canciller Pero López de Ayala. Rimado de Palacio: esbozo de una edición crítica*, Valencia, Biblioteca Valenciana, 2010.

Lázaro Carreter, Fernando, *Diccionario de términos filológicos*, Madrid, Gredos, 1952.

López Estada, Francisco, *Introducción a la literatura medieval española*, Madrid, Gredos, 1979.

Manley, John M. y Edith Rickert, *The Text of the Canterbury Tales*, Chicago, U. Chicago Press, 1940.

Manuscrito = *Manuscrito de Per Abbat. Cantar de Mio Cid*, Madrid, Biblioteca Nacional de España, 1998.

Marcos Marín, Francisco, *Libro de Alexandre*, Madrid, Alianza Editorial, 1987.

Marín Ocete, A., «El estado actual de la crítica de textos», *Boletín de la Universidad de Granada*, 4 (1932), 349-361.

Moorman, Charles, *Editing the Middle English Manuscript*, Jackson, U.P. of Mississippi, 1975.

Nitti, John, «Computers and the Old Spanish Dictionary», *Computers and the Humanities*, 12 (1978), 43-52.

O'Neill, John, *Electronic Text and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings*, Madison-New York, HSMS, 1999.
Onís, Federico, *Torres Villarroel. Vida*, Madrid, La Lectura, 1912.
Robinson, Peter M. W., «The Collation and Textual Criticism of Icelandic Manuscripts (l): Collation», *Literary and Linguistic Computing*, 4/2 (1989a), 99–105.
—, «The Collation and Textual Criticism of Icelandic Manuscripts (2): Textual Criticism», *Literary and Linguistic Computing*, 4/3 (1989b) 174-181.
Ruiz Albi, Irene, «Un fragmento de los *Bocados de oro* en el archivo de la Real Chancillería de Valladolid», en *«Alma littera»: Estudios dedicados al profesor José Manuel Ruiz Asencio*, ed. Herrero de la Fuente, Valladolid, Universidad, 2014, 579-593.
Ruiz, Elisa (1985), «Crítica textual, edición de textos», en *Métodos de estudio de la obra literaria*, ed. J. M.ª Díez Borque, Madrid, Taurus, 67-120.
—, *Manual de codicología*, Salamanca, Fundación Germán Sánchez Ruipérez, 1988.
Sánchez-Prieto Borja, Pedro, *Cómo editar los textos medievales. Criterios para su presentación gráfica*. Madrid, Arco/Libros, 1998.
—, *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*, San Millán de la Cogolla, Cilengua, 2011.
Santiago, Ramón, «Crítica textual y edición de textos en el *Diccionario de términos filológicos*: la primera descripción del método lachmanniano en España», en *Palabras, norma, discurso. En memoria de Fernando Lázaro Carreter*, Salamanca, Universidad, 2005, 1105-1119.
—, «Acerca de los primeros pasos de la crítica textual en la filología española: Menéndez Pidal y el Centro de Estudios Históricos», en *El legado de Ramón Menéndez Pidal (1869-1968) a principios del siglo XXI*, ed. Inés Fernández-Ordóñez, Madrid, CSIC, I, 2020, 221-256.
Simón Díaz, José, *Bibliografía de la literatura hispánica*, Madrid, CSIC, 1950-1993, 16 vols.
Wehrli, Max, *Introducción a la ciencia literaria*, Buenos Aires, Nova, 1951.
West, Martin L., *Textual Criticism and Editorial Technique Applicable to Greek and Latin Texts*, Stuttgart, B.G. Teubner, 1973.

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

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)

Stefano Bazzaco (Università di Verona)
Ana Milagros Jiménez Ruiz (Universidad de Zaragoza)
Mónica Martín Molares (Universidade da Coruña)
Ángela Torralba Ruberte (Universidad de Zaragoza)

Abstract

El trabajo presenta los recientes logros en el campo del reconocimiento de textos llevado a cabo en 2021 gracias a la colaboración entre los siguientes proyectos: Progetto Mambrino (Univ. de Verona), BIDISO (Univ. de A Coruña) y COMEDIC (Univ. de Zaragoza). En concreto, en la primera parte del artículo se describe el estado de la cuestión de los sistemas de transcripción automática en relación con los textos impresos de la Edad Moderna, se relatan las primeras experiencias llevadas a cabo con la plataforma Transkribus (READ Coop) y los resultados preliminares obtenidos. En la segunda parte se presentan dos modelos de HTR que consienten la transcripción automática de textos en letra gótica y redonda de la Edad Moderna (siglos XV-XVII). En dos apéndices finales se describen según las normas tipobibliográficas actuales los documentos empleados para la creación de ambos modelos.
Palabras clave: Humanidades Digitales; Transkribus (READ Coop); HTR (Handwritten Text Recognition); impresos de la Edad Moderna; Siglos de Oro

The work presents the recent achievements in the field of text recognition carried out in 2021 thanks to the collaboration between the following projects: Progetto Mambrino (Univ. of Verona), BIDISO (Univ. of A Coruña) and COMEDIC (Univ. of Zaragoza). Specifically, the first part of the article describes the state of the art of automatic transcription systems in relation to the recognition of printed texts of the Modern Age, the first experiences carried out with the Transkribus platform (READ Coop) and the preliminary results obtained. In the second part, we present two HTR models that allow the automatic transcription of early printed texts in gothic and round scripts of the Modern Age (15th-17th centuries). In two final appendices, the documents used for the creation of both models are described according to current typobibliographical standards.
Keywords: Digital Humanities; Transkribus (READ Coop); HTR (Handwritten Text Recognition); early printed documents; Siglos de Oro

Stefano Bazzaco, Ana Milagros Jiménez Ruiz, Mónica Martín Molares, Ángela Torralba Ruberte, «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 (2022) Humanidades Digitales y estudios literarios hispánicos, pp. 67-125.
DOI: https://doi.org/10.13136/2284-2667/1190 - ISSN: 2284-2667.

Premisa[*]

En el primer capítulo del libro *From Gutenberg to Google*, Peter Shillingsburg, uno de los precursores y más importantes estudiosos del fenómeno de la migración de textos a un entorno digital, deja constancia de un problema que afecta a cualquier transmisión de prácticas de escritura en la web: la falta de fiabilidad de los contenidos informativos que se encuentran en línea. En esas provechosas páginas, de sumo interés para cualquier especialista en Humanidades, se hallan claras evidencias de un constante choque entre la tensión del filólogo hacia la reconstrucción exhaustiva del texto como acto informativo y, por otra parte, el aspecto que adquiere un texto, cualquiera que sea su forma de fijación, dentro del heterogéneo y volátil espacio del *World Wide Web*.

Para abordar la cuestión, el autor considera que cualquier editor tiene una responsabilidad compleja, fundada en la necesidad de declarar todos los aspectos relativos al trabajo editorial y a la interacción del nuevo texto con sus precedentes evolutivos, es decir, sus múltiples concreciones físicas a lo largo del tiempo. Por esta razón, la nueva edición debería guardar información fiable con respecto a dónde ha sido encontrado, qué procedimientos editoriales se le aplicaron, cuáles son las diferencias entre la obra editada y los materiales fuente, teniendo en cuenta constantemente que ningún acto editorial es una operación neutral, sino deliberadamente electiva (2006, 19 y ss.).

En opinión de Schillingsburg, todas estas cuestiones adquieren aún más importancia en la época digital, inicialmente gobernada por un clima de entusiasmo general que ha ocultado en parte las fisuras intrínsecas del proceso de migración de los textos a la red. Al respecto, según el estudioso,

[*] El presente trabajo se desarrolla en el marco de los siguientes proyectos: *Proyecto PRIN 2017 Mapping Chivalry. Spanish Romances of Chivalry from Renaissance to 21st Century: a Digital Approach* (2017JA5XAR), investigador principal Anna Bognolo, Università di Verona (2017-2023); Progetto di Eccellenza «Le Digital Humanities applicate alle lingue e letterature straniere», Università di Verona (2018-2022); Proyecto de Investigación *Catálogo de Obras Medievales Impresas en Castellano (COMEDIC)*, PID2019-104989GB-I00, financiado por MCIN/AEI/10.13039/501100011033, que se inscribe en el grupo investigador Clarisel y cuenta con la participación económica del Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón; proyecto de I+D+i *Biblioteca Digital Siglo de Oro 6* (BIDISO 6), referencia: PID2019-105673GB-I00 financiado por MCIN/AEI/10.13039/501100011033/.

los resultados obtenidos a principios del nuevo milenio por el Proyecto Gutenberg demuestran una inconsistencia de fondo, tanto que le llevan a preguntarse:

> does anyone believe that a Project Gutenberg electronic text could be relied upon to be accurate? Do these productions state accurately what the source text was? Do they describe the bibliographic features of the source text? Did the 'editors' pick as a source text one that has any sort of authority or historical importance? Did they indicate in any way how the editing or transcribing of scanning involved changed the text? (2006, 21).

Constatando que los textos del Proyecto Gutenberg que ha consultado son inservibles para su trabajo filológico, Schillingsburg declara que el trasvase de un documento a un entorno digital conlleva las mismas responsabilidades que afectan a su edición en formato analógico, como tener conciencia del texto que se está editando, de los métodos que se emplean para su transmisión y de los diferentes estadios que llevan a su nueva representación, como la colación y el *proofreading*. En ausencia de tales responsabilidades, contaríamos tan solo con ediciones imprecisas y poco útiles para cualquier propósito de investigación académica (2006, 22)[1].

Es evidente que, si extendemos la mirada a proyectos afines al Project Gutenberg, los mismos inconvenientes afectan a la mayoría de los repositorios textuales de largo alcance presentes en la red, puesto que en ellos la reproducción de obras literarias está sometida generalmente a dudosas prácticas ecdóticas. Como resultado, Internet ha llegado a configurarse como un cosmos interminable de *fast food libraries*, donde las ediciones correctas, preciadas como joyas, son muy difíciles de detectar y consultar (Italia, 2020, 38-39).

[1] «My only point at this time is that bringing a text from an old book or manuscript into the twenty-first century takes more than a computer – with or without a scanner or digital camera. It takes thoughtfulness about text, an exercise of care and good judgment about methods, and an old-fashioned devotion to sight collation and proofreading that tends to dampen enthusiasm. In the absence of these responsibilities, what will we have? Noisy texts, without any doubt. Misleading texts, very likely. Text useless for scholarly purposes, of course» (2006, 22). Estas mismas preocupaciones se encuentran tratadas de forma más general en relación con el concepto de «infoesfera» en Floridi (2014).

Son muchas las cuestiones interesantes sugeridas por Schillingsburg. Por un lado, sus reflexiones abarcan aspectos todavía por resolver, como el papel del filólogo dentro del contexto de la migración masiva de documentos que se publicaron impresos o manuscritos a la red. Por otro, sus palabras apuntan a la necesidad de rehabilitar el espacio virtual con la recreación de materiales textuales fiables, correctos y bien identificables por medio de metadatos que informen sobre los actores del proceso editorial y el texto proporcionado, por ejemplo, indicando el nombre del autor, el nombre del editor, cuál es el ejemplar transcrito, etc.

Es interesante que, con respecto a las técnicas que determinan la forma del objeto digitalizado y que han influido en su escasa fiabilidad hasta ahora, Schillingsburg insista en la exigencia de declarar cuál es la tecnología que ha sido empleada para su digitalización[2]. Al respecto, es improbable que el estudioso americano se refiera aquí a la conversión en formato imagen de los libros, ya que en sí no puede representar una fuente de riesgo: por supuesto, las primeras experimentaciones en este campo llevaron a la proliferación de imágenes facsímil de mala calidad, pero sería algo forzado sugerir que esto sea el problema principal que afecta a los textos en la web. Por otra parte, parece más lógico pensar que Shillingsburg se refiera aquí a los inconvenientes derivados de la transcripción masiva con sistemas defectuosos de reconocimiento de textos, una tendencia que ha llevado a resultados nefastos en varios ámbitos de la edición digital.

Esta desconfianza hacia la transcripción automatizada no sorprende: la crítica ha sugerido en varias ocasiones que la dificultad en localizar ediciones literarias fiables en la red procede en gran medida del uso impropio que se hace de las herramientas de transcripción automática. En efecto, los avances en este campo, que en el siglo pasado contaba solamente con los sistemas de reconocimiento óptico de caracteres (*Optical Character Recognition*, de aquí en adelante OCR) y que ahora ha ido enriqueciéndose con nuevos sistemas de reconocimiento de textos manuscritos basados en redes neuronales (*Handwritten Text Recognition*, de

[2] Entendemos aquí digitalización en su sentido extenso como acto de conversión de un objeto analógico al entorno digital.

aquí en adelante HTR), han permitido una migración masiva del patrimonio textual analógico a la web; pero ¿de qué manera se dio la colonización del espacio digital por medio de estas herramientas?

A pesar de que existen estudios que aseguran cierta fiabilidad de los textos OCRizados presentes en la red (por ejemplo, Kichuk, 2019)[3], es incuestionable que la conversión de los libros antiguos en objetos digitales ha dado pie a una proliferación de documentos electrónicos de dudosa calidad, de los cuales desconocemos con frecuencia características fundamentales como la procedencia, las fuentes y las normas de transcripción adoptadas. En otras palabras, ha ido consolidándose en varios contextos la tendencia a emplear herramientas de transcripción automatizada de forma no supervisada con resultados preocupantes, tanto en la publicación de *ebooks* y ediciones comerciales, que se distribuyen repletos de errores, como en la indexación de los objetos informativos, que cuentan frecuentemente con unos metadatos derivados de un proceso de extracción con OCR descontrolado y asistemático. A todo ello, hay que añadir que la figura del humanista, inicialmente escéptico hacia los resultados generados por los sistemas de transcripción automatizada, ha quedado inevitablemente fuera de la ecuación, hasta ser un mero intérprete y no un actor del proceso de migración digital de los textos, que siguen multiplicándose en la web de modo desordenado y exponencial.

El presente trabajo trata la posibilidad de invertir esta tendencia y ofrecer al humanista un punto de acceso válido para la difusión masiva de textos fiables en la red, jugando en el mismo campo tecnológico que ha favorecido su exclusión del proceso, es decir, el de la transcripción automatizada.

En particular, en estas páginas se presentan los primeros resultados obtenidos por un proyecto colaborativo de transcripción de impresos hispánicos instituido en 2021 por medio de la colaboración entre los siguientes proyectos de investigación: Progetto Mambrino (Università di Verona); BIDISO (Universidade da Coruña); COMEDIC (Universidad de Zaragoza).

En la primera parte del artículo, se ofrece una introducción a los

[3] Sin embargo, hay que reparar en el hecho de que el estudio de Kichuk se ocupa de la distribución masiva de textos en la red para fines no científicos.

sistemas de reconocimiento de caracteres, destacando los principales aspectos relativos a su evolución, su estado actual y los desafíos futuros que suponen con respecto al estudio de los impresos antiguos. La segunda parte está dedicada a la descripción del proyecto de colaboración, del cual se detallan los objetivos y los primeros logros en el ámbito de la transcripción automatizada de impresos de la Edad Moderna. En concreto, se consideran los siguientes aspectos relativos al proyecto: la publicación de dos modelos de transcripción automatizada para los impresos hispánicos en gótica y redonda, estrenados a finales de 2021 y disponibles en acceso abierto dentro de la plataforma Transkribus (READ Coop SCE); las posibilidades de explotación de las transcripciones obtenidas; los principales canales de difusión de los resultados del proyecto; la progresiva alimentación de los modelos de reconocimiento publicados.

La sección conclusiva está dedicada a la presentación de los dos modelos extendidos de HTR *SpanishGothic* (Apéndice 1) y *SpanishRedonda* (Apéndice 2). De ambos se ofrece una ficha de síntesis que proporciona información acerca de las características principales, las instituciones académicas y los estudiosos participantes en su creación, la última versión publicada del modelo y las indicaciones para citarla. Sigue una descripción detallada de las obras que se emplearon para el entrenamiento de la máquina y que constituyen el *dataset* para la creación de los dos modelos: de cada obra se proporcionan las principales informaciones gráficas y bibliográficas con la intención de asistir a los especialistas en la comprensión, utilización y alimentación de los recursos descritos[4].

[4] La *Premisa*, los epígrafes 1 y 2 y las *Conclusiones* son obra de Stefano Bazzaco (Univ. di Verona); el *Apéndice 1* de Ana Milagros Jiménez Ruiz y Ángela Torralba Ruberte (Univ. de Zaragoza); el *Apéndice 2* de Mónica Martín Molares (UDC).

1. Los sistemas de reconocimiento de textos y los impresos antiguos: estado de la cuestión

1.1. Breve historia del reconocimiento de textos

La historia de los sistemas de reconocimiento de textos es amplia y está marcada por precisos saltos tecnológicos. Los estudiosos coinciden en que los primeros pasos en este campo se deben buscar en la invención del *Retina scanner*, allá por 1870, por parte de Carey, es decir, en la creación de un dispositivo que permitía la lectura de imágenes simulando la acción del ojo humano. Este aparato estimuló la experimentación en el campo del reconocimiento automático, lo que produjo la creación de medios electrónicos que pudiesen sustentar la lectura de los ciegos. A partir de él, rápidamente aparecieron el Optófono, un aparato ideado por Fournier d'Albe en la primera década del siglo XX capaz de vocalizar las palabras impresas en una pantalla, y la *Reading Machine* de Tauschek, estrenada por primera vez en 1928 y considerada a todos los efectos el antepasado de los actuales sistemas de transcripción automatizada. Se trata, en efecto, de un aparato que busca la coincidencia entre caracteres impresos y unas representaciones modélicas de los mismos colocadas en un disco rodante: al encontrar una correspondencia, la máquina imprime ese carácter en una nueva hoja y continúa con el reconocimiento del carácter sucesivo.

A partir de este momento, la línea evolutiva de los sistemas de reconocimiento de textos coincide esencialmente con la historia de las herramientas de OCR. En concreto, hablamos de sistemas de OCR en sentido estricto solamente a partir de los años 50, cuando este campo empieza a relacionarse con los intereses de empresas comerciales que por primera vez vislumbraron la posibilidad de ejercer un control generalizado sobre enormes cantidades de datos textuales. La primera generación de *hardware* OCR, es decir, de artefactos físicos que consentían transcribir de forma automatizada documentos impresos, como el IBM 1418[5], se produciría solamente en la década posterior, con la creación de unos

[5] Producido por la Endicott, este dispositivo se lanzó el 12 de septiembre de 1960. Para más detalles, remitimos a la web de IBM, y en especial al siguiente enlace: <https://www.ibm.com/ibm/history/exhibits/endicott/endicott_chronology1960.html> (cons. 21/02/2022).

prototipos con funcionalidades muy limitadas porque consentían la interpretación de un *set* concreto de letras (Narang *et al*., 2020, 5119-5121). En este período, paralelamente, aparecían también unos especiales *typefaces* conocidos como OCR A y OCR B, que eran sistemas gráficos específicamente dibujados para ser interpretados por medio de las tecnologías muy limitadas del período y que constituyeron un primer avance en el área de la transcripción no supervisada.

Llegamos pues a los años 70, un momento determinante para la evolución de los sistemas de reconocimiento de texto porque, junto con la aparición de una segunda generación de *hardwares* OCR que consentían la transcripción de documentos *multifont*, es decir, de mpresos que mezclaban distintos sistemas gráficos, se registran los primeros atisbos de una evolución en el campo del reconocimiento de textos manuscritos. En principio, la máquina pudo interpretar solamente números o letras aisladas y poco complejas como los códigos postales. Necesariamente, debemos fijar aquí el nacimiento de esta subárea del reconocimiento de textos, cuya historia en parte sigue solapándose con la de los OCR hasta adquirir en años recientes un estatuto propio y registrar una creciente y rápida expansión.

Con la reducción del coste del *hardware* y la consiguiente distribución de los *personal computers*, asistimos a la aparición de los primeros paquetes *software* de OCR, que constituyen un verdadero avance porque remiten el problema de la transcripción automatizada a la comunidad de usuarios. Como acaece con frecuencia en el contexto del desarrollo de herramientas digitales, el hecho de que los usuarios tengan acceso a una nueva tecnología constituye un punto de inflexión en la evolución de la misma, puesto que ejercicios e intuiciones particulares son el motor de nuevas experimentaciones. Como directa consecuencia, desde la mitad de los 70, las prestaciones de los sistemas de OCR se incrementan de forma notable: los *softwares* de reconocimiento de textos, sobre todo impresos, llegan a descifrar conjuntos de caracteres muy distintos, mientras que paralelamente se empieza a prestar atención a la interpretación de documentos complejos, por ejemplo, los textos multilengua.

La última etapa evolutiva de los *softwares* de reconocimiento, que podríamos colocar desde los primeros años del 2000 a la época actual, está

caracterizada por los avances más notables. Al mismo tiempo que aparecen proyectos de digitalización de largo alcance y se consolidan los intereses de grandes empresas privadas (Terras, 2010), los sistemas de transcripción automatizada siguen perfeccionándose, sustentados por la ilusión de que en un futuro ya próximo se llegue a transformar todo el patrimonio textual analógico en texto electrónico que la máquina pueda medir y manejar. Con la introducción de nuevos procedimientos de la inteligencia artificial basados en arquitecturas *Long Short Term Memory* (LSTM), como el *deep learning* y las redes neurales, los *softwares* de reconocimiento llegan a interpretar distintas grafías cada vez más complejas (*complex scripts*), incluso textos no occidentales, impresos antiguos y documentos manuscritos.

Sin embargo, si por un lado podríamos decir que el reconocimiento de textos impresos de la actualidad (posteriores a 1930) se considera un problema solucionado, con los impresos antiguos y los manuscritos la situación está lejos de resolverse. Los resultados más alentadores se están dando solo en los últimos años, con el florilegio de plataformas de transcripción automatizada de HTR que constituyen una verdadera revolución para convertir los contenidos textuales de bibliotecas y archivos al espacio virtual de la web. Es interesante notar cómo el nombre de estas herramientas guarda un cambio notable: el interés ha pasado de la interpretación de caracteres aislados (*Optical Character Recognition*) a la búsqueda de *patterns* recurrentes en porciones o líneas de texto (*Handwritten Text Recognition*), prometiendo resultados de reconocimiento que hace 10 años no habríamos podido ni imaginar.

1.2. Los sistemas de OCR/HTR y los estudios humanísticos

Describir la relación entre los humanistas y los sistemas de reconocimiento de texto equivale a trazar una historia de promesas desatendidas.

En la época del desarrollo de los primeros *softwares* de OCR que, como vimos, coincidió con la eclosión de grandes proyectos de digitalización del patrimonio textual, los sistemas de reconocimiento

fueron bien aceptados por parte de la comunidad científica, con filólogos y expertos de documentación a la cabeza; sin embargo, a la luz de unos primeros resultados no propiamente significativos, la ilusión pronto se convirtió en frustración. De hecho, los humanistas empezaron a percibir los sistemas de reconocimiento como instrumentos no fiables para la investigación porque proporcionaban transcripciones repletas de errores, lo cual tuvo como consecuencia una neta distinción entre *clean transcription*, es decir, la transcripción manual realizada con métodos tradicionales, y *dirty OCR*, o sea, los textos generados de forma automatizada.

La crítica subraya cómo este prejuicio está en la base de una renuncia sustancial a la utilización de los sistemas de reconocimiento de textos por parte de los humanistas (Smith-Cordell, 2018, 10-11). En efecto, se trató de una sospecha de fondo difícil de extirpar que persistió hacia los años iniciales del nuevo milenio, a pesar de que en el ámbito informático se dieran progresivos avances tecnológicos en este campo de investigación.

Una primera vuelta a los sistemas de reconocimiento en el ámbito humanístico se dio solamente 20 años más tarde por medio de la creación de grandes repositorios de textos digitalizados en formato imagen, principalmente en el ámbito del proyecto *Google Book Search*, lanzado por la empresa de Mountain View en 2004 con ocasión de la Feria del libro de Frankfurt[6]. Se trata de una forma mediada porque en esta ocasión se emplearon sistemas de OCR más competitivos por parte de los técnicos de Google, sobre todo derivados de un constante refinamiento de los resultados obtenidos con la plataforma de acceso abierto *Tesseract* (desarrollada por Hewlett-Packard entre 1984-94). Sin embargo, la aplicación de esta herramienta se limitaba a la disposición de un estrato OCR oculto, favorable para la búsqueda de palabras clave internas al repositorio. De este modo, el humanista podía tranquilamente desconocer de dónde procedían los resultados de búsqueda, pero implícitamente en sus investigaciones documentales estaba ya sirviéndose de un sistema de reconocimiento de textos, y, quizás movido por los fascinantes resultados que obtenía de la explotación de materiales recolectados en formato digital, parecía también olvidarse de su escasa fiabilidad.

[6] Para una historia del proyecto lanzado por Google remito a los imprescindibles trabajos de Roncaglia (2009; 2010).

Fortalecidos por el interés de perfeccionar las funciones de búsqueda de palabras clave, los sistemas de transcripción automatizada pudieron entonces pasar por una rehabilitación, convirtiéndose en un verdadero punto de referencia para los estudiosos del texto que experimentaban en esos mismos años una inédita atracción por los trabajos de análisis cuantitativo (Moretti, 2005; 2022). Los tiempos eran maduros para la remoción del prejuicio inicial y la vuelta a la experimentación en el campo de la transcripción automatizada con la idea de que pudiera asegurar logros de sumo relieve en distintos campos de las Humanidades. El panorama reciente de los sistemas de OCR/HTR, que sigue enriqueciéndose día a día, es justamente el resultado del cambio de percepción que acabamos de señalar y de un sorprendente florecimiento tecnológico posterior, capaz de sustentar nuevos proyectos de digitalización como el que describimos en estas páginas.

1.3. La transcripción automática de impresos antiguos: estado de la cuestión

Trazar un estado de la cuestión de lo que ha llegado a ser en la actualidad el reconocimiento de textos impresos no es una tarea simple. En general, esto se debe a dos clases de problema.

De entrada, el primer asunto es que la bibliografía relacionada generalmente con este ámbito de estudio es de naturaleza muy variada. De hecho, los estudios sobre OCR/HTR tratan distintas áreas del conocimiento, que van desde la informática pura hasta las ciencias de la documentación y los estudios históricos y literarios. Por consiguiente, la producción de artículos referidos a este campo de indagación es bastante heterogénea: hay artículos técnicos, que relatan el desarrollo de una tecnología determinada en los campos del preprocesamiento de imágenes, la segmentación de imágenes (o *Layout Analysis*), el reconocimiento de documentos complejos (textos no occidentales o multilingües); artículos científicos que tratan casos de aplicación de herramientas de transcripción automática a un corpus de estudio concreto dentro de proyectos editoriales de largo alcance; artículos de carácter más generalista que dan cuenta de proyectos locales de digitalización y de pequeños resultados

obtenidos en una esfera de aplicación muy limitada. Evidentemente, encontrar referencias concretas dentro de un conjunto de estudios tan amplio implica ciertas complicaciones.

En segundo lugar, son pocos los trabajos de investigación que intentan ofrecer una mirada más extensa, que abarque los últimos quince años de actividades en el campo del reconocimiento de textos impresos. Los estudios más sugerentes al respecto vienen de humanistas digitales que, guiados por el objetivo de tratar un caso de estudio concreto, se dedican a reconstruir parte de la tradición bibliográfica relativa a las herramientas digitales adoptadas. Sin embargo, en muchas ocasiones, estos trabajos tienen una finalidad específica, acabando por ofrecer una panorámica muy limitada acerca de otras subáreas de investigación. El ejemplo más relevante al respecto son los artículos producidos por el grupo de investigación alemán que se ha formado bajo el magisterio de Christian Reul y Uwe Springmann en Würzburg[7]. Estos trabajos, a pesar de fundarse en una perspectiva crítica adecuada, acaban por centrarse únicamente en el reconocimiento de textos impresos en *Fraktur*, una grafía empleada por los periódicos alemanes a principios del siglo XX, y no plantean una visión de conjunto.

Para encontrar una investigación que intente tratar de forma exhaustiva el reconocimiento de documentos impresos hay que volver al imprescindible volumen *Electronic Textual Editing* de 2006, editado por Burnard, O'Keeffe y Unsworth con el patrocinio de la MLA (*Modern Language Association*). Efectivamente, la obra, que constituye un hito fundamental dentro de los estudios de Humanidades Digitales, contiene una sección «Practices and procedures», donde aparece un artículo de Gifford Fenton y Duggan (2006, 241-253) que intenta trazar un cuadro general de lo que ha llegado a ser en ese momento histórico el reconocimiento de textos en relación con la filología de los documentos manuscritos e impresos. Este artículo nos servirá de guía para resaltar los avances que se han dado recientemente en este campo.

En la introducción, la autora presenta su experiencia acerca de la

[7] A este grupo se debe el desarrollo de *software* de reconocimiento de textos como OCRopy/OCRopus, Calamari, OCR4All. Para un listado de las publicaciones remito a la bibliografía que se encuentra en Reul *et al.* (2018, 6).

conversión de documentos impresos del repositorio JSTOR, mientras en las siguientes secciones principales aborda los fundamentos del procesamiento con OCR y segmentación automática (*Layout Analysis* o *Zoning*), las características de las fuentes impresas y errores más frecuentes, y los factores decisivos en la adopción de sistemas de OCR en un proyecto de investigación.

Al tratar los fundamentos del procesamiento de textos con sistemas de OCR, Gifford Fenton aclara que la transcripción automática de documentos impresos estaba prometiendo significativos avances, hasta convertirse en lugar común para proyectos de digitalización de largo alcance. La estudiosa considera entonces el flujo de trabajo tradicional de cualquier sistema de OCR, centrándose en tres asuntos principales: la digitalización en formato imagen, los problemas derivados de la segmentación no supervisada y la detección de un orden de lectura de las zonas segmentadas.

Con respecto a las imágenes digitalizadas, se puede apreciar cómo ya en esa época los estudiosos estaban convencidos de que los resultados del reconocimiento con OCR dependían en gran medida de la calidad de los materiales escaneados. En la época actual, gracias a los avances que se han dado en la gestión de imágenes, que en su mayoría se difunden en formatos de alta calidad que alcanzan por lo menos los 300 dpi, también las prestaciones de los sistemas de transcripción automática se han intensificado. Por otra parte, si se considera la segmentación e interpretación de la página digitalizada, notamos que no todos los problemas se han resuelto. Gifford Fenton señala la cuestión de la siguiente manera: «while [...] zoning task would be simple for a human reader with the ability to interpret semantic clues, it can present a variety of challenges for a machine» (2006, 247-248). Es por ello por lo que, a pesar de contar en época reciente con la aplicación de redes neuronales convolucionales (típicas del entrenamiento profundo), el campo sigue presentando inconvenientes. Esto es evidente para cualquier estudioso que se acerque a las herramientas de OCR/HTR disponibles en la actualidad, puesto que las manchas y los desgastes presentes en la fuente digitalizada siguen siendo descifrados por parte del ordenador como porciones de texto y las zonas segmentadas no siempre son interpretadas

siguiendo el orden de lectura correcto.

Al respecto, los avances en el campo de la inteligencia artificial más interesantes residen en la posibilidad de entrenar la máquina sobre modelos de *layouts* preconcebidos, lo cual proporciona resultados cada vez más fiables; pero la impresión general es que faltan todavía unos años para que la cuestión esté resuelta de forma definitiva.

Al analizar las causas de errores más frecuentes en el contexto de la transcripción automatizada, Gifford Fenton destaca varios elementos de las fuentes empleadas que inciden de forma negativa en el reconocimiento: debilitamiento de la tinta, tamaño de las letras, elementos gráficos (a veces ubicados en transparencia bajo el texto), columnas de texto muy pegadas. Sin embargo, contrariamente a lo que se señaló en el caso de la segmentación de la página digitalizada, se trata de problemas resueltos en la época actual. Por lo que atañe a la presencia de elementos gráficos, en el caso de los impresos antiguos normalmente es preferible contar con imágenes de las fuentes en colores. La razón reside en que la detección del *layout* de la página está basada en la densidad de los píxeles y, por tanto, es más probable que los elementos gráficos que no son de interés y que constituyen parte del ruido que obstaculiza el reconocimiento sean excluidos del proceso. El tamaño de las letras y la cercanía de las columnas, por otra parte, no ocasionan problemas relevantes.

Finalmente, al tratar la adopción de sistemas de reconocimiento de textos dentro de proyectos de Humanidades, Gifford Fenton se centra en ocho factores decisivos: seleccionar un enfoque para promover los objetivos del proyecto; definir las características de los materiales fuente; adoptar medidas de control de calidad de los textos transcritos; definir la extensión del proyecto (*scalability*); externalizar procedimientos a un sujeto tercero; considerar el gasto de tiempo que conlleva la producción de texto electrónico; considerar la duración total del proyecto; y poner atención en los costes.

Evidentemente, muchos de estos asuntos son de interés, pero no aplicables al contexto contemporáneo. Con el reciente desarrollo de herramientas de OCR/HTR de acceso abierto, los costes de la transcripción automática no parecen ser una cuestión determinante a la hora de planear un proyecto digital; lo que sí es fundamental es el control

final de los textos transcritos. El argumento de interés entre los señalados por Gifford Fenton es el tercero, que trata el control de la calidad, es decir, las operaciones posteriores al reconocimiento. Al respecto, se están llevando a cabo importantes experimentaciones en distintas áreas de las Humanidades Digitales. En el campo de la automatización se están integrando diccionarios y sistemas de procesamiento del lenguaje natural que agilizan el trabajo del filólogo que revisa los textos. Por otro lado, en el ámbito de la colaboración, se están promoviendo recientemente planes de transcripción en grupo, como el *Transcribathon* de Europeana <https://www.transcribathon.com/en/> (cons. 15/05/2022) y la revisión múltiple en *crowdsourcing*. Se trata de unas prácticas de largo alcance que miran por la formación de estudiantes y colaboradores para la producción de datos fiables y certificados, generalmente realizados bajo la supervisión de especialistas de la materia que validan el trabajo y gestionan todo el flujo de producción. El resultado es que de estos proyectos pronto se obtendrá una cantidad ingente de transcripciones certificadas, que podrán constituir el entrenamiento de *softwares* de reconocimiento de textos progresivamente más precisos y fiables.

1.4. De la función instrumental a la revolución heurística: ¿cuál puede ser el futuro de los sistemas de reconocimiento de textos?

Si miramos retrospectivamente a la evolución de los sistemas de reconocimiento de textos y partimos de una mirada menos ilusionada de la que tenían los humanistas de los 80, más que de sueños frustrados parece que deberíamos hablar de una actitud sustentada por premisas erróneas y caracterizada por cierta impaciencia. Es habitual que, para adaptarse a las metodologías tradicionales de las disciplinas humanísticas, una nueva tecnología digital necesite en principio un tiempo de acomodamiento y de reflexión crítica que regularice su utilización indiscriminada (Orlandi, 1994, 7 y ss.)[8].

[8] Al respecto, propongo retomar las palabras de Tito Orlandi, padre fundador de la escuela romana de Humanidades Digitales, quien afirma: «È accaduto a varie riprese che siano state prodotte delle macchine, nuove e potenti, ma con scopi limitati e praticamente semplici, e che soltanto dopo esse

Por ejemplo, esto pasó con la codificación en lenguaje XML de las ediciones académicas digitales, donde el modelo proporcionado por la TEI (*Text Encoding Initiative*) se asentó como estándar solamente después de un tiempo muy largo de reflexión crítica que abarcó las últimas dos décadas del siglo XX. A pesar de que ahora está padeciendo legítimas críticas por parte de algunos detractores, no hay duda de que se trata de un estándar altamente productivo, que sigue representando un punto de referencia esencial en el campo de la publicación digital por haber sido contextualizado y variadamente explotado por tantos estudiosos de Humanidades desde su aparición.

Lo mismo está ocurriendo con los estudios cuantitativos de la literatura, que ahora, quince años después de la publicación de los primeros trabajos de lectura distante realizados por Moretti (2005), están demostrando su solidez y fertilidad, abriendo la vía a nuevos caminos interpretativos, sobre todo para la estilometría (Hernández Lorenzo, 2019; Calvo-Tello, 2021) y, dentro de ella, la atribución de autoría (García-Reidy, 2019; Bazzaco, 2022).

No se dieron las mismas condiciones en el campo del reconocimiento de textos. Quizás se deba a un entusiasmo inicial por parte de los humanistas, que veían en la nueva tecnología la posibilidad de simplificar trabajos engorrosos como la transcripción manual y la metadatación, y a una sucesiva frustración derivada del empleo de una tecnología, en esos años poco madura, que cometía muchos errores. La consecuencia natural de esta situación fue el prematuro abandono de las experimentaciones de transcripción automatizada de manuscritos e impresos antiguos, calificándolas como poco rentables para el filólogo. Sin embargo, si consideramos que recientemente el campo del reconocimiento de textos ha experimentado unos notables avances, ¿podemos rectificar el prejuicio que caracterizaba esta área de estudio de las Humanidades Digitales? o, en otras palabras, ¿es legítimo suponer que la adopción de sistemas de reconocimiento de textos podría constituir un recurso de interés para la reorganización del trabajo filológico? y, junto a

abbiano generato una riflessione teorica, che dunque ha seguito e non preceduto l'innovazione tecnologica. […] Tuttavia sono state proprio le riflessioni teoriche che hanno mostrato il vero significato dei risultati che si potevano ottenere con le macchine» (1994, 8).

ello, ¿podemos imaginar posibles aplicaciones de herramientas de transcripción automática para acrecentar la presencia de documentos fiables en la red, invirtiendo la tendencia actual?

Si para contestar a estas preguntas nos limitamos a una percepción que se consolidó hace más de 30 años, corremos el riesgo de valorar de forma errónea el problema. Los sistemas de reconocimiento de textos, en efecto, han llegado en tiempos recientes a prometer unos resultados de transcripción automatizada muy alentadores, que se acercan al 1% de error para los impresos antiguos y al 5% para los manuscritos: se hace necesaria una rehabilitación de este campo de trabajo que, en la estela de los avances tecnológicos, enseñe los nuevos caminos que pueden abrirse para la labor filológica.

Originariamente el reconocimiento de textos se fundamenta en la idea de aligerar y acelerar de forma notable el proceso de transcripción manual. Ya señalamos en otra ocasión (Bazzaco, 2020, 539 y ss.) cómo esta función, que llamaríamos de tipo *instrumental*, estaba ya en la base de los planes de digitalización del patrimonio cultural que surgieron durante los años 80 del siglo XX. Melissa Terras (2010) recuerda cómo los primeros proyectos de escaneo de fuentes documentales tenían como premisa la transformación de los archivos de imágenes en un texto electrónico; lo cual nos lleva a pensar que el mismo proceso de digitalización estaba en su nacimiento íntimamente relacionado con el desarrollo de sistemas de OCR fiables que permitiesen volcar los datos textuales extraídos en un formato que la máquina pudiera medir y manejar (*machine readable form*).

Es cierto que esta es la función primaria que le asignamos a los sistemas de transcripción automatizada: prometer una rápida transformación de las imágenes digitales en textos electrónicos explotables, que pueden ser reutilizados después en varios campos de la investigación en Humanidades, como la edición digital, la extracción de corpora y lemas, los procedimientos de análisis cuantitativo, etc.

Proyectos de Humanidades Digitales que emplean sistemas de transcripción automática de forma masiva son, por ejemplo,

TranscribeBentham (2013-2017)[9], TrAIN (*Tracing Authorship In Noise*, 2018)[10], *Entangled Histories: Ordinances of the Low Countries*[11], CREMMA (*Consortium pour la Reconnaissance d'Écriture Manuscrite des Matériaux Anciens*, 2020-2021)[12], o, en relación con los estudios hispánicos, el prestigioso proyecto ETSO (Estilometría aplicada al Teatro del Siglo de Oro, 2018) dirigido por Germán Vega y Álvaro Cuéllar[13]. En todos estos casos, sin embargo, el reconocimiento de textos es visto como un medio para apresurar y aligerar el trabajo manual, no como un momento fundamental dentro del flujo ecdótico y editorial.

Aun considerando que la función instrumental sigue siendo la razón esencial para la adopción de sistemas de OCR/HTR en el interior de un proyecto de edición de textos, es necesario avanzar en la reflexión teórica y apuntar los posibles caminos para el futuro empleo de estas herramientas que aclaren hasta dónde se puede llegar.

Como punto de partida, conservan su validez las reflexiones de Raul Mordenti sobre el procedimiento ecdótico en ambiente informático. El estudioso, en efecto, habla en su obra de transcripción, que propone considerar como un momento más del acto de codificación del documento, o sea el primer eslabón de un proceso editorial que pretende fijar gráficamente el texto; si por un lado la maquetación del texto electrónico es «momento crucialissimo» porque corresponde a la «immissione nella macchina dell'informazione da cui dipenderanno tutti i successivi trattamenti e manipolazioni», por otra parte se califica la transcripción como «momento forte», «decisivo» porque corresponde al procedimiento «più costoso in termini di tempo/uomo», capaz de configurar todos los pasos posteriores (2001, 29). Cuando transcribimos en formato máquina un texto estamos en el nivel de una primera codificación, es decir, la transformación del texto contenido en los

[9] Transcribe Bentham <https://blogs.ucl.ac.uk/transcribe-bentham/> (cons. 15/05/2022). Véase Causer-Terras (2014).
[10] TrAIN <http://www.etrap.eu/research/tracing-authorship-in-noise-train/> (cons. 15/05/2022). Véase Franzini *et al.* (2018).
[11] Entangles Histories <https://lab.kb.nl/dataset/entangled-histories-ordinances-low-countries> (cons. 15/05/2022).
[12] CREMMA <https://www.dim-map.fr/projets-soutenus/cremma/> (cons. 15/05/2022).
[13] ETSO <https://etso.es/> (cons. 15/05/2022).

materiales escaneados en una secuencia de bits. A esta primera codificación, sigue una segunda codificación de las informaciones más relevantes que el documento transmite, es decir, la modelización de los elementos que el editor quiere conservar, sean estos aspectos semánticos, materiales, lingüísticos, etc. del texto fuente.

Con respecto al proceso ecdótico, los sistemas de reconocimiento de textos permiten automatizar –o, en otros términos, delegar a la máquina– el primer tipo de codificación, es decir, la «digitización» (del inglés *to digit*, o sea «teclear») del texto fuente, su versión en un «magnetoescripto» (Mordenti, 2001, 49). Al respecto, apreciamos cómo en la ecdótica tradicional los dos momentos se dan de forma sincrónica, puesto que al acto de transcripción manual se integra la necesidad de explicitar cuáles son las características principales del texto que se edita[14]. Sin embargo, dentro del espacio digital, que impone ordenar y formalizar de modo secuencial las operaciones (tanto que *divide et impera* ha llegado a ser el mantra de los filólogos digitales), la conversión no supervisada de imágenes en texto electrónico y su modelización corresponden a dos procedimientos distintos y sucesivos: primero se transcribe un documento, que en un segundo momento se maqueta de acuerdo con estándares compartidos. La interpretación se limita, pues, a la segunda codificación, mientras que la primera correspondería a un acto aparentemente neutral, que quizás podemos asemejar en el contexto analógico a la realización de una edición diplomática, donde a partir de unos criterios fijos de transcripción, que evidentemente se establecen con prioridad respecto al trabajo, se proporciona un texto que es una reproducción fiel de lo que aparece en la fuente.

Por lo que atañe a la segunda codificación, que llamamos interpretativa (o, lo que es lo mismo, crítica), los sistemas de reconocimiento de textos todavía no ofrecen una solución viable de

[14] Con respecto a este asunto, ténganse en cuenta las palabras de Mordenti: «[è] possible vedere come il concetto di edizione critica contenesse in sé e unificasse molte cose assai diverse fra loro. Quando noi trascriviamo un testo noi compiamo un'operazione di ricodifica che in realtà sovrappone e mescola nel gesto del trascrivere (che ci appare, del tutto erroneamente, semplice) diverse funzioni» (2001, 47). Entre las funciones mencionadas se encuentran: la conservación, la reproducción, la corrección, el acercamiento al lector; en el caso del texto digital se suma a estas funciones la necesidad de re-codificación en formato informático para explotar después el poder de cálculo de la máquina.

automatización del trabajo, porque la maquetación tiene que pasar por una *selección* de las características textuales explícitas. Por otra parte, en el caso de la codificación del texto en formato máquina, los recientes logros en el campo de la transcripción automatizada pueden representar un avance considerable: en primer lugar, porque permiten obtener un texto fiable, que respeta con regularidad todos los signos gráficos que aparecen en la fuente; en segundo lugar porque, simplificando la transcripción manual, dejan al editor centrarse principalmente en el acto de codificación interpretativa; y, en tercer lugar, porque potencian exponencialmente la posibilidad de contrarrestar la difusión de contenidos poco fiables en la red, agilizando de modo considerable la creación de ediciones digitales que, por ser muy apegadas al texto fuente, son también más respetuosas.

Basándose en estos presupuestos, la transcripción automatizada podría afectar al mismo procedimiento ecdótico, porque permite transcribir varios testimonios de un texto en un tiempo en que normalmente la transcripción manual no llegaría ni a ofrecer la transcripción completa de un ejemplar. De tal manera, podemos imaginar que en un futuro muy cercano existirán métodos que faciliten la corrección de los testimonios transcritos y la colación inmediata de todos ellos, de cara a una publicación digital que consienta navegar a través de las variantes textuales[15]. Se trata, en otras palabras, de concebir el proceso ecdótico de otra forma, rebajando la necesidad de un texto único reconstruido y centrando la atención en la tradición textual de una obra, que es el fruto de sus distintas concreciones a lo largo del tiempo[16].

A manera de ejemplo, y para sintetizar cuanto acabamos de decir, piénsese en un proyecto que supone la edición de un texto bastante largo que cuenta con más de cinco ediciones. En el contexto analógico, necesariamente sometido al paradigma de la página, el editor seleccionaría un ejemplar fiable, lo transcribiría, cotejaría las variantes presentes en otras ediciones y ejemplares, proporcionaría un texto reconstruido que tuviera

[15] Pienso, por ejemplo, en la visualización sinóptica de variantes que ofrece la herramienta EVT2, segunda versión de la herramienta Edition Visualization Technology proporcionada por el grupo de la universidad de Turín dirigido por Roberto Rosselli del Turco. Véase Rosselli Del Turco *et al.* (2019).

[16] Al respecto, es constante la publicación de trabajos que insisten en una inédita *primacía del documento* en el contexto de la filología digital como, por ejemplo, Mordenti (2001, 31 y ss.), Pierazzo (2015) y Allés Torrent (2017, 69 y ss.).

en cuenta (utópicamente) todas las variantes y, finalmente, relegaría al aparato las lecciones alternativas encontradas. Al revés, el medio digital promete unos cambios metodológicos significativos con respecto al mismo proceso, porque no impone la selección de un testimonio ni la fijación de un texto artificialmente reconstruido. El editor, en tales condiciones, puede considerar conjuntamente todas las ediciones (incluso todos los ejemplares supervivientes), transcribir los testimonios de forma no supervisada con herramientas de OCR/HTR, corregir el texto (por medio también de *specific domain dictionaries*), detectar semi-automáticamente las variantes en las transcripciones obtenidas, publicar una edición que de modo simultáneo y dinámico permita navegar entre las distintas cristalizaciones del documento.

De tal manera la remediación digital condiciona la misma heurística del trabajo editorial, porque no se fundamenta en la intención de solucionar viejos problemas (función instrumental), sino en la posibilidad de abrir nuevas vías para la reproducción de los textos analógicos en un entorno digital. No hay duda de que los sistemas de reconocimiento de textos, por las razones indicadas, podrían convertirse en herramientas de interés para el trabajo ecdótico y tener un papel determinante en la migración web de documentos impresos.

Esto es el presupuesto principal que ha llevado a la constitución del proyecto colaborativo que tratamos a continuación.

2. Modelos de HTR para el reconocimiento de impresos hispánicos de la Edad Moderna

2.1. Primeras experimentaciones en el campo de la transcripción automatizada de impresos de la Edad Moderna

El Progetto Mambrino nació en 2003 para llevar a cabo una exploración de las continuaciones italianas de los ciclos caballerescos castellanos de Amadís y Palmerín. El grupo de investigación veronés se acercó al campo de la transcripción automática de impresos antiguos con la plataforma Transkribus a partir del año 2017; sin embargo, la idea de

transcribir y editar las obras de interés del proyecto entraba en los planes de sus directores desde hacía tiempo[17].

Hacia 2010 hubo un primer empuje debido a la posibilidad de financiar a dos becarias de investigación durante un año, con un proyecto de digitalización de los ejemplares del corpus conservados en bibliotecas locales, sobre todo la del ayuntamiento de Verona. En tal ocasión se pudieron crear unos recursos digitales que contenían las imágenes de las obras en alta calidad (formato RAW) y que salieron en DVD en una publicación unitaria de 20 discos. Al mismo tiempo se publicaron los recursos en un formato apto para la difusión en línea en el sitio web del proyecto[18], que entonces surgió precisamente como recolector de datos para censar y distinguir los ejemplares registrados, y para alojar las colecciones digitales que teníamos preparadas. En aquella ocasión, cada una de las becarias transcribió una obra del ciclo amadisiano a partir de las imágenes escaneadas de las fuentes, realizando una edición de ese ejemplar, especialmente valiosas siendo las primeras trascripciones manuales de estos libros de caballerías italianos[19].

Este primer intento fue muy importante porque puso de relieve los principales límites de sostenibilidad del proyecto: se trataba, en concreto, de transcribir un grupo de obras muy extensas –cada una de más de 900 hojas–, lo cual suponía unos costes muy elevados en términos económicos y, sobre todo, de tiempo.

Frente a esta dificultad, se consideró la oportunidad de automatizar parte de la tarea de transcripción gracias al empleo de herramientas de reconocimiento de textos. No obstante, para los impresos antiguos no existían todavía sistemas de OCR fiables. En principio, se llevaron a cabo algunos experimentos con el *software* ABBY FineReader, pero la aplicación ofrecía un reconocimiento por caracteres aislado, por lo tanto, aun entrenando el *software* con la transcripción manual de parte del texto, los resultados de transcripción no fueron prometedores. Debido a la

[17] Los primeros pasos del proyecto se describen en Bazzaco (2018).

[18] Web del Progetto Mambrino, Sección *Collezioni digitali* <https://www.mambrino.it/it/collezioni-digitali/biblioteca-civica-di-verona> (cons. 23/03/2022).

[19] Paola Bellomi editó la continuación al quinto libro de Amadís titulada *Il secondo libro delle prodezze di Splandiano* (Venezia, Michele Tramezzino, 1564); Federica Colombini editó la continuación al décimo libro, la *Aggiunta al Florisello (Le prodezze di don Florarlano)* (Venezia, Michele Tramezzino, 1564).

materialidad de las fuentes, es decir, impresos del Renacimiento en formato octavo, altamente manejados y, en ocasiones, muy desgastados por el paso del tiempo, junto con la presencia de efectos impropios de iluminación y contraste derivados del escaneo manual, se obtuvieron unos resultados inservibles, con transcripciones que tenían un margen de error elevado (Mancinelli, 2016; Bazzaco, 2018).

A partir del año 2016, establecimos los primeros contactos con el Proyecto Europeo READ (Recognition and Enrichment of Archival Documents). READ nacía de otro proyecto financiado por la Unión Europea llamado TranScriptorium[20], que tenía el objetivo de poner a disposición de los usuarios una refinada tecnología de HTR que permitiera la digitalización en formato texto electrónico de documentos de archivo, sobre todo manuscritos. Persiguiendo los mismos objetivos y tras la experiencia de este primer proyecto, READ (2016-2019), al amparo de otra financiación europea, produjo y difundió la plataforma de HTR Transkribus. Tal aplicación contaba desde su nacimiento con una consistente comunidad de referencia[21] y prometía resultados alentadores no solo con los textos manuscritos, sino también con los impresos antiguos, porque ambos compartían unos problemas parecidos en cuanto a variedad gráfica de las letras (*fluctuation*).

Guiados por la idea de considerar los impresos antiguos como si fueran documentos manuscritos muy regulares, llevamos a cabo unas pruebas con la plataforma Transkribus, que presentamos durante las jornadas de estudios «Transcribing. Towards an OCR for old fonts» (5-6 junio de 2018)[22]. La iniciativa, que se concluyó con un taller práctico para aprender a usar Transkribus, vio la participación del director del proyecto READ, el Dr. Günter Mühlberger, y de unos colaboradores del grupo de

[20] El proyecto TranScriptorium (2013-2015) fue llevado a cabo gracias a la colaboración entre la Universidad de Innsbruck, la University of London, la Universidad Politécnica de Valencia y otras instituciones. Para más detalles véase el siguiente enlace: <https://cordis.europa.eu/project/id/600707/it> (cons. 24/03/2022).

[21] Participaron en la fundación del proyecto más de diez instituciones, entre las cuales figuran, junto con las universidades ya citadas, la Universidad de Rostock, la Technische Universität de Wien, el University College de Londres. Para más detalles véase: <https://cordis.europa.eu/project/id/674943/it> (cons. 24/03/2022).

[22] Enlace a la iniciativa: <https://www.dlls.univr.it/?ent=iniziativa&id=7892&lang=it> (cons. 24/03/2022).

investigación veronés.

Para entonces, dentro del Progetto Mambrino ya se habían llevado a cabo los primeros entrenamientos sobre la cursiva veneciana del siglo XVI, lo cual había permitido obtener las transcripciones de un primer conjunto de textos del ciclo italiano de *Amadís de Gaula* con un margen de error muy bajo[23].

La labor de transcripción automática que se había llevado a cabo constó de las siguientes fases[24]:

1. Censo de los ejemplares digitalizados y descarga de las imágenes facsímiles.
2. Subida de las imágenes a la plataforma Transkribus.
3. Segmentación (*Layout Analysis*) de las imágenes en distintas áreas (correspondientes a caja y líneas de texto).
4. Transcripción manual (*Ground Truth production*) de una porción del texto segmentado (alrededor de 2000 palabras por cada obra).
5. Entrenamiento de la máquina y creación de un modelo de HTR individual (específico para cada obra).
6. Transcripción automática de las obras del corpus, en principio las que constituyen el Ciclo italiano de *Amadís de Gaula* (alrededor de 20 obras, entre traducciones y continuaciones originales);
7. Extracción de las transcripciones en formato DOC, TXT y XML para la modelización de cada edición según el estándar TEI.

Los resultados obtenidos en la cursiva gracias a Transkribus fueron muy alentadores, porque con alrededor de 2000 palabras transcritas manualmente para cada libro se podía generar una transcripción muy fiable con un índice de error (en Transkribus: *Character Error Rate*) que no superaba el 2%. Véase al respecto la siguiente tabla que indica los resultados obtenidos con los seis volúmenes del *Sferamundi di Grecia*, libro trece del *Amadís*:

[23] Para un inventario de los resultados obtenidos, consúltese Bazzaco (2018, 265-268).

[24] Una descripción completa del flujo de trabajo que supone Transkribus se encuentra en Mühlberger *et al.* (2019, 957 y ss.).

Libro	Localización ejemplar	Letra	Resultados (*CER*)
13/1 *Sferamundi. Prima parte*. 1558	Madrid, Biblioteca Nacional de España, 5-4978	cursiva	1.57%
13/2 *Sferamundi. Seconda parte*. 1560	Madrid, Biblioteca Nacional de España, 5-4978	cursiva	1.21%
13/3 *Sferamundi. Terza parte*. 1563	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 20)	cursiva	1.80%
13/4 *Sferamundi. Quarta parte*. 1563	München, Bayerische Staatsbibliothek, P.o.hisp. 105 k-4	cursiva	1.11%
13/5 *Sferamundi. Quinta parte*. 1565	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 22)	cursiva	1.59%
13/6 *Sferamundi. Sesta parte*. 1565	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 22)	cursiva	1.71%

Tabla 1. Primeros resultados de transcripción automática con el *Sferamundi di Grecia* (Venecia, s. XVI)

En un segundo momento, las transcripciones obtenidas, revisadas por unos especialistas y etiquetadas según el estándar XML TEI compondrían el corpus piloto de la Biblioteca Digital del Progetto Mambrino, un proyecto de edición digital de los libros de caballerías italianos que se está llevando a cabo hoy en día gracias a una financiación concedida en 2018 por el Ministerio Italiano de Universidad dentro de un proyecto más amplio llamado *Mapping Chivalry: Spanish Romances of Chivalry from Renaissance to 21st Century: a digital approach* (en el que participa un equipo que incluye cuatro unidades: de Verona, Trento, Roma y Salerno) y otra financiación que obtuvo en el mismo año el Departamento de Lenguas y Literaturas Extranjeras de la Universidad de Verona, como Departamento de Excelencia.

En el tiempo en que estábamos transcribiendo de forma automática la letra cursiva, se fortalecieron los contactos con los proyectos de investigación españoles BIDISO y COMEDIC, que propusieron poner a prueba la plataforma Transkribus con textos castellanos de su interés.

En el interior del Progetto Mambrino ya se estaba pensando en extender el reconocimiento automático a los textos españoles, gracias a la elaboración de unos modelos de reconocimiento para obras caballerescas impresas en el XVI en letra gótica. Las primeras pruebas con la gótica fueron muy alentadoras, ya que con un *dataset* muy limitado de páginas

transcritas manualmente (alrededor de 1500 palabras por obra) pudimos entrenar un modelo de HTR individual para los libros de caballerías castellanos con el que experimentamos[25]. Finalmente, una vez acabado el entrenamiento de la máquina, transcribimos las obras caballerescas en gótica con un margen de error cercano al 2%. Poco más tarde fueron los primeros intentos de transcripción automática de documentos impresos en letra redonda, que aseguraron unos resultados parecidos (Tabla 2).

Libro	**Localización ejemplar**	**Letra**	**Resultados (*CER*)**
Leandro el Bel. Toledo, Ferrer, 1563	Madrid, Biblioteca Nacional de España, R/9030	gótica	1.43%
Florando de Inglaterra. Lisboa, Gallarde, 1545	London, British Library, C62 H14	gótica	2.13%
Silves de la Selva. Sevilla, De Robertis 1549	Madrid, Biblioteca Nacional de España, R/865	gótica	1.58%
Libro de los Siete Sabios de Roma. Barcelona, Andreu, 1678	Madrid, Biblioteca Nacional de España, R/530	redonda	2.30%

Tabla 2. Primeros resultados de transcripción automática de documentos en gótica y redonda

Si se tiene en cuenta que estas pruebas se realizaron en un momento en que la plataforma Transkribus estaba todavía en su desarrollo tecnológico y contaba con un número mínimo de documentos procesados, se puede comprender cuáles son las posibilidades reales que ofrece la herramienta. En la actualidad, con la introducción de nuevos sistemas de reconocimiento (del HTR básico a los métodos HTR+/PyLaya) y una cantidad interminable de documentos procesados, podemos imaginar que los resultados serían aún más convencedores[26].

En la estela de estos primeros logros y guiados por la idea de extender el reconocimiento a un conjunto más extenso de textos hispánicos de la Edad Moderna, se generó una red de colaboración internacional entre

[25] Los colaboradores del proyecto que participaron fueron: Stefano Bazzaco (*Leandro el Bel*), Stefano Neri (*Florando de Inglaterra*), Giada Blasut (*Silves de la Selva*). Otras pruebas con la redonda fueron realizadas por Bazzaco en el mismo año.

[26] Recordamos al respecto que el reconocimiento en Transkribus, fundándose en procedimientos de *machine learning*, incrementa sus prestaciones según aumenta la cantidad de documentos procesados (Mühlberger *et al.*, 2019, 957).

investigadores de los tres proyectos, con el fin de proporcionar unos recursos que fueran de utilidad para toda la comunidad científica.

2.2. Modelos de HTR para la transcripción automática de impresos hispánicos en gótica y redonda

El proyecto de reconocimiento de impresos hispánicos de la Edad Moderna nace formalmente en 2021 de la colaboración entre tres proyectos de investigación distintos: el Progetto Mambrino (Universtà di Verona), BIDISO (Universidade da Coruña) y COMEDIC (Universidad de Zaragoza).

En un primer seminario formativo impartido por Stefano Bazzaco, se constituyó una red de colaboración que vio la participación de una docena de investigadores en un trabajo colectivo para la transcripción manual y el entrenamiento de la plataforma Transkribus en relación con distintos textos impresos en gótica y en redonda. Los investigadores involucrados en esta colaboración fueron: Giada Blasut, Federica Zoppi, Manuel Garrobo Peral (Progetto Mambrino); Ana Milagros Jiménez Ruiz, Ángela Torralba Ruberte, Nuria Aranda García, Daniela Santonocito, Gaetano Lalomia (COMEDIC); Carlota Fernández Travieso, Mónica Martín Molares (BIDISO). Los resultados del proyecto se dieron a conocer por primera vez en el Congreso Internacional «Humanidades Digitales y estudios literarios hispánicos. De los impresos de la Edad Moderna a las ediciones académicas digitales», que se celebró en la Universidad de Verona en junio de 2021.

Objetivos del proyecto colaborativo: la finalidad del proyecto, como ya se anticipaba, era poner a disposición de otros estudiosos unos modelos de HTR que fueran aplicables directamente, sin la necesidad de entrenar la máquina cada vez que se necesitaba un nuevo texto transcrito. Por esta razón, buscamos la vía para generar unos modelos de reconocimiento extendidos que pudieran abarcar un conjunto muy amplio de obras. Con respecto a los modelos de HTR individuales, que están basados en una sola obra, los modelos de HTR extendidos se fundamentan pues en obras

distintas y permiten transcribir todos los textos que presenten unas características tipográficas parecidas con un buen grado de fiabilidad.

Flujo de trabajo: para empezar, se escogieron los documentos que constituirían el *dataset* del modelo, es decir, obras de naturaleza variada que abarcasen distintas representaciones gráficas de caracteres del mismo tipo. La creación de unos modelos de HTR extendidos constó entonces de tres etapas principales: transcribir manualmente, según unos mismos criterios, un número determinado de páginas (en nuestro caso, alrededor de 20 páginas) pertenecientes a distintas obras; entrenar la máquina sobre el conjunto de transcripciones realizadas; generar un modelo de HTR único que transcribiera automáticamente otros textos que no pertenecieran al conjunto inicial. Para cumplir con este flujo de trabajo, Transkribus ha sido un recurso fundamental, porque se presenta como una plataforma con varias funciones de automatización y realmente colaborativa, ya que consiente la interacción de los usuarios de forma simplificada y asíncrona[27].

En principio, se fijaron unos criterios de transcripción rigurosos para el conjunto de transcripciones manuales que se realizaron para la producción de los dos modelos extendidos, uno para la gótica (siglos XV-XVI) y uno para la redonda (siglos XVI-XVII). Tales criterios se regían por tres necesidades principales: preservar todos los signos gráficos que presentaba la fuente; aligerar el proceso de postproducción de los textos exportados; fomentar la solidez (*consistency*) del modelo, o sea, mantener una regularidad total en las transcripciones manuales para que la máquina pudiera aprender unos patrones recurrentes de identificación de las letras[28]. Finalmente, elegimos unos criterios de transcripción semi-diplomática que respetaran la variabilidad textual de las fuentes, pero que

[27] Entre otras señalamos las siguientes funciones: Text2Image, para importar en la plataforma textos ya transcritos; P2PaLA, para la segmentación de la página digitalizada con modelos de *layout* predefinidos; Keyword Spotting (KWS), para la extracción de palabras clave a partir de la forma gráfica de las mismas. Para más detalles véase la página de recursos <https://readcoop.eu/it/transkribus/resources/> (cons. 24/03/2022).

[28] Las convenciones de transcripción establecidas por Transkribus están disponibles en la web de Read Coop, en el siguiente enlace: <https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/> (cons. 24/03/2022).

a la vez consintieran reducir intervenciones posteriores en fase de revisión, sobre todo en el caso de las abreviaturas (Tabla 3):

Criterios de transcripción [29]
1. Signos de interpunción: se mantienen como aparecen en el texto fuente. 2. Acentos: se mantienen como aparecen en el texto fuente. 3. Signo tironiano: se trascribe como 'e' comercial (&). 4. 's' larga (ſ): se transcribe como 's' simple. 5. Abreviaturas: se desarrollan.

Tabla 3. Criterios de transcripción adoptados

Resultados: una vez acabadas las transcripciones manuales (*Ground Truth*) que habrían constituido la base del entrenamiento dentro de la plataforma, generamos los dos modelos de HTR que se describen brevemente a continuación (Tabla 4).

SpanishGothic_XV-XVI_extended [30]	**SpanishRedonda_XVI-XVII_extended** [31]
Versión actual: 1.0.0 Dataset: 16 textos, 150˙137 palabras Fiabilidad: 99.08%	Versión actual: 1.0.0 Dataset: 14 textos, 61˙938 palabras Fiabilidad: 98.93%

Tabla 4. Descripción sintética de los dos modelos de HTR publicados

Aun teniendo en cuenta que la fiabilidad de los dos modelos se basa en los textos que constituyen el *Dataset* y que, por consiguiente, en los textos externos al modelo el porcentaje de letras transcritas correctamente puede disminuir, hay que subrayar que las primeras pruebas efectuadas

[29] Los criterios están disponibles en GitHub, en los siguientes enlaces: <https://github.com/stefanobazzaco/HTR-model-SpanishGothic_XV-XVI_extended#transcription-criteria> / <https://github.com/stefanobazzaco/HTR-model-SpanishRedonda_XVI-XVII_extended#transcription-criteria> (cons. 24/03/2022).

[30] Para una descripción detallada del modelo remito al *Apéndice 1* del presente trabajo.

[31] Para una descripción detallada del modelo remito al *Apéndice 2* del presente trabajo.

han asegurado cierta consistencia con respecto a los resultados obtenidos, con índices de error muy bajos[32].

Distribución: los modelos de HTR creados están disponibles en abierto desde julio de 2021 dentro de la plataforma Transkribus en la sección *Public Models*; pueden, por lo tanto, ser empleados por cualquier usuario que tenga acceso a la plataforma.

Para la utilización de los modelos, el estudioso tiene que atender al siguiente flujo de trabajo:

a) de entrada, subir las imágenes digitalizadas de la obra de interés a la plataforma;
b) ejecutar la segmentación de las imágenes (*Layout Analysis*) de forma automatizada o manual[33];
c) lanzar el reconocimiento con uno de los modelos a disposición (duración: unos minutos por página).

Por medio de estos pasajes, el usuario podrá pasar a la exportación de las transcripciones obtenidas en varios formatos (DOC, PDF, TXT, XML)[34].

Contribución y puesta al día: los *datasets* que constituyen la base del entrenamiento de los modelos están disponibles en el repositorio Zenodo con acceso limitado[35], según el protocolo establecido por la licencia Creative Commons CC BY-NC-ND 4.0, que impide la creación de objetos derivados y su distribución con finalidades comerciales[36]. En el futuro, se piensa implementar progresivamente cada modelo de HTR por medio de

[32] Al respecto, véanse las primeras pruebas efectuadas por Giada Blasut en esta misma publicación (pp. 175-193).
[33] Se dispone también de un modelo de P2PaLA que permite la segmentación automatizada de textos en doble columna. Para más información contactar con los autores.
[34] Para una descripción exhaustiva del flujo de trabajo de la plataforma Transkribus, véase Bazzaco (2018) o bien la web de READ Coop: <https://readcoop.eu/transkribus/resources/how-to-guides/> (cons. 24/03/2022).
[35] Dataset del modelo de HTR SpanishGothic_extended_sXV-XVI: <https://zenodo.org/record/4888927#.Yj2hSerMI2w>. Dataset del modelo de HTR SpanishRedonda_extended_sXVI-XVII: <https://zenodo.org/record/4889218#.Yj2hS-rMI2w> (cons. 24/03/2022).
[36] Para más información, véase el siguiente enlace: <https://github.com/stefanobazzaco/HTR-model-SpanishRedonda_XVI-XVII_extended#copyright-statement> (cons. 24/03/2022).

la inserción de otras porciones de textos transcritos manualmente por distintos estudiosos que quieran colaborar en el proyecto colaborativo. En este sentido, Zenodo se revela como un recurso de fundamental importancia porque permite que los *datasets* implementados se pongan al día con regularidad cada semestre, incrementando las prestaciones de los modelos publicados y manteniendo un mismo identificador unívoco (DOI) para ambos recursos.

Conclusiones

El proyecto colaborativo descrito se fundamenta, evidentemente, en una constante actualización que pueda proporcionar resultados cada vez más satisfactorios para el filólogo. El punto de partida, como se ha señalado, era el de contrarrestar los efectos de la migración descontrolada de textos no especializados en la red; sin embargo, los resultados obtenidos y el flujo de trabajo establecido permiten añadir unas consideraciones ulteriores.

En concreto, observamos cómo el área del reconocimiento de texto está prometiendo beneficios de gran interés para la investigación: no hay duda de que la tecnología HTR ha alcanzado recientemente un nivel de desarrollo muy alto, permitiendo la transcripción automática de textos impresos con un nivel de fiabilidad muy elevado y apuntando a la reducción de trabajo para la publicación de ediciones científicas. Sin embargo, para invertir la tendencia de una difusión masiva de materiales textuales dudosos y alimentar la red con textos controlados, en primer lugar, es necesario reparar en el hecho de que la tecnología de por sí no es suficiente: debe estar guiada por una mirada especializada que supervise el desarrollo de los *softwares* de forma concienzuda.

Al respecto, las preocupaciones de Padre Roberto Busa, quien percibió muy pronto el desarrollo informático y computacional como un riesgo para la producción de datos sólidos y eficientes, aptos para la investigación humanística, parecen aún más fundamentales y urgentes. En opinión del jesuita, la nueva velocidad que se impuso al procesamiento del lenguaje y a la digitalización de documentos textuales no constituye de por

sí una respuesta a la degradación de los contenidos informativos de la web, sino que tiene que acompañarse con una interpretación ponderada de carácter inductivo, que busque la evidencia empírica y la producción de unos datos que puedan constituir una documentación fiable y replicable. En otras palabras,

> he foresaw that the wide availability of large collections of digitized textual data and of tools for processing them automatically would run the risk of being incorrectly exploited. Busa believed the greatest danger lay in considering Computational Linguistics (and Digital Humanities, too) not as a discipline aimed at doing things better, but rather as a tool to do things faster, both in the phase of collecting data and in that of exploiting data. He feared that the computational linguists and the digital humanists of the third millennium would cease caring for the quality of data and lose the humility to check them carefully, preferring instead to process huge masses of texts quickly and approximately, without even reading a line (Rockwell-Passarotti, 2019, 26).

En la línea de evitar que la explotación de grandes masas de datos sustituya a la recolección y al enriquecimiento de datos apropiados, los trabajos colaborativos del tipo que tratamos en estas páginas adquieren un notable interés. Estos representan una respuesta concreta a la producción de objetos digitales fiables y, a la vez, se sustentan en una visión nítida del fenómeno, que prevé que el trabajo de los humanistas esté sujeto principalmente al atento análisis de los datos y no en asuntos que atañen por la mayoría a la mejora de las máquinas. Y esto porque solo a partir de datos más consistentes y refinados, la implementación tecnológica podrá sustentar la investigación científica de forma correcta y rigurosa.

La importancia de los tres proyectos implicados y la gran experiencia en tema de edición de impresos de la Edad Moderna de los colaboradores involucrados constituye finalmente la *conditio sine qua non* para el desarrollo del proyecto y un aspecto determinante para que la colaboración proporcione resultados correctos, replicables y explotables por otros estudiosos, capaces de alimentar un conocimiento responsable, documentable y sistemático de la textualidad en el ámbito digital.

Apéndice 1. Modelo de HTR SpanishGothic_extended_sXV-XVI

Descripción Dataset:
Tipo de documentos: impresos
Nr. de palabras: 150 137
Nr. de líneas: 16 816
CER Training Set: 0.45%
CER Validation Set: 0.92%

Autores:
Stefano Bazzaco (coord.), Giada Blasut, Federica Zoppi, Nuria Aranda García, Ángela Torralba Ruberte, Ana Milagros Jiménez Ruiz, Pedro Monteiro

Versión actualmente disponible: versión 1.0.0 (julio 2021)

Cómo citar: Stefano Bazzaco (coord.), Federica Zoppi, Giada Blasut, Nuria Aranda García, Ángela Torralba Ruberte, Ana Milagros Jiménez Ruiz, & Pedro Monteiro. (2021). HTR model SpanishGothic_XV-XVI_extended DATASET (1.0.0) [Data set]. Zenodo.
DOI: <https://doi.org/10.5281/zenodo.4888927>

Enlaces:
https://github.com/stefanobazzaco/HTR-model-SpanishGothic_XV-XVI_extended
https://zenodo.org/record/4888927#.YLX6xqgzY2w
https://readcoop.eu/model/spanish-gothic-15th-16th-century/

Presentación del corpus

Los principios que rigieron la selección del corpus estuvieron condicionados por los intereses del Progetto Mambrino, COMEDIC y BIDISO. En general, perseguimos dos objetivos principales: la aplicación del modelo a un grupo de ediciones muy variado desde un punto de vista editorial (Grupo Misceláneo) y el uso de la herramienta en géneros editoriales homogéneos y consolidados en la época de la imprenta manual (Grupo de Libros de Caballerías, Grupo de Historias breves de caballería).

Desde finales del siglo XV hasta el segundo tercio del XVI, concretamente hasta 1560, el uso de tipos góticos será el hegemónico en la imprenta hispánica. Dentro de este marco temporal y tipográfico, el número de ediciones que forman nuestro corpus asciende a dieciséis (Tabla 5). A partir del género editorial de estos impresos[37], hemos establecido tres partes principales.

En primer lugar, encontramos un grupo misceláneo constituido por un conjunto de ediciones de diversa disposición bibliográfica y género literario. Estas son: el *Doctrinal de los caballeros* de Alonso de Cartagena, *La Fiameta* de Juan Boccaccio, la *Crónica del Rey Don Rodrigo* de Pedro del Corral, también conocida como *Crónica Sarracena*, el *Retablo de la Vida de Cristo* de Juan de Padilla, el «Cartujano», y una nueva edición de la *Tragicomedia de Calisto y Melibea* de Fernando de Rojas.

El segundo grupo está formado por cinco ediciones de libros de caballerías, un género literario cuyas obras, al presentar una disposición editorial homogénea, muestran errores de reconocimiento similares. Las obras escogidas de este género son: el *Lisuarte de Grecia* de Juan Díaz, el anónimo *Florando de Inglaterra*, el *Silves de la Selva* de Pedro de Luján, el *Lisuarte de Grecia* de Feliciano de Silva y el *Leandro el Bel* de Pedro de Luján.

Por último, son seis las obras del género de las historias breves de caballerías que componen el tercer grupo: tres ediciones diferentes de *El libro del conde Partinuplés* y, respectivamente, una de la *Historia de la linda Magalona*, la *Historia de la reina Sebilla* y la *Historia del rey Canamor*.

[37] Entre los muchos trabajos sobre la cuestión del «género editorial» de Víctor Infantes (especialmente, 1992), consideramos muy acertada la síntesis que presenta en Infantes (2003).

Título	Año	Lugar de impresión	Tipología	Formato	Disposición del texto	n.º de ficha en COMEDIC
Doctrinal de los caballeros	1487	Burgos. Fadrique de Basilea	Miscelánea	Folio - 168 h.	Línea tirada	82
La Fiameta	1497	Salamanca. Impresor de la Gramática de Nebrija	Miscelánea	Folio - 44 h.	Dos columnas	147
Crónica del Rey Don Rodrigo	1499	Sevilla. Meinardo Ungut y Estanislao Polono	Miscelánea	Folio - 227 h.	Dos columnas	53
Retablo de la Vida de Cristo	1510	Sevilla. Juan Cromberger	Miscelánea	Folio - 58 h.	Dos columnas	309
Tragicomedia de Calisto y Melibea	[1512 - 1515]	Roma. Marcelo Silber	Miscelánea	4º - 80 h.	Línea tirada	322
Lisuarte de Grecia	1526	Sevilla. Jacobo y Juan Cromberger	Libro de caballerías	Folio - 123 h.	Dos columnas	/
Lisuarte de Grecia	1550	Sevilla. Jácome Cromberger	Libro de caballerías	Folio - 109 h.	Dos columnas	/
Florando de Inglaterra	1545	Lisboa. Germán Gallarde	Libro de caballerías	Folio - 172 h.	Dos columnas	/
Silves de la Selva	1549	Sevilla. Dominico de Robertis	Libro de caballerías	Folio - 150 h.	Dos columnas	/
Leandro el Bel	1563	Toledo. Miguel Ferrer	Libro de caballerías	Folio - 128 h.	Dos columnas	/
El libro del conde Partinuplés	1519	Sevilla. Jacobo Cromberger	Historia breve de caballerías	4º - 95 h.	Línea tirada	106
El libro del conde Partinuplés	1558	Burgos. Herederos de Juan de Junta	Historia breve de caballerías	4º - 86 h.	Línea tirada	106
El libro del conde Partinuplés	1563	Burgos. Felipe de Junta	Historia breve de caballerías	4º - 86 h.	Línea tirada	106
Historia de la linda Magalona	1519	Sevilla. Jacobo Cromberger	Historia breve de caballerías	4º - 63 h.	Línea tirada	213
Historia de la reina Sebilla	1551	Burgos. Felipe de Junta	Historia breve de caballerías	4º - 70 h.	Línea tirada	/
Historia del rey Canamor	1527	Valencia. Jorge Costilla	Historia breve de caballerías	4º - 110 h.	Línea tirada	347

Tabla 5. Listado de las obras del corpus para el modelo *SpanishGothic*

a) Grupo misceláneo

Las ediciones que integran este grupo presentan como común denominador el haber sido impresas en los primeros decenios de implantación y consolidación de la imprenta hispánica, tres de ellas durante el periodo incunable y dos en el periodo postincunable. Durante este periodo de evolución editorial (González-Sarasa Hernáez, 2013), el formato de las ediciones determina en gran medida la configuración de los elementos que constituyen la caja de escritura –cabeceras, letras xilográficas y lombardas, grabados, foliación– y estos, a su vez, condicionarán la aplicación del *layout*. Por tanto, la exposición de nuestro caso práctico está regida por la realidad bibliográfica de las ediciones.

En lo que se refiere a los incunables, a pesar de la variedad en su género literario, tanto la edición del *Doctrinal de los Caballeros* impresa en Burgos en 1487 (el 20 de junio) por Fadrique de Basilea[38], como *La Fiameta* publicada en Salamanca en 1497 y atribuida al Impresor de la Gramática de Nebrija[39], y la *Crónica del Rey Don Rodrigo* (*Crónica Sarracina*) impresa en Sevilla en 1499 por Meinardo Ungut y Estanislao Polono[40] presentan un formato en folio. La caja de escritura del *Doctrinal* muestra el texto literario en una columna de 35 líneas –a pesar de que el título y las tablas están dispuestos a doble columna–, mientras que el texto de *La Fiameta* y la *Crónica Sarracina* se divide en dos columnas de 48 y 47 líneas, respectivamente.

La aplicación del análisis de la plana –*Layout Analysis*– ha sido exitosa en casi todas las partes de la caja tipográfica de los ejemplares del *Doctrinal* y de la *Crónica* –incluida la cabecera–. En el caso de *La Fiameta*, se ha tenido que reconstruir de forma manual en más de la mitad de las hojas, puesto que la digitalización del ejemplar presenta una baja calidad y multitud de manchas, fruto del deterioro del ejemplar. La segmentación ha funcionado solamente en la identificación de las dos columnas, pero no en la

[38] Se ha trabajado con la digitalización del ejemplar conservado por la Real Academia Española, con la signatura Inc. San Román 6. Digitalización en color.

[39] Se ha trabajado con la digitalización del ejemplar localizado en la Pierpont Morgan Library de Nueva York (Incunable Collection-Oversize: INCUNOS1, ChL 1742: PML 667). Digitalización en blanco y negro.

[40] Se ha trabajado con la digitalización del ejemplar localizado en la Hispanic Society of America (signatura Inc. 84).

separación de las líneas de cada una y, de hecho, ha habido varias hojas que han resultado irreconocibles para el modelo.

En los tres casos, se ha hallado dificultad solamente en el reconocimiento de las letras xilográficas y lombardas mayúsculas, que dan comienzo al texto, y que se han transcrito manualmente. En *La Fiameta*, no se ha reconocido la xilográfica mayúscula inicial, pero en el resto de lombardas el programa ha detectado un elemento no identificado para el que ha dejado un espacio libre –a completar por el usuario del programa. El reconocimiento de estas letras ilustradas es una de las limitaciones de Transkribus ya que, para una correcta identificación de las mismas –especialmente las xilográficas–, sería preciso emplear programas destinados al reconocimiento de grabados[41].

En el caso del *Retablo de la Vida de Cristo*, obra de gran éxito editorial por el elevado número de ediciones[42], se ha escogido la edición impresa en Sevilla en 1510 por Juan Cromberger[43]. Presenta un formato en folio y una disposición en la caja de escritura muy elaborada: además de la cabecera, el texto principal está dividido en dos columnas con 52 líneas (máx.) que incluyen grabados xilográficos, junto con los nombres de los personajes ficcionales que intervienen y que aparecen en ambos márgenes. De este modo, en total, el *Layout Analysis* había de reconocer cuatro columnas. Sin embargo, esta compleja distribución sumada a la baja calidad de la digitalización ha provocado que el programa presente problemas en la identificación de la cabecera, los grabados y las *marginalia*, que se identificaban dentro de la misma *baseline* del texto principal (Figura 1). Todos estos casos han tenido que resolverse de forma manual.

[41] A este respecto, conviene destacar los programas de OCR de ilustraciones empleados por grupos de investigación como 15cBookTrade de la Universidad de Oxford <http://15cbooktrade.ox.ac.uk/> (cons. 15/05/2022) o The Illustrated book in Lyon 1480-1600 - Equipex Biblissima, dirigido por la Dra. Barbara Tramelli.

[42] Véase la ficha 309 en Comedic: *Catálogo de obras medievales impresas en castellano hasta 1600* [en línea] <http://grupoclarisel.unizar.es/comedic/> (cons. 15/05/2022).

[43] Se ha trabajado con el ejemplar de la Biblioteca Nacional de España (signatura R/31133-3).

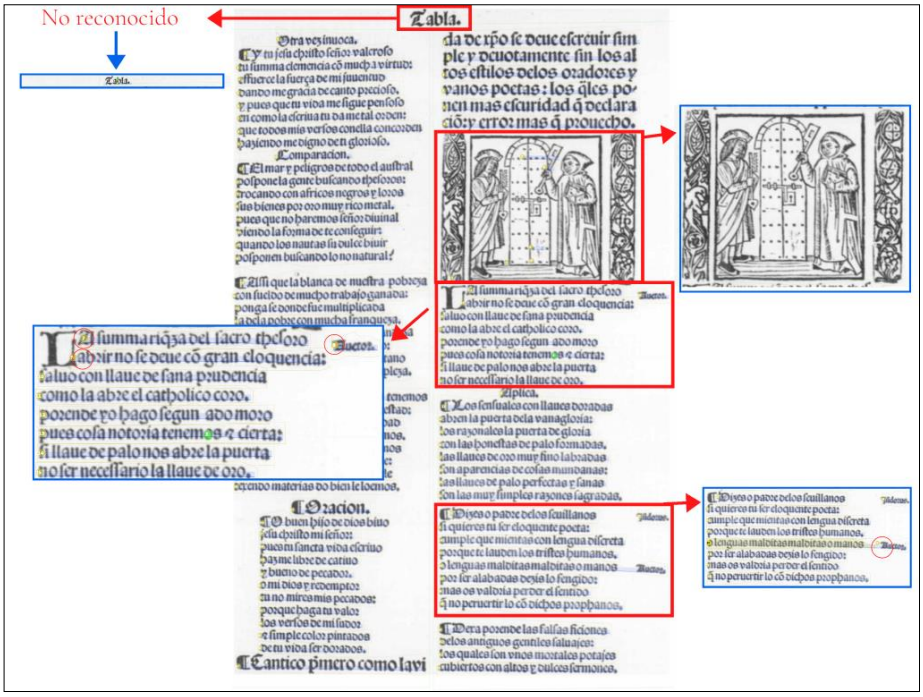


Fig. 1. Muestra del *Layout Analysis* fallido en el ejemplar del *Retablo de la Vida de Cristo*

Respecto a la *Tragicomedia de Calisto y Melibea* de Fernando de Rojas impresa en Roma entre 1512-1515 por Marcello Silber[44], presenta un formato en cuarto con una caja de escritura constituida por una columna de 37 líneas. El proceso de *Layout Analysis* ha reconocido todas las *baselines*, incluida la secuenciación de personajes que aparece al inicio de cada auto a modo de *dramatis personae*. Al presentar los mismos tipos que el texto principal, el programa ha leído el compendio de personajes como una línea de texto al uso; por lo que se ha tenido que indicar de forma manual que constituye otro campo. Al igual que los grabados del *Retablo*, las características figuritas factótum celestinescas se han identificado como campos textuales. Se trata de un error que solamente se ha producido en una ocasión, ya que el programa ha sabido discriminar correctamente los grabados y obviarlos en el resto de páginas.

[44] Se ha trabajado con la digitalización del único ejemplar conocido, y localizado en la Biblioteca Estatal de Ulm (signatura: Schad 4434). Disponible en: <http://openaccess-stadtbibliothek.ulm.de/pdf/Reproduktionen/Schad_4434/?C=M;O=A> (cons. 10/05/2022).

b) Textos de temática caballeresca

Se han empleado once textos de temática caballeresca en el entrenamiento del modelo *Gothic Extended* en el programa *Transkribus*. Por sus rasgos materiales y de *dispositio* textual, podemos dividirlos en dos grupos diferentes, que nos permiten explicarlos en conjunto.

Por un lado, cinco textos que forman parte del género de los libros de caballerías. En primer lugar, dos ediciones diferentes del *Lisuarte de Grecia*: la impresa por Jacobo y Juan Cromberger en 1526 (BNE: R/71) y la edición publicada a cargo de Jácome Cromberger en 1550 (BNE: R/13138(2)). La edición del *Silves de la Selva* de Pedro de Luján impresa en Sevilla por Dominico de Robertis en 1549 (BNE: R/865); la del *Florando de Inglaterra*, impreso en Lisboa por Germán Gallarde en 1545 (British Library: C.62.h.14); y, por último, el ejemplar de la Biblioteca Nacional de España (R/9030) del *Leandro de Bel* de Pedro de Luján en la edición impresa en Toledo por Miguel Ferrer en 1563.

Por otro lado, seis impresos que constituyen parte del conjunto de textos que en los últimos años se han denominado «historias breves de caballerías»[45]. Dentro de esta serie, hemos empleado tres ediciones diferentes de *El libro del conde Partinuplés*: la edición sevillana impresa por Jacobo Cromberger en 1519 (BNL: Res. 401/18) y dos testimonios burgaleses. El primero data de 1558 y fue impreso por los herederos de Juan de Junta (BNE: R/31364/38) y el segundo en 1563 por Felipe de Junta (British Library: C.55.d.4). Además, también hemos utilizado el ejemplar *unicum* de la *Historia de la linda Magalona*, impreso en Sevilla por Jacobo Cromberger en 1519 (British Library: C.7.a.18); la edición de la *Historia de la reina Sebilla* impresa en 1551 por Felipe de Junta (Bibliothèque Nationale de France: Rés. Y2849) y, por último, la edición valenciana de 1527 de la *Historia del rey Canamor*, impresa por Jorge Costilla (Universidad de Oviedo: CEA-227).

[45] Véanse al respecto: Baranda (1991, 183-191); Infantes (1991, 165-182 y 1996, 127-132).

Libros de caballerías

Desde el punto de vista del cuerpo del texto, los cinco textos presentados, al pertenecer al género literario caballeresco, se caracterizan por ser impresos en formato folio, tamaño idóneo para la edición de obras extensas. De hecho, «el formato del género editorial caballeresco que se imprime en talleres peninsulares se limita al folio» (Lucía Megías, 2000, 431). En cuanto a la disposición del interior del libro de caballerías, el texto de este tipo de obras aparece siempre distribuido en dos columnas, así como sucede con las tablas de capítulos, mientras que el de los preliminares legales (privilegio, licencia, aprobación, fe de erratas y tasa) y el de los prólogos aparece siempre a línea tirada (2000, 448). Asimismo, estos textos presentan los títulos o cabeceras centradas en la parte superior de cada una de las planas. Por último, en la esquina superior derecha se coloca la foliación de las páginas mediante números romanos (Figuras 2 y 3).

Fig. 2 y 3. Ejemplos de cabeceras en el *Lisuarte de Grecia* (Sevilla, 1526), fols. 4v-5r.

Como se observa en las imágenes, estas obras presentan el modelo más habitual de cabecera: el título se reparte entre el vuelto y el recto de los folios, de modo tal que al tener abierto el impreso podamos leer el título completo en la parte superior. De acuerdo con la clasificación de Lucía Megías, se trata de las «cabeceras que indican el lugar que ocupan el texto en una serie más amplia de libros» (2000, 451).

Ambos rasgos propios de los libros de caballerías, tanto la división del texto en dos columnas como la introducción de la cabecera, no generan ningún problema con el programa Transkribus al aplicar el modelo de segmentación *2columns+heading* creado por medio de la función P2PaLA (*Page to Page Layout Analysis*), como se puede ver en la siguiente imagen (Fig. 4):

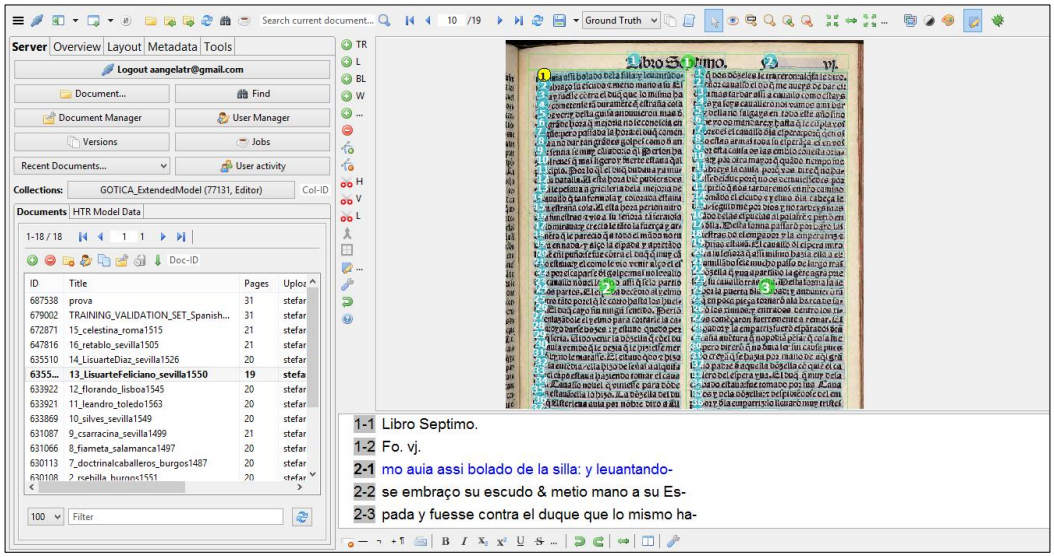


Fig. 4. Segmentación con modelo de P2PaLA del *Lisuarte de Grecia* (Sevilla, 1550)

Asimismo, en todos los casos la primera línea del epígrafe de cada capítulo presenta un tipo diferente y de mayor tamaño. Como ejemplo paradigmático que se puede observar en las figuras, ofrecemos la edición del *Lisuarte de Grecia* de 1526 que emplea para la primera línea del epígrafe una letrería G158 (T:2(C)) y para el resto del texto un tipo G98 (T:8b).

Tras dichos epígrafes que encabezan cada capítulo, el texto se inicia

con una capital xilográfica, que en la edición que hemos tomado como referencia aparece recuadrada en un ribete que ocupa cuatro líneas.

Además, tras varios epígrafes, se inserta un pequeño grabado en blanco y negro que ilustra el contenido de dicho capítulo y ocupa once líneas. Por último, también las portadas aparecen decoradas con un grabado. Por ejemplo, el *Lisuarte de Grecia* ha ilustrado su portada con una xilografía a dos tintas, negra y roja, que ha enmarcado por una orla. La ilustración de la portada contiene seis grabados que representan a diferentes caballeros del ciclo (Fig. 5):

Fig. 5. Portada del *Lisuarte de Grecia* (Sevilla, 1526)

Estos elementos decorativos y explicativos, tanto los que se ubican en la portada como los del interior del relato, han sido pasados por alto por el programa, puesto que la detección del *layout* con imágenes en colores no considera relevantes las zonas con densidad de píxeles menor que la del cuerpo del texto.

Historias breves de caballerías

A diferencia de los extensos libros de caballerías, esta serie de textos se caracteriza por la brevedad, ya que no llegan habitualmente a ocho pliegos (64 páginas) y suelen estar impresos en formato 4º, en lugar de folio. Como resultado de la reducción material, tampoco requieren un excesivo material gráfico ni una especial disposición impresa: el texto se presenta a línea tirada, desaparecen las cabeceras y, por último, la paginación, también en números romanos, se traslada a la esquina inferior derecha de la página. La nueva distribución de la *mise en page* implica que no hay necesidad de aplicar modelos preconcebidos de P2PaLA en Transkribus (Fig. 6).

Por otro lado, de manera similar a las ediciones del *Lisuarte de Grecia*, la primera línea del epígrafe presenta un tipo diferente y de tamaño mayor. En la *Historia del rey Canamor* se emplean tipos de tres fundiciones: para el título Gótica-38: c.190-G; para el epígrafe inicial Gótica-15B: c.132-G y para el texto restante Gótica-22: 100-G (Fig. 7).

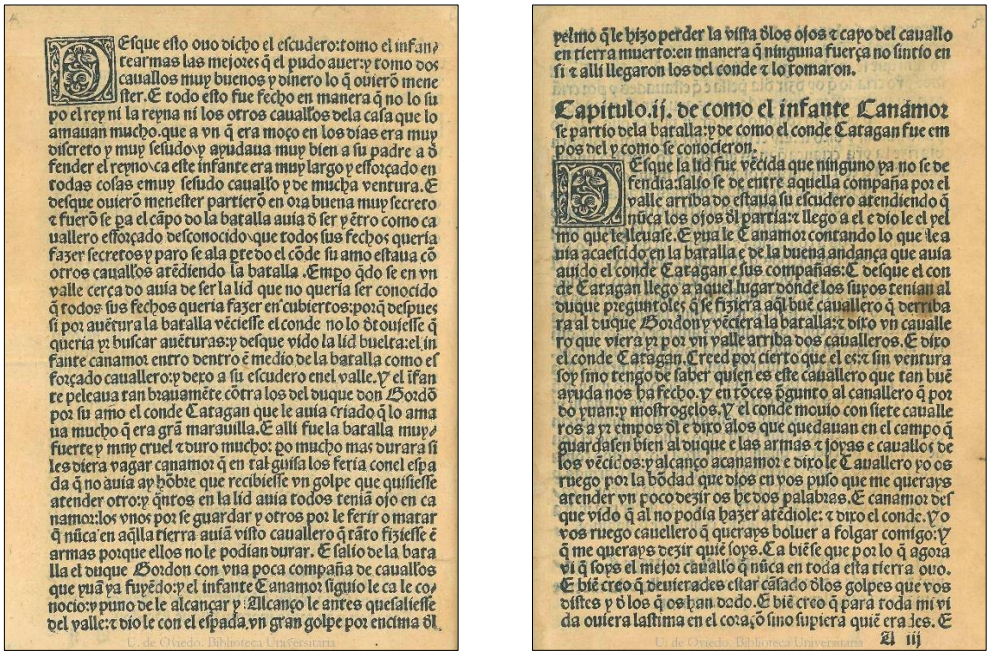


Fig. 6 y 7. Distribución, tamaño de texto y epígrafe en el *Libro del Rey Canamor* (Valencia, 1527)

Por último, todos estos textos incluyen un grabado únicamente en la portada del relato. Es decir, a diferencia de los ejemplares del *Lisuarte de Grecia*, no se ha ilustrado el interior de los testimonios.

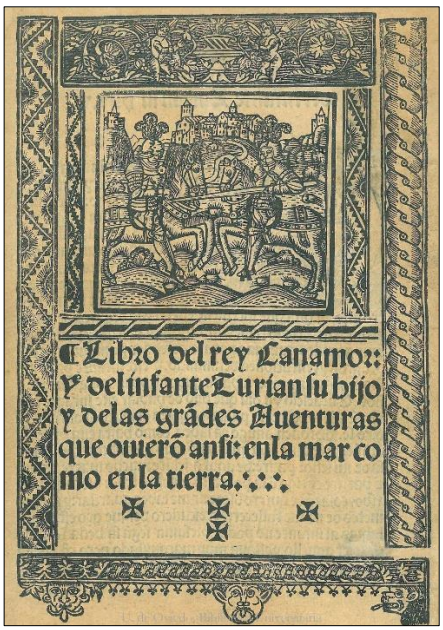


Fig. 8. Portada del *Libro del Rey Canamor* (Valencia, 1527)

El grabado, que representa un combate de dos caballeros con una ciudad de fondo con una pieza, se encuentra dentro de un marco formado por cuatro piezas xilográficas y encima del título.

A pesar de la distinción en cuanto a la *mise en page* entre los libros de caballerías y las ediciones que se corresponden con las historias breves de caballerías, en todos los casos Transkribus ha delimitado correctamente el *layout*. De hecho, el reconocimiento de los grabados de portada e interiores, en el caso de los libros de caballerías, ha sido omitido de manera automática; mientras que las capitales xilográficas –debido a su composición artística–, no han sido reconocidas por el programa y se han excluido manualmente de la delimitación de líneas.

Recapitulación letra gótica

En conclusión, realizado el recorrido por estos textos de acuerdo con los principios descriptivos de la bibliografía, observamos lo siguiente: las características editoriales y bibliográficas de los dieciséis impresos en letra gótica que constituyen el *dataset* del modelo son muy variadas, pero no afectan de forma significativa al reconocimiento del texto. Podríamos dividir estos textos impresos en folio (*Lisuarte de Grecia, Doctrinal de los Caballeros, La Fiameta, Crónica del Rey Don Rodrigo, Retablo de la Vida de Cristo, Florando de Inglaterra, Silves de la Selva* y *Leandro el Bel*) o en cuarto (*Partinuplés, Magalona, Reina Sebilla, Tragicomedia de Calisto y Melibea*); en ediciones con el texto presentado a doble columna, el cual plantea más problemas por la segmentación de la imagen (*Lisuarte de Grecia*, *La Fiameta, Crónica del Rey Don Rodrigo, Retablo de la Vida de Cristo, Florando de Inglaterra, Silves de la Selva* y *Leandro el Bel*) o a línea tirada (*Partinuplés, Magalona, Reina Sebilla, Tragicomedia de Calisto y Melibea, Doctrinal de los Caballeros*); o por contener grabados internos (*Lisuarte, Tragicomedia de Calisto y Melibea*) o prescindir de ellos. Si bien, dicha casuística no supone ningún problema insalvable en el manejo de Transkribus, solamente aparecen dificultades cuando las digitalizaciones son defectuosas o de baja calidad, o el ejemplar está dañado.

Apéndice 2. Modelo de HTR SpanishRedonda_extended_sXVI-XVII

Descripción Dataset:
Tipo de documentos: impresos
Nr. de palabras: 61 938
Nr. de líneas: 7 675
CER Training Set: 0.21%
CER Validation Set: 1.07%

Autores:
Stefano Bazzaco (coord.), Gaetano Lalomia, Daniela Santonocito, Manuel Garrobo Peral, Mónica Martín Molares, Carlota Cristina Fernández Travieso

Versión actualmente disponible: versión 1.0.0 (julio 2021)

Cómo citar: Stefano Bazzaco (coord.), Gaetano Lalomia, Daniela Santonocito, Manuel Garrobo Peral, Mónica Martín Molares, & Carlota Cristina Fernández Travieso (2021). HTR model SpanishRedonda_XVI-XVII_extended DATASET (1.0.0) [Data set]. Zenodo.
DOI: <https://doi.org/10.5281/zenodo.4889218>

Enlaces:
https://github.com/stefanobazzaco/HTR-model-SpanishRedonda_XVI-XVII_extended
https://zenodo.org/record/4889218#.YmUUF9pBw2x
https://readcoop.eu/model/spanish-redonda-round-script-16th-17th-century/

Al igual que los dieciséis impresos seleccionados para el corpus en gótica, para la creación del *dataset* que constituye la base del entrenamiento del modelo *SpanishRedonda* en el programa Transkribus se empleó una quincena de impresos, que podemos dividir en dos grupos.

Por un lado, se identifican cinco obras de carácter histórico-caballeresco, publicadas entre 1578 y 1607. Todas ellas escritas en verso con una disposición textual variada (ya sea en una o dos columnas). Hacemos referencia a dos obras de Cristóbal de Mesa: el libro primero de la *Restauración de España* (impreso en Madrid por Juan de la Cuesta en 1607) y el argumento del canto primero de *Las navas de Tolosa* (Madrid, viuda de P. Madrigal, 1594). A estas dos sumamos la *Historia de las hazañas y hechos del invencible caballero Bernardo del Carpio*, de Agustín Alonso (Toledo, Pero López de Haro, 1585); el canto primero de *Las lágrimas de Angélica*, de Luis Barahona de Soto (Granada, Hugo de Mena, 1586) y el canto primero de la segunda parte del *Libro del Orlando determinado que prosigue la materia de Orlando el Enamorado*, compuesto por Don Martín de Bolea y Castro (Lérida, Miguel Prats, 1578).

Por otro lado, constituye el grueso del corpus una decena de relaciones de sucesos, sobre las que nos detendremos en el apartado siguiente para comentar sus características bibliográficas. Y es que, con motivo de poder ofrecer la transcripción de estos textos en acceso abierto en el portal del proyecto *Biblioteca Digital Siglo de Oro (BIDISO)*[46], se consideró fundamental la creación de un modelo de transcripción en redonda. Para ello, se fue elaborando un corpus documental sobre distintas colecciones vinculadas con las fuentes primarias de las bibliotecas digitales del proyecto BIDISO.

Relaciones de sucesos

Entendemos por *relaciones de sucesos* «aquellos documentos noticieros que solían versar sobre asuntos muy diversos y cuya forma era variada. Podían ser manuscritas o impresas, estar en verso o prosa, y constar de un

[46] Se pretende, además, que estos textos transcritos sean enriquecidos con marcación formal y semántica, a través de una codificación en XML-TEI, para la creación de ediciones académicas digitales. Véase, en este mismo número, Fernández Travieso y Garrobo Peral (2022).

solo pliego o, incluso, llegar a tener las dimensiones de un libro voluminoso»[47]. Por tanto, se evidencia ya en su propia definición el carácter tan heterogéneo, y a su vez complejo, que presenta este género editorial[48]. Como vemos, se hace mención a distintos aspectos: la temática variada, la modalidad del discurso o la extensión e, incluso, la forma de difusión (que, si bien puede ser manuscrita o impresa, por razones obvias al estudiar una tipografía concreta, nos centramos solo en esta última).

Habida cuenta de las particularidades del género, se buscó una selección lo más representativa posible de estos textos (Tabla 6). De ahí que se hayan empleado relaciones de sucesos con diferente temática, impresos por distintos tipógrafos en talleres ubicados en emplazamientos variados –peninsulares, como Madrid, Valencia, Sevilla o Cuenca, y extranjeros como Roma, Bruselas o Lima– y que abarcasen un abanico de años lo más amplio posible.

Tipología

Buena parte de los acontecimientos festivos, políticos y sociales de la Edad Moderna fueron plasmados en los impresos noticieros para dejar memoria de lo ocurrido. En este corpus de relaciones que hemos empleado (Tabla 6), vemos cómo destaca una tipología: las relaciones de ceremonias o festejos. De estas conservamos relaciones más extensas, que daban lugar a los llamados *libros de fiestas*. Aunque parezca una obviedad, tal particularidad influye en la configuración misma de los volúmenes, tanto en el contenido como en la forma. Debido a esto, y ante las necesidades iniciales de proveer a Transkribus del mayor número de páginas posible para poder crear el modelo, optamos por servirnos de este subgénero caracterizado por una extensión mayor principalmente por ser «recomendable transcribir [al menos] una veintena de páginas manualmente» (Bazzaco, 2020, 548).

[47] Información disponible en el portal BIDISO <https://www.bidiso.es/estaticas/ver.htm?id=6> (cons. 15/05/2022). Para una aproximación a las Relaciones de Sucesos (RdS), véanse Pena Sueiro (2001) y Pena Sueiro y Ruiz Astiz (2019).
[48] Infantes (1996, 208).

Título	Año	Lugar de impresión	Tipología	Form.	Columnas	CBDRS
Relación de la solemne entrada hecha en Ferrara a los 13 días de noviembre MDXCVIII por la serenísima Margarita de Austria	1598	Roma. Nicolás Mucio	Ceremonias y festejos. Entrada	4º 12 h.	1 (prosa)	0001160 B
Relacion del aparato que se hizo en la ciudad de Valencia para el recebimiento de la serenissima reina doña Margarita de Austria	1599	Valencia. Pedro Patricio Mey	Ceremonias y festejos. Entrada	8º 16 h.	1 (prosa)	0002620 A
Relación del nacimiento del nuevo infante y de la muerte de la reina nuestra señora	1612	Cuenca. Salvador Viader	Ceremonias y festejos. Nacimiento y exequias	4º 2 h.	2 (verso)	0002764 A
Relación verdadera del acompañamiento y bautismo de la serenísima princesa Margarita María Catalina	1623	Madrid. Diego Flamenco	Ceremonias y festejos. Bautismo	Folio 2 h.	1 (prosa)	0004060 B
Fiesta que se hizo en Aranjuez a los años del Rey Nuestro Señor	1623	Madrid. Juan de la Cuesta	Ceremonias y festejos.	4º 26 h.	1 (prosa)	0007066 A
Relación verdadera en que se da cuenta de todo el daño que causó las crecientes del rio Guadalquivir en la ciudad de Sevilla y Triana	1626	Lima. Gerónimo de Contreras	Suceso extraordinario de la naturaleza	Folio 2 h.	1 (prosa)	0007022 A
Relacion de las fiestas que se han hecho en la fidelissima Ciudad de Napoles por el nazimiento del Principe N.S. […], hasta cinco de Março de este año de 1658	[1658]	[s.l.] [s.n.]	Ceremonias y festejos. Nacimiento	4º 12 h.	1 (prosa)	0007308 A
Primera parte de la relación de las reales disposiciones […] jornada a la provincia de Guipuzcoa a entregar a la serenísima	1660	Sevilla. Juan Gómez de Blas	Acontecimiento político. Relaciones de viajes	4º 4 h.	1 (prosa)	0003549 A
Segunda parte de la relación diaria del itinerario que su Majestad ha seguido desde que salió de Madrid hasta llegar a Fuenterrabía	[1660]	Sevilla. Juan Gómez de Blas	Acontecimiento político. Relaciones de viajes	4º 4 h.	1 (prosa)	0002944 A
Relación de un nuevo milagro obrado por intercesión del glorioso apóstol de las Indias, san Francisco Xavier	1663	Bruselas	Suceso extraordinario. Milagro	4º 4 h.	1 (prosa)	0003722 A

Tabla 6. Listado de relaciones de sucesos empleadas para el modelo *SpanishRedonda*

En la selección aparecen cuatro libros de fiestas, es decir, cuatro relaciones que exceden el volumen de un pliego de cordel: la *Relación de la solemne entrada hecha en Ferrara [...]*, de Ioan Paolo Mocante (Roma, Nicolás Mucio, 1598)[49]; la *Relación del aparato que se hizo en la ciudad de Valencia […]*, de Juan Bautista Confalionero (Valencia, Patricio Mey, 1599)[50]; la *Fiesta que se hizo en Aranjuez a los años del rey nuestro señor don Felipe IIII [...],* de don Antonio de Mendoza (Madrid, Juan de la Cuesta, 1623)[51] y la *Relación de las fiestas que se han hecho en la fidelísima ciudad de Nápoles por el nacimiento del príncipe […]* ([1658])[52]. En total, estaríamos hablando de 101 páginas, lo que supone alimentar a Transkribus con 3249 líneas con solo estas cuatro ediciones.

Portada y grabados

Además, son justo «estos libros [de fiestas], más que ninguna otra clase de relaciones, [los que] suelen adornarse con ilustraciones de grabados xilográficos o calcográficos» (López Poza, 1999, 220). Constatan esta afirmación las portadas de las ediciones que transcribimos (Fig. 9).

Generalmente, en ellas se incluían elementos heráldicos (reales, papales o nobiliarios), dependiendo del editor o promotor en algunos casos, o bien los motivos xilográficos que poseía un impresor en su taller. Así, podemos ver en la portada de la izquierda el escudo papal y el real, en medio el del Reino de Valencia y, a la derecha, el escudo de Felipe IV (Figs. 9a y 9b). La edición sobre las fiestas de Nápoles no incorpora ningún grabado, pero, por la disposición de la portada, no parece descabellado sospechar que podría haberse reservado el espacio inferior de la misma para incluir uno (Fig. 9c).

---

[49] Ejemplar conservado en la Biblioteca Valenciana, sign. XVI/F-33. Disponible en: <https://bivaldi.gva.es/es/catalogo_imagenes/grupo.cmd?presentacion=pagina&posicion=1&path=1004644&registrardownload=0&texto_busqueda=&interno=S> (cons. 25/04/2022).

[50] Ejemplar conservado en la British Library, sign. 9930.aa.9. Disponible en: <https://www.bl.uk/treasures/festivalbooks/pageview.aspx?strFest=0142&strPage=001> (cons. 25/04/2022).

[51] Ejemplar conservado en la Biblioteca Nacional de España, sign. R/15515. Disponible en: <http://bdh-rd.bne.es/viewer.vm?id=0000052047&page=1> (cons. 25/04/2022).

[52] Ejemplar conservado en la Biblioteca Nacional de España, sign. VE/1558/11. Disponible en: <https://archive.org/details/relaciondelasfie00napl/mode/2up> (cons. 25/04/2022).


RELACION
DE LA SOLENE ENTRADA HECHA EN FERRARA
a los 13. dias de Noviembre M. D. XCVIII.
Por la Sereniſſima S. Doña Margarita de Auſtria Reyna de Eſpaña, y del Conſiſtorio publico con todos los aparatos que ſu S.S.N. Clemente Papa VIII. mando hazer y hizo para tal affecto.
Con particular auiſo de la Miſſa Pontifical que canto ſu Beatitud, y de las Ceremonias de los deſpoſorios que ſe hizieron en la Iglefia Cathedral de dicha Ciudad Domingo alos 15. del meſmo con la ceremonia de la Roſa que ſu S. acabada la Miſſa dio alla Reyna.
Compueſta por Ioan Paolo Mecante Vno de los Maeſtros de ceremonias de la Capilla de ſu S. en lengua Italiana, y traducida, e interpretada, en lengua Eſpañola por el Doctor Andres Catalan natural de la Ciudad de Toro.
CON LICENCIA DE LOS SVPERIORES.
EN ROMA.
Impreſſa por Nicolas Mucio. Año M. D. XCVIII.
En plaça Capranica.

RELACION DEL
APARATO QVE SE
HIZO EN LA CIVDAD DE
Valencia para el recebimiento de la Sereniſſima Reyna Doña Margarita de Auſtria deſpoſada con el Catholico y potentiſſimo Rey de Eſpaña Don Phelipe Tercero deſte nombre.
Vendenſe en caſa de Franciſco Miguel a la calle de Caualleros.

RELACION
DELAS FIESTAS
Que ſe han hecho en la Fideliſsima Ciudad de Napoles.
POR EL NAZIMIENTO DEL
PRINCIPE N.S.
Que DIOS guarde,
Haſta cinco de Março de eſte año de 1658.

Fig. 9. Portadas de algunas RdS incluidas en el corpus

Las portadas de las demás relaciones constan de un encabezado a modo de título. Este podía ocupar desde tres líneas, como en el caso de la *Relación verdadera del acompañamiento y bautismo de la serenísima princesa Margarita María Catalina* (Madrid, Diego Flamenco, 1623)[53] hasta casi media plana, como en la *Relación de un nuevo milagro obrado por intercesión del glorioso apóstol de las Indias, san Francisco Xavier, en 2 de septiembre de 1662* (Palermo, 1663)[54]. En estos encabezados suele destacarse tipográficamente –generalmente con el empleo de mayúsculas– algunos elementos textuales del título y se puede incluir otra información sobre la edición como el pie de imprenta (lugar de impresión, nombre del impresor y/o del costeador, fecha), los datos legales (mención del privilegio, de la licencia, de la tasa, etc.) u otras menciones o alusiones (mención de edición, autor, traductor). Así, en el ejemplo de la relación sobre el milagro de san Francisco Javier, se añade: *en Palermo de Sicilia, aprobado por el ilustrísimo arzobispo de dicha ciudad. Según la copia italiana, impresa en Palermo el*

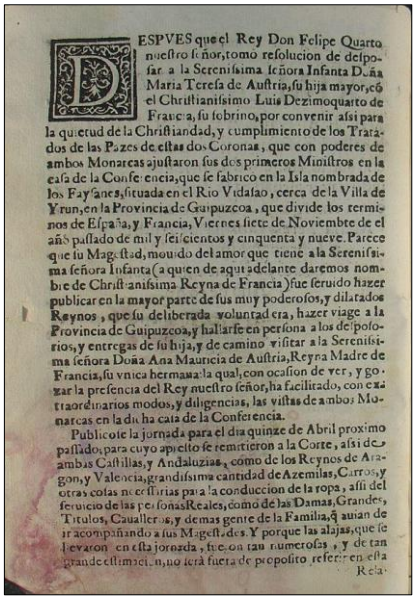
[53] Ejemplar conservado en el fondo antiguo de la Biblioteca de la Universidad de Sevilla, sign. A 109/085(033). Disponible en: <https://archive.org/details/A109085109> (cons. 25/04/2022). Puede consultarse la edición crítica anotada y un estudio sobre las tres ediciones de esta relación en Martín Molares (2021).

[54] Ejemplar conservado en el fondo antiguo de la Biblioteca de la Universidad de Sevilla, sign. A 111/025(17). Disponible en: <https://archive.org/details/A11102517> (cons. 25/04/2022).

mes de agosto de 1663 y sacada de la copia francesa, impresa en Bruselas a 5 de septiembre de dicho año.

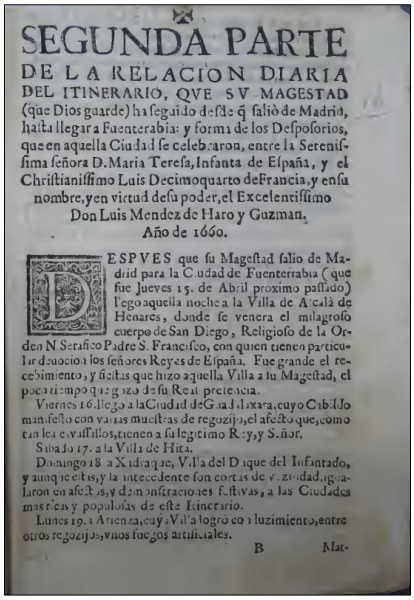
No tan frecuente es el caso de la *Primera parte de la relación de las reales disposiciones y majestuosos aparatos con que su Majestad […] se ha servido hacer jornada a la provincia de Guipúzcoa, a entregar a la serenísima señora doña María Teresa Bibiana de Austria, su hija, al cristianísimo Luis decimocuarto de Francia, su esposo* (Sevilla, Juan Gómez de Blas, 1660). En esta breve relación, de cuatro hojas en formato 4.º, se reserva el primer recto a la portada. En ella se incluye el título ya citado, el escudo real xilográfico, la mención a la licencia y, tras un filete, el pie de imprenta. Se entiende que pueda deberse al reclamo editorial por tratarse de una relación festiva sobre un acontecimiento monárquico tan relevante como la unión de Luis XIV con María Teresa de Austria. Sin embargo, la *Segunda parte de la relación diaria* […], en línea con lo anteriormente mencionado, ocupa aproximadamente un tercio del primer recto.

En el interior de las relaciones de sucesos con las que se ha trabajado no encontramos otros grabados. Solo adornaban algunos textos las letras capitulares xilográficas al inicio del impreso. Cinco de las diez relaciones, todas ellas político-festivas, añaden este adorno tipográfico. Por ejemplo, las dos relaciones sevillanas –la primera y la segunda parte del viaje para la entrega de la novia– comienzan por la misma letra (*D*), por lo que el impresor Juan Gómez de Blas utilizó el mismo grabado xilográfico (Fig. 10).



DESPVES que el Rey Don Felipe Quarto nuestro señor, tomo resolucion de desposar a la Serenissima señora Infanta Doña Maria Teresa de Austria, su hija mayor, cõ el Christianissimo Luis Dezimoquarto de Francia, su sobrino, por convenir assi para la quietud de la Christiandad, y cumplimiento de los Tratados de las Pazes de estas dos Coronas, que con poderes de ambos Monarcas ajustaron sus dos primeros Ministros en la casa de la Conferencia, que se fabrico en la Isla nombrada de los Faysanes, situada en el Rio Vidasao, cerca de la Villa de Yrun, en la Provincia de Guipuzcoa, que divide los terminos de España, y Francia, Viernes siete de Noviembre de el año passado de mil y seiscientos y cinquenta y nueve. Parece que su Magestad, mouido del amor que tiene a la Serenissima señora Infanta (a quien de aqui adelante daremos nombre de Christianissima Reyna de Francia) fue seruido hazer publicar en la mayor parte de sus muy poderosos, y dilatados Reynos, que su deliberada voluntad era, hazer viage a la Provincia de Guipuzcoa, y hallarse en persona a los desposorios, y entregas de su hija, y de camino visitar a la Serenissima señora Doña Ana Mauricia de Austria, Reyna Madre de Francia, su vnica hermana: la qual, con ocasion de ver, y gozar la presencia del Rey nuestro señor, ha facilitado, con extraordinarios modos, y diligencias, las vistas de ambos Monarcas en la dicha casa de la Conferencia.

Publicose la jornada para el dia quinze de Abril proximo passado, para cuyo apresto se remitieron a la Corte, assi de ambas Castillas, y Andaluzias, como de los Reynos de Aragon, y Valencia, grandissima cantidad de Azemilas, Carros, y otras cosas necessarias para la conduccion de la ropa, assi del seruicio de las personas Reales, como de las Damas, Grandes, Titulos, Caualleros, y demas gente de la Familia, q̃ auian de ir acompañando a sus Magestades. Y porque las alajas, que se llevaron en esta jornada, fueron tan numerosas, y de tan grande estimacion, no será fuera de proposito referir en esta Rela-



SEGUNDA PARTE
DE LA RELACION DIARIA
DEL ITINERARIO, QVE SV MAGESTAD (que Dios guarde) ha seguido desde q̃ salió de Madrid, hasta llegar a Fuenterabia: y forma de los Desposorios, que en aquella Ciudad se celebraron, entre la Serenissima señora D. Maria Teresa, Infanta de España, y el Christianissimo Luis Decimoquarto de Francia, y en su nombre, y en virtud de su poder, el Excelentissimo Don Luis Mendez de Haro y Guzman.
Año de 1660.

DESPVES que su Magestad salio de Madrid para la Ciudad de Fuenterrabia (que fue Jueves 15. de Abril proximo passado) llego aquella noche a la Villa de Alcalà de Henares, donde se venera el milagroso cuerpo de San Diego, Religioso de la Orden N. Serafico Padre S. Francisco, con quien tienen particular deuocion los señores Reyes de España. Fue grande el recebimiento, y fiestas que hizo aquella Villa a su Magestad, el poco tiempo que gozo de su Real presencia.

Viernes 16. llego a la Ciudad de Guadalaxara, cuyo Cabildo manifesto con varias muestras de regozijo, el afecto que, como tan leal vassallos, tienen a su legitimo Rey, y Señor.

Sabado 17. a la Villa de Hita.

Domingo 18. a Xidraque, Villa del Dique del Infantado, y aunque estas, y la antecedente son cortas de vezindad, signalaron en afectos, y demonstraciones festivas, a las Ciudades mas ricas, y populosas de este Itinerario.

Lunes 19. a Arienza, cuya Villa logro con luzimiento, entre otros regozijos, vnos fuegos artificiales.

B Mar-

Fig. 10. Inicial grabada en dos relaciones seriadas

Formato y extensión

La mayor parte de los textos transcritos –siete de las diez relaciones– presentan un formato en 4.º, muy frecuente en las hojas o pliegos sueltos. No sorprende, por tanto, que la extensión de estos impresos sea breve: una relación de dos hojas, tres de cuatro hojas, dos de doce hojas y solo una tiene una extensión superior al ocupar 26 hojas.

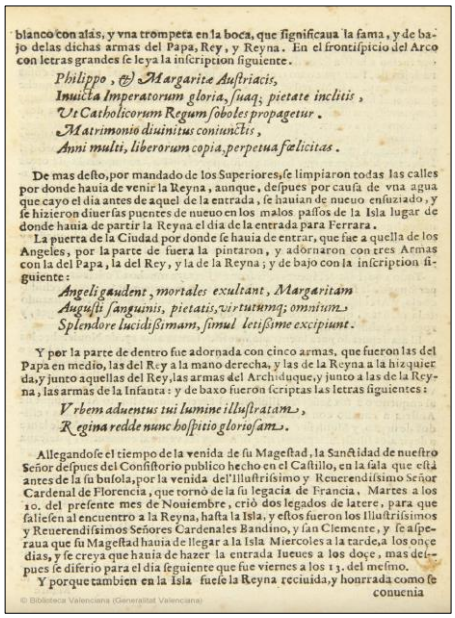
Las otras tres relaciones restantes, que ocupan dos hojas, están en formato folio (*Relación verdadera del […] bautismo de la serenísima princesa Margarita María Catalina* y la *Relación verdadera en que se da cuenta de todo el daño que causó las crecientes del rio Guadalquivir en la ciudad de Sevilla y Triana*)[55], y solo una en formato 8.º (*Relación del aparato que se hizo en la ciudad de Valencia para el recibimiento de la serenísima reina doña Margarita de Austria*)[56].

[55] Ejemplar conservado en la Biblioteca Nacional de España, sign. VE/59/64. Disponible en <http://bdh-rd.bne.es/viewer.vm?id=0000075249&page=1> (cons. 25/04/2022).

[56] Esta distribución podríamos entender que es también representativa de los formatos más empleados para este género editorial. Así, el *Catálogo y Biblioteca Digital de Relaciones de Sucesos (CBDRS)* devuelve resultados similares en cuanto a la preferencia de los formatos: en 4.º se registran 2 475 ediciones; en formato folio, 1 420 ediciones; en 8.º, 191 y en 12.º, 13 ediciones.

Disposición del texto

En cuanto a la disposición del texto, la mayor parte de las relaciones escogidas son en prosa y solo una es en verso: la *Relación del nacimiento del nuevo infante y de la muerte de la reina nuestra señora* (Cuenca, Salvador Viader, 1612)[57]. No obstante, en las relaciones, una u otra forma no eran excluyentes puesto que pueden aparecer combinadas. De este modo, en alguno de los volúmenes impresos –por ejemplo, en los libros de fiesta– se introducen en la narración los distintos versos que se recitaban en las justas o certámenes, así como las composiciones que adornaban los elementos efímeros.



blanco con alas, y vna trompeta en la boca, que significaua la fama, y de baxo delas dichas armas del Papa, Rey, y Reyna. En el frontispicio del Arco con letras grandes se leya la inscription siguiente.

Philippo, & Margaritæ Austriacis,
Inuicta Imperatorum gloria, suaq; pietate inclitis,
Vt Catholicorum Regum soboles propagetur.
Matrimonio diuinitus coniunctis,
Anni multi, liberorum copia, perpetua fœlicitas.

De mas desto, por mandado de los Superiores, se limpiaron todas las calles por donde hauia de venir la Reyna, aunque, despues por causa de vna agua que cayo el dia antes de aquel de la entrada, se hauian de nueuo enxuziado, y se hizieron diuersas puentes de nueuo en los malos passos de la Isla lugar de donde hauia de partir la Reyna el dia de la entrada para Ferrara.

La puerta de la Ciudad por donde se hauia de entrar, que sue a quella de los Angeles, por la parte de fuera la pintaron, y adornaron con tres Armas con la del Papa, la del Rey, y la de la Reyna; y de bajo con la inscription siguiente:

Angeli gaudent, mortales exultant, Margaritam
Augusti sanguinis, pietatis, virtutumq; omnium,
Splendore lucidissimam, simul lætissime excipiunt.

Y por la parte de dentro sue adornada con cinco armas, que fueron las del Papa en medio, las del Rey a la mano derecha, y las de la Reyna a la hizquierda, y junto aquellas del Rey, las armas del Archiduque, y junto a las de la Reyna, las armas de la Infanta: y de baxo fueron scriptas las letras siguientes:

Vrbem aduentus tui lumine illustratam,
Regina redde nunc hospitio gloriosam.

Allegandose el tiempo de la venida de su Magestad, la Sanctidad de nuestro Señor despues del Consistorio publico hecho en el Castillo, en la sala que está antes de la su bulola, por la venida del'Illustrissimo y Reuerendissimo Señor Cardenal de Florencia, que tornó de la su legacia de Francia, Martes a los 10. del presente mes de Nouiembre, crió dos legados de latere, para que saliesen al encuentro a la Reyna, hasta la Isla, y estos fueron los Illustrissimos y Reuerendissimos Señores Cardenales Bandino, y san Clemente, y se aspe-raua que su Magestad hauia de llegar a la Isla Miercoles a la tarde, a los onze dias, y se creya que hauia de hazer la entrada Iueues a los doze, mas despues se diferio para el dia siguiente que sue viernes a los 13. del mesmo.

Y porque tambien en la Isla sue sue la Reyna reciuida, y honrrada como se conuenia

Fig. 11. Ejemplo de RdS con alternancia de prosa y verso

Lo que sí evidenciamos en este ejemplo es que se emplean también distintas letras dentro de la misma relación (redonda y cursiva), así como lenguas diferentes. Por ejemplo, los fragmentos en latín de las ceremonias

[57] Ejemplar conservado en la Biblioteca Nacional de España, sign. R/12676. Disponible en: <http://bdh-rd.bne.es/viewer.vm?id=0000061653&page=1> (cons. 25/04/2022).

religiosas (Fig. 11).

Por último, y a diferencia de los libros de caballerías en gótica, las relaciones no suelen tener titulillos en los que se incluyan los nombres de los epígrafes del capítulo. Solo encontramos un caso, el de las *Fiestas de Aranjuez*, en donde en los rectos encabeza el titulillo «de Aranjuez» y en los versos aparece «Fiestas». Además, es el único ejemplo del texto que presenta apostillas marginales.

En definitiva, la heterogeneidad de un género como las relaciones de sucesos ofrece múltiples posibilidades de estudio bibliográfico, por lo que parece ser el contexto ideal para la explotación de herramientas de reconocimiento de textos.

§

Bibliografía citada

Allés Torrent, Susanna, «Tiempos hay de acometer y tiempos de retirar: literatura áurea y edición digital», *Studia Aurea*, 11 (2017), pp. 13-30.

Baranda, Nieves, «Compendio bibliográfico sobre narrativa caballeresca breve», en *Evolución narrativa e ideológica de la literatura caballeresca*, ed. M.ª Eugenia Lacarra, Bilbao, Servicio Editorial de la Universidad del País Vasco, 1991, pp. 183-191.

Bazzaco, Stefano, «El Progetto Mambrino y las tecnologías OCR: estado de la cuestión», *Historias Fingidas*, 6 (2018), pp. 257-272. DOI: <https://doi.org/10.13136/2284-2667/89> (cons. 13/05/2022).

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

—, «Experimentos de estilometría en el ámbito de los libros de caballerías. El caso de atribución de un original italiano: *Il terzo libro di Palmerino d'Inghilterra* (Portonari, 1559)», *Actas de la Asociación Hispánica de Literatura Medieval*, 2022, en prensa.

Bognolo, Anna y Stefano Bazzaco, «Tra Spagna e Italia: per un'edizione digitale del Progetto Mambrino», *eHumanista/IVITRA*, 16 (2019), pp. 20-36.

Burnard, Lou; O'Brian O'Keeffe, Katherine; Unsworth, John (eds.), *Electronic Textual Editing*, New York, MLA, 2006.

Calvo-Tello, José, *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*, Transcript, Verlag, 2021.

Causer, Tim; Terras, Melissa, «"Many hands make light work. Many hands together make merry work": Transcribe Bentham and crowdsourcing manuscript collections», en *Crowdsourcing our Cultural Heritage*, Ashgate, Farnham, 2014, pp. 57-88.

Floridi, Luciano, *The 4th Revolution. How the Infosphere is Reshaping Human Reality*, Oxford, Oxford University Press, 2014.

Franzini, Greta; Kestemont, Mike; Rotari, Gabriela; Jander, Melina; Ochab, Jeremi K.; Franzini, Emily; Byszuk, Joanna; Rybicki, Jan, «Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm», *Frontiers in Digital Humanities*, 5 (2018), s.p.

García-Reidy, Alejandro, «Deconstructing the Authorship of *Siempre ayuda la verdad*: a play by Lope de Vega?», *Neophilologus*, 103 (2019), 493-510.

Gifford Fenton, Eileen; Duggan, Hoyt N., «Effective methods of producing machine-readable text from manuscript and print sources», en *Electronic Textual Editing*, eds. Lou Burnard, Katherine O'Brian O'Keeffe y John Unsworth, New York, MLA, 2006, pp. 241-261.

González-Sarasa Hernáez, Silvia, *Tipología editorial del impreso antiguo español*, Tesis doctoral, dir. Fermín de los Reyes Gómez, Universidad Complutense de Madrid, 2013.

Hernández-Lorenzo, Laura, «Poesía áurea, estilometría y fiabilidad: métodos supervisados de atribución de autoría atendiendo al tamaño de las muestras», *Caracteres: estudios culturales y críticos de la esfera digital*, n. 8/1 (2019), pp. 189-229.

Infantes, Víctor, «La narración caballeresca breve», en *Evolución narrativa e ideológica de la literatura caballeresca,* ed. María Eugenia Lacarra, Bilbao, Servicio Editorial Universidad del País Vasco, 1991, pp. 165-182.

—, «La prosa de ficción renacentista: entre los géneros literarios y el 'género editorial'», en *Actas del X Congreso de la Asociación Internacional de Hispanistas*, ed. Antonio Vilanova, Barcelona, PPU, 1992, vol. 1, pp. 467-474.

—, «¿Qué es una relación?: divagaciones varias sobre una sola divagación», in *Las «Relaciones de sucesos» en España (1500-1750). Actas del primer Coloquio Internacional (Alcalá de Henares, 8, 9 y 10 de junio de 1995)*, eds. María Cruz García de Enterría, Henry Ettinghausen, Víctor Infantes de Miguel y Agustín Redondo, Alcalá de Henares, Editorial Universidad de Alcalá - Publications de la Sorbonne, 1996, pp. 203-216.

—, «El *género editorial* de la narrativa caballeresca breve», *Voz y letra. Revista de literatura,* 7/2 (1996), pp. 127-132.

—, «La tipología de las formas editoriales», en *Historia de la edición y de la lectura en España 1472-1914*, dir. Víctor Infantes, François López y Jean-François Botrel, Madrid, Fundación Germán Sánchez Ruipérez, 2003, pp. 39-49.

Italia, Paola, *Editing 2000. Per una filologia dei testi digitali*, Roma, Salerno Editrice, 2020.

Kichuk, Diana, «Quantità e qualità dei testi online: il problema della digitalizzazione di massa», en *Teoria e forme del testo digitale*, ed. Michelangelo Zaccarello, Roma, Carocci Editore, 2019, pp. 135-166.

López Poza, Sagrario, «Las peculiaridades de las relaciones festivas en forma de libro», en *La fiesta. Actas del II Seminario de Relaciones de Sucesos (A Coruña, 1998)*, ed. Sagrario López Poza y Nieves Pena Sueiro, A Coruña, Sociedad de Cultura Valle Inclán, 1999, pp. 213-222.

Lucía Megías, José Manuel, *Imprenta y libros de caballerías*, Madrid, Ollero & Ramos, 2000.

Mancinelli, Tiziana, «Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work», *Historias Fingidas*, 4 (2016), pp. 255-260. DOI: <https://doi.org/10.13136/2284-2667/65> (cons. 13/05/2022).

Martín Molares, Mónica, «El bautismo de la princesa Margarita María Catalina de Austria (1623): tres ediciones de una relación», en *Buenas noticias. Relaciones de sucesos en los siglos XVI-XVIII: estudios y textos*, eds. Gabriel Andrés y Sandra M.ª Peñasco González, Pesaro, Metauro Edizioni, coll. Ispanica urbinate n. 4, 2021, pp. 65-89.

Mordenti, Raul, *Informatica e critica dei testi*, Roma, Bulzoni, 2001.

Moretti, Franco, *Graphs, Maps, Trees: Abstract Models for a Literary History*, Londres - Nueva York, Verso, 2005.

—, *Falso movimento. La svolta quantitativa nello studio della letteratura*, Milano, Nottetempo, 2022.

Mühlberger, Günter *et al.*, «Transforming scholarship in the archives through Handwritten Text Recognition. Transkribus as a case study», *Journal of Documentation - Emerald Publishing*, 75/5 (2019), pp. 954-976.

Narang, Sonika Rani; Jindal, M. K.; Kumar, Munish, «Ancient text recognition: a review», *Artificial Intelligence Review*, 53 (2020), pp. 5517-5558.

Orlandi, Tito (ed.), *Il problema della formalizzazione*, Roma, Accademia Nazionale dei Lincei, 1994.

Pena Sueiro, Nieves, «Estado de la cuestión sobre el estudio de las Relaciones de sucesos», *Pliegos de Bibliofilia*, 13/1 (2001), pp. 43-66.

Pena Sueiro, Nieves; Ruiz Astiz, Javier, «Las relaciones de sucesos: producto y género editorial en la Monarquía Hispánica», *Memoria y Civilización. Anuario de Historia*, 22 (2019), pp. 371-380.

Pierazzo, Elena, *Digital scholarly editing: Theories, models and methods*, Aldershot, Ashgate, 2015.

Reul, Christian; Springmann, Uwe; Wick, Christoph; Puppe, Frank, «State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open-Source Engines», *ArXiv e-prints*, 2018. URL: <https://arxiv.org/abs/1810.03436> (cons. 24/03/2022).

Rockwell, Geoffrey; Passarotti, Marco, «The Index Thomisticus as a Digital Humanities Big Data Project», *Umanistica Digitale*, 5 (2019), pp. 13-34. DOI: <http://doi.org/10.6092/issn.2532-8816/8575> (cons. 13/05/2022).

Roncaglia, Gino, «Google Book Search e le politiche di digitalizzazione libraria», *DigItalia web. Rivista del Digitale nei Beni Culturali*, 2 (2009), pp. 17-35.

—, *La quarta rivoluzione. Sei lezioni sul futuro del libro*, Roma-Bari, Laterza, 2010.

Rosselli del Turco, Roberto; di Pietro, Chiara; Martignago, Chiara, «Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo», *Umanistica Digitale*, 7 (2019), pp. 5-21. DOI: <http://doi.org/10.6092/issn.2532-8816/9322> (cons. 13/05/2022).

Shillingsburg, Peter L., *From Gutenberg to Google. Electronic Representations of Literary Texts*, Cambridge, Cambridge University Press, 2006.

Smith, David A., Ryan Cordell, *A Research Agenda for Historical and Multilingual Optical Character Recognition*, NULab, Northeastern University, 2018.

Terras, Melissa, «The Rise of Digitization: An Overview», en *Digital Libraries*, ed. Rico Rukowski, Olanda, Sense Publishers, 2010, pp. 3-20.

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Humanidades Digitales y literatura medieval española: la integración de Transkribus en la base de datos COMEDIC

Nuria Aranda García

(École Normale Supérieure de Lyon)*

Abstract

La base de datos COMEDIC (Catálogo de obras medievales impresas en castellano) nace en 2012 con el objetivo de analizar la recepción y transmisión de los textos literarios medievales en la imprenta quinientista. Los avances en la aplicación de herramientas de las Humanidades Digitales y las sucesivas renovaciones del proyecto lo han conducido al empleo de estos útiles en la investigación que desarrolla. En el presente artículo se hace un estado de la cuestión y una proyección sobre los objetivos que se pretenden conseguir mediante la aplicación del programa Transkribus.
Palabras clave: COMEDIC; Humanidades Digitales; Transkribus; ediciones digitales académicas; literatura medieval española

The database COMEDIC (Catálogo de obras medievales impresas en castellano) was created in 2012 with the aim of analysing the reception and transmission of medieval literary texts in the 16th century print. Advances in the application of Digital Humanities tools and the successive renewals of the project have led it to apply these tools to its research data. This paper offers a state of the art and a projection of the goals to be achieved in the application of the Transkribus software.
Keywords: COMEDIC; Digital Humanities; Transkribus; Digital Scholar Editions; Spanish Medieval Literature

§

* Este trabajo se ha realizado en el marco del Proyecto de Investigación PID2019-104989GB-I00, financiado por MCIN/AEI/10.13039/501100011033, y se inscribe en el grupo investigador CLARISEL, que cuenta con la participación económica del Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón.

Introducción

Las Humanidades Digitales debieron su nacimiento al deseo por potenciar la divulgación y visibilidad de los resultados científicos de esta disciplina mediante el apoyo proporcionado por el medio digital. Este no solo les ha servido para ganar reconocimiento frente a otras disciplinas en el panorama investigador actual, sino que las ha empujado a la interdisciplinariedad y a la formación de equipos de trabajo donde se dan la mano especialistas de estudios diversos y donde es necesaria la comunicación con grupos de informáticos (Spence, 2014). Pese a que estas reflexiones, sencillas en su conceptualización teórica, han comenzado a aplicarse en la práctica en muchos trabajos, las Humanidades Digitales apenas han comenzado a despertar en nuestro país, aunque el balance que puede realizarse hasta el momento es bastante positivo e indica que caminamos hacia la buena dirección (González Blanco-García, 2013; López Poza, 2015)

En el ámbito hispánico, lejano queda aparentemente en el tiempo BOOST *(The Bibliography of Old Spanish Texts)*, actualmente la base de datos *PhiloBiblon*, considerado como primer proyecto de Humanidades Digitales orientado al tratamiento de textos en español, y desarrollado por Charles Faulhaber y Francisco Marcos Marín en la década de los 70 en el seno del *Hispanic Seminar of Medieval Studies* de la Universidad de Wisconsin[1]. Las Humanidades comenzaron a incorporar herramientas informáticas en España en la década de los 80, momento en que puede situarse la labor de los pioneros de nuestro país, cuyos trabajos sirvieron como punto de partida para el desarrollo de los primeros proyectos especialmente en los departamentos de filología (Canet, 2014). El incremento de subvenciones por parte de organismos públicos e instituciones privadas potenció las Humanidades Digitales en la década de los 90, y esto se tradujo en una proliferación de trabajos cuya subsistencia y sostenibilidad se debió y se debe a esta financiación y a la proporcionada a los proyectos dentro del

[1] Son muchos los trabajos que actualmente ofrecen un estado de la cuestión sobre las Humanidades Digitales en España, sirvan como ejemplo Rojas Castro (2013), Spence y González-Blanco García (2014), López Poza y Pena Sueiro (2014), Morrás y Rojas Castro (2015), López Poza (2015; 2019) Toscano, Rabadán, Ros y González-Blanco García (2020) y Hernández Lorenzo (2020).

programa estatal de I+D+I (Toscano, Rabadán, Ros y González-Blanco García, 2020). Sirvan como ejemplos representativos de esta etapa inicial ADMYTE de Charles Faulhaber y Francisco Marcos Marín (1992) (Faulhaber y Marcos Marín, 1992) y el portal LEMIR de José Luis Canet (1995) (Canet, 2014, 12). A todo lo anterior se sumaron los fondos FEDER, que promovieron una colaboración interdisciplinar entre investigadores de distintas áreas humanísticas y la incorporación de grupos de informáticos (López Poza, 201, 138) que culminó en la creación de bases de datos de autores y obras, la edición y estudio de determinados géneros y la creación de bibliotecas digitales de corpus específicos que en tiempos más recientes han incorporado el lenguaje de marcado (López Poza, 2015: 3).

Con la llegada del nuevo milenio, y fundamentalmente a partir de 2005 y 2006, comienzan ya a proliferar las publicaciones con teorizaciones sobre esta nueva disciplina y sobre la edición de textos electrónicos y el desarrollo de ediciones digitales (Hernández Lorenzo, 2020, 566-568), mientras la etiqueta «Humanidades Digitales» ya queda perfectamente acuñada en España. Sustituta de la lejana «Informática humanística», que respondía a la inseguridad de estos investigadores hacia los aspectos técnicos, el nuevo sintagma pone en primer lugar a la figura del humanista que debe adquirir las destrezas de las nuevas tecnologías (López Poza, 2019, 127). Desde entonces, las Humanidades Digitales han seguido un proceso de consolidación hasta institucionalizarse en 2011, momento en que el panorama inicial de falta de comunicación entre proyectos afines e iniciativas inconexas se transforma gracias a la sucesión de eventos y encuentros relacionados con las Humanidades Digitales que culminan en la creación de la asociación *Humanidades Digitales Hispánicas. Sociedad Internacional* (HDH)[2]. Ahora las publicaciones van a centrarse en aplicaciones concretas del medio digital en los estudios humanísticos, va a potenciarse la formación en Humanidades Digitales mediante su inclusión en los planes de estudios y van a sucederse los encuentros con especialistas de este campo (Hernández Lorenzo, 2020, 572-579). Proliferan las bases de datos de autores, obras y géneros en el ámbito literario y también

[2] <https://humanidadesdigitaleshispanicas.es/> (cons. 05/11/2021). Las actas se publicaron en López Poza y Pena Sueiro (2014).

aparecen en línea las primeras ediciones que incorporan lenguajes de marcado como el XML-TEI[3].

La base de datos COMEDIC en las Humanidades Digitales españolas

La base de datos COMEDIC, acrónimo de «Catálogo de obras medievales impresas en castellano», nace en el año 2012, precisamente dentro de este contexto de mayor impulso de las Humanidades Digitales y de incremento del número de bases de datos. Fruto de la iniciativa de unos profesores del Departamento de Filología Española de la Universidad de Zaragoza pertenecientes al grupo de investigación consolidado CLARISEL, financiado por el Gobierno de Aragón, se desarrolla gracias a la concesión de un proyecto de investigación I+D+I dependiente del por aquel entonces Ministerio de Economía y Competitividad dentro del VI Plan Nacional de Investigación Científica, Desarrollo e Innovación[4]. El objetivo que perseguían era claro: la realización de un catálogo que recogiese todas las obras literarias medievales que han sobrevivido gracias a la imprenta, originales y/o traducidas, sin excluir los textos con una notable tradición manuscrita pero dejando a un lado la «literatura gris». El nexo de unión de estos investigadores era estudiar la relación entre la literatura medieval y la renacentista, concentrándose en la producción estrictamente castellana tanto dentro como fuera de la Península, lo que llevó a la incorporación de dos investigadores internacionales[5]. La premisa de la que se partió fue la no existencia de una ruptura entre la literatura medieval y la producida a partir de 1500[6].

[3] Resulta pionero el trabajo de edición digital de las *Soledades* de Góngora realizado por Rojas Castro (2015; 2017)

[4] En el equipo inicial se encontraban María Jesús Lacarra, al frente del proyecto, Juan Manuel Cacho Blecua, Alberto del Rio, María Carmen Marín Pina, José Aragüés Aldaz, María Sanz Julián y la doctoranda María Coduras Bruna.

[5] Estos fueron Amaia Arizaleta (Université de Toulouse 2-Jean Jaurés) y Gaetano Lalomia (Università degli Studi di Catania). La primera se ocuparía de las obras en castellano impresas en Francia y el segundo de aquellas estampadas en territorio italiano.

[6] Véanse al respecto las reflexiones que ya habían realizado Whinnom (1967) y Simón Díaz (1988).

Se sentaban así las bases para el inicio del trabajo, esto es, el estudio de la difusión, evolución, transformación y recepción de la literatura medieval española en el Quinientos, y la posibilidad de analizar la supremacía de algunos géneros literarios en detrimento de otros, así como las manipulaciones que sobre los textos operaron los tipógrafos para orientarlos a los nuevos lectores e introducirlos en los nuevos paradigmas de lectura (Santonocito, 2013). En relación con el *corpus*, cronológicamente se encuadró entre 1470, fecha en la que ve la luz la primera obra en castellano salida de una imprenta hispana, el *Sacramental* de Clemente Sánchez de Vercial, y 1600, momento en que se produjo el cambio hacia una recepción más moderna de los textos. El principal requisito que estos debían cumplir era haber conocido existencia previa a 1500.

El proyecto fue presentado ese mismo año en la *SEMYR* (Sociedad de Estudios Medievales y Renacentistas) y en el *Seminario de Estudios sobre Narrativa Caballeresca* de la Universidad Nacional Autónoma de México y, al año siguiente, en la *Asociación Hispánica de Literatura Medieval*. Los inicios de la creación de la base de datos fueron complejos, y se comenzó el trabajo sobre fichas creadas en *FileMaker* que contenían los principales campos que se incluirían, para después contar con una interfaz alojada en los servidores web de la Universidad de Zaragoza que permitiría hacer visibles los resultados al mismo tiempo que garantizaba a los investigadores seguir trabajando sobre las fichas. Además, el análisis de los textos no era solo textual o material, sino que implicaba un planteamiento multidisciplinar al prestar atención a las ilustraciones e imágenes, a su recepción y a los elementos paratextuales que acompañan a las imágenes, a lo que se sumaba la dificultad añadida de la lenta respuesta de algunas bibliotecas a la hora de proveer las digitalizaciones. Dos metas se perseguían: incrementar el número de fichas disponibles y aumentar la visibilidad de la base de datos (Lacarra, 2019, 421)[7].

Para ello fue fundamental su integración en la red ARACNE en 2017,

[7] El equipo del proyecto contaba ya con una experiencia previa en el uso de bases de datos y herramientas informáticas, puesto que desde finales de los 90 han sacado adelante tres bases de datos bibliográficas: *Amadís*, sobre literatura caballeresca, *Sendebar*, sobre cuentística, y *Heredia*, sobre literatura aragonesa medieval. Están agrupadas en CLARISEL. Bases de datos bibliográficas, <https://clarisel.unizar.es/> (cons. 15/10/2021). Se suma la puesta en línea de DINAM. Diccionario de nombres del ciclo amadisiano, <http://dinam.unizar.es/> (cons. 15/10/2021).

la primera red de Humanidades Digitales y Letras Hispánicas en España y un hito en la investigación en Humanidades. Como iniciativa, ARACNE surge en julio de 2011 en el seno del *I Seminario Internacional sobre Bibliotecas Digitales y Bases de Datos especializadas para la investigación en Literaturas Hispánicas (BIDESLITE)* organizado por el Grupo de Estudios de Prosa hispánica bajomedieval y renacentista, dirigido por aquel entonces por Mercedes Fernández Valladares, con la colaboración del proyecto DIALOGYCA-BDDH y la ayuda del Instituto Universitario Menéndez Pidal[8]. El debate que se suscitó tomaba como punto de partida la dispersión que presentaban los proyectos de investigación en la aplicación de los conceptos y herramientas propuestos por las Humanidades Digitales puesto que, pese a tratarse de proyectos diversos en contenidos, objetivos, metodología, géneros literarios y periodos temporales, se habían enfrentado a problemas muy similares (Baranda Leturio y Rodríguez López, 2014, 101-102). La finalidad fue, entonces, intentar dar respuesta a la necesidad de coordinación y unificación de criterios percibida por los equipos de investigación de algunos proyectos de literatura española, precisamente ante una falta de unanimidad en la utilización de estas nuevas herramientas digitales (Pena Sueiro, 2017, 408)[9]. En 2012 consigue financiación bajo la dirección de Pedro Ruiz Pérez[10] y se encamina hacia tres líneas diferentes: la consecución de una mayor visibilidad y difusión de los trabajos realizados, que esa visibilidad fuese un elemento de identificación y desarrollo de líneas de investigación con base tecnológica de calidad y la conveniencia de generar instrumentos de enlace y mejora en la recuperación de los datos y el desarrollo de herramientas que permitiesen proporcionar información de calidad para la creación de otras herramientas en el área (Baranda Leturio y Rodríguez López, 2014, 101).

El resultado ha sido la creación por parte del Laboratorio de Bases

[8] Las actas de este encuentro están disponibles en: <https://webs.ucm.es/BUCM/blogs/Foliocomplutense/4104.php> (cons. 16/10/2021).

[9] Los seis proyectos fundadores de ARACNE fueron BIESES (UNED), BSF (UDC), BIDISO (UDC), CLARISEL (UZ), DIALOGYCA (UCM) y PHEBO (UCO).

[10] Ministerio de Economía y Competitividad (Gobierno de España), referencia de proyecto FFI2011-15606-E. El equipo inicial de trabajo estuvo formado por Pedro Ruiz e Ignacio García Aguilar (UCO), Nieves Baranda y María Dolores Martos (UNED), Ana Vian, Consolación Baranda y Mercedes Fernández Valladares (UCM), José Luis Villacañas Berlanga (UCM), Juan Manuel Cacho Blecua y María Jesús Lacarra y María Carmen Marín Pina (UZ), Sagrario López Poza y Nieves Pena Sueiro (UDC).

de Datos de la Universidade da Coruña de un portal que facilita el acceso a catorce recursos de investigación. Sobre este portal se ha diseñado un buscador de metadatos estándar Dublin Core que facilita una consulta simultánea de todos los contenidos ofrecidos por los distintos proyectos que lo componen gracias al protocolo OAIPMH (*Open Archive Initiative-Protocol for Metadata Harvesting*), que permite que el contenido de las colecciones contenidas en Aracne sea visible en recolectores virtuales como Hispana o Europeana (Arrigoni y Rodríguez López, 2014, 250-251)[11]. La incorporación de la base de datos COMEDIC no fue sino una continuación por parte de estos investigadores de la línea seguida por sus bases de datos CLARISEL, en los orígenes de la red, y un paso más hacia la difusión de los contenidos investigadores. El punto culminante llegaría en 2019 con la concesión por parte del Ministerio de la Red de Excelencia ARACNE NODUS, bajo la dirección de Nieves Pena Sueiro, que pretende seguir con la búsqueda de una mayor visibilidad de los resultados de sus colecciones (Pena Sueiro, 2017, 411)[12].

Fruto de renovaciones posteriores del proyecto, la base de datos ha ido implementándose, así como ha aumentado progresivamente el número de obras actualmente visibles para los investigadores, que se cuenta en 83. La interfaz que se presenta al usuario por cada una de las fichas que la componen es sencilla: se recogen de manera general las variantes del título que presenta la obra en todas sus ediciones conservadas, así como los segundos autores, destinatarios o comisionarios, las fechas de redacción y/o traducción y los testimonios manuscritos conservados. Se da cuenta también de las ediciones modernas de la obra, los testimonios de lectura y, punto especialmente trascendental, la reescritura que sufre esta precisamente con motivo de su tradición impresa. Una sección destinada a la bibliografía cierra el conjunto. En el interior de las fichas se abre un indicador por cada edición documentada

[11] Sobre su descripción y funcionamiento, ver Alvite Díez y Pena Sueiro (2020).

[12] <https://www.red-aracne.es/presentacion> (cons. 11/11/2021). Ministerio de Ciencia, Innovación e Universidades, referencia de proyecto RED2018-102755-T. La concesión de la Red de Excelencia vino de la mano de la incorporación de cinco nuevos grupos, cuatro de perfil humanista y uno tecnológico. Estos son: BeCLar (dir. Antonio Moreno Hernández, UNED), PARNASEO (dir. Marta Haro Cortés, UV), Progetto Mambrino (dir. Stefano Neri, UNIVR), SILEM (dir. Carlos María Collantes Sánchez, US) y el LBD (dir. Ángeles Saavedra Places, UDC).

como fiable. De estas se refieren los repertorios en los que aparecen recogidas, tanto en formato papel como en formato digital; los ejemplares conservados, incluidos los digitalizados en bibliotecas digitales[13]; los facsímiles, en el caso en que los haya; y, finalmente, los paratextos entendidos según el concepto de Genette (2001): legales, socioliterarios, editoriales y grabados interiores y de portada[14].

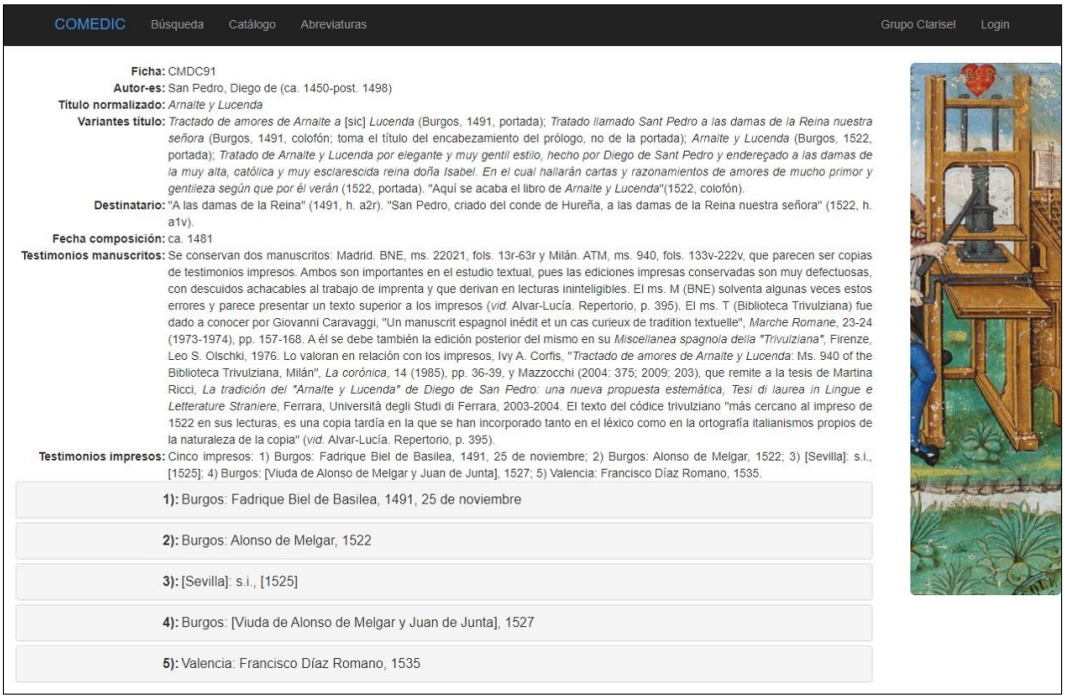


Fig. 1. COMEDIC. Ficha de *Arnalte y Lucenda.* Autora: María Carmen Marín Pina

Una última implementación se ha llevado a cabo, estrechamente relacionada con uno de los objetivos a corto plazo que la red ARACNE se estableció en su fundación: el «diseño de un modelo de criterios de evaluación de calidad técnica y científica de los productos resultantes, con la finalidad de homogeneizar protocolos y propiciar su aceptación para el

[13] Está disponible también para los usuarios un manual donde se recogen las principales referencias abreviadas empleadas en los repertorios bibliográficos citados y las principales bibliotecas patrimoniales y digitales donde se encuentran depositados los ejemplares.
[14] Véase Cacho Blecua (2014) y Lacarra (2017).

reconocimiento y evaluación de los trabajos realizados» (Baranda y Rodríguez López, 2014, 106). La ausencia de criterios que permitan la evaluación de la calidad investigadora presente en la elaboración de materiales científicos dentro de las Humanidades Digitales por parte de las agencias estatales[15] ha intentado solventarse en cada una de las entradas de la base de datos mediante la asignación de un DOI y un ISSN que las haga evaluables[16].

Transkribus y COMEDIC: una propuesta de integración

Transkribus y los modelos de trabajo

Con la llegada del nuevo milenio se incrementó el número de tecnologías avanzadas destinadas al tratamiento de imágenes digitalizadas, triunfando los *softwares* de reconocimiento de caracteres, gracias a este aumento de la financiación desde los gobiernos nacionales y autonómicos, los proyectos a nivel internacional y europeo y la intervención de instituciones y fundaciones privadas. Los OCR (*Optical Character Recognition*) se definen como sistemas de conversión de *inputs* en forma de texto en un formato codificado por la máquina. Frecuentes en *softwares* comerciales u *Open Source,* dependen principalmente de la extracción de rasgos y la discriminación y/o clasificación de estos rasgos basados en una serie de patrones (Memon, Sali *et al.,* 2020). Los sistemas de reconocimiento de caracteres son herramientas específicamente creadas para la obtención de transcripciones a partir de imágenes escaneadas de las fuentes en papel, y aplicados sobre una imagen de una página impresa, proceden a fragmentarla en caracteres que se comparan con conjuntos de

[15] «Igualmente parece poco razonable que quien pone al servicio de la comunidad el fruto de su trabajo publicándolo en acceso abierto enseguida (lo cual puede ayudar a acelerar otras investigaciones) sea menos valorado que el que publica un libro cuya distribución es deficiente. Es chocante —y delata ignorancia— que, en las evaluaciones de producción científica, se siga pidiendo como condición que una publicación lleve un número de ISBN, como si ello fuera un marchamo de calidad (cuando no es más que un control de ventas que cualquiera puede solicitar a la Federación de Gremios de Editores de España) y no se midan otros parámetros más científicos» (López Poza, 2015, 5).

[16] Lucía Megías (2010, 1253) aboga precisamente por acercarse al modelo de patentes propuesto por las disciplinas científicas pero adaptado al contexto de las Humanidades.

rasgos abstractos que describen caracteres prototípicos aprendidos previamente a partir de un entrenamiento sobre las tipografías. Este se basa en metologías de Aprendizaje Automático (*Machine Learning*) y Aprendizaje Profundo (*Deep Learning*) que se aplican primero sobre material preparado y anotado. Se trata de un proceso que funciona especialmente bien con tipografías modernas, puesto que las similitudes entre las fuentes aprendidas y reconocidas, la clara separación entre caracteres uniformes sobre un fondo blanco impoluto y una ortografía moderna y estandarizada contribuyen al buen reconocimiento (Springman y Lüdeling, 2017). Su empleo reciente sobre caligrafía manuscrita ha recibido bastante atención, especialmente por su contribución a la digitalización de manuscritos medievales (Memon, Sali *et al.,* 2020).

Con todo, los sistemas OCR no han dejado de estar exentos de riesgos y problemáticas. Las manchas en las páginas de los testimonios, debidas a la humedad o al deterioro por el paso del tiempo, sumadas a las deformaciones procedentes del escaneo manual o del estado defectuoso de las fuentes empleadas han dificultado la aplicación del reconocimiento sobre ciertos soportes (Springman y Lüdeling, 2017). Esto afecta directamente al tratamiento de textos anteriores a la época contemporánea, especialmente a las obras impresas en el siglo XVI, centuria en la que se inscribe el corpus de obras trabajado en COMEDIC, donde convivían diferentes tipografías (gótica, redonda, cursiva) que podían adquirir incluso variaciones internas dependiendo de los distintos juegos de tipos y fundiciones de cada impresor[17]. Se suman otras dificultades como el uso frecuente de abreviaturas, el empleo de signos tironianos, las ligaduras entre caracteres y factores externos como los desgastes de los tipos y los defectos de entintado (Bazzaco, 2020, 545). Además, las tecnologías OCR imponen una correspondencia directa entre el signo ortográfico y una letra en el texto transcrito, lo que obliga a entrenar un único modelo individual para un libro individual y una tipografía concreta. Este mecanismo, que podía funcionar para el libro o texto que había servido de entrenamiento, se volvía impracticable al aplicarlo a otros diferentes puesto que, aunque la fuente resultase similar

[17] Al respecto, ver también Barbuti y Caldarola (2013) y Berk-Kirpatrick y Klein (2014).

para el ojo humano, los resultados arrojados no eran positivos (Sprigmann y Lüdeling, 2017). Para paliar este defecto, se han implementado las plataformas de HTR (*Handwritten Text Recognition*), donde se integra Transkribus, nacidas para la lectura e interpretación de textos manuscritos, pero aplicables también a textos impresos de esta primera época de implantación de la imprenta en nuestro país. Esta plataforma, surgida del Proyecto READ (*Recognition and Enrichment of Archival Documents*) dentro del programa europeo Horizon 2020[18], se basa en la colaboración entre usuarios que proveen los materiales digitalizados para generar *inputs* mientras un equipo de informáticos se encargan del procesamiento de los datos. Dichos materiales han tenido que pasar por un tratamiento previo de la imagen para asegurar su calidad y garantizar el correcto reconocimiento.

El trabajo que pretende realizarse en COMEDIC se integra en los avances realizados por el Progetto Mambrino en la aplicación de Transkribus. Con el objetivo de elaborar ediciones digitales académicas de las obras que componen el ciclo italiano del *Amadís de Gaula,* comenzaron los primeros proyectos experimentales en el año 2016 con la tipografía cursiva que caracteriza a estos impresos (Mancinelli, 2016; Bazzaco, 2018). El objetivo ahora es la elaboración de un modelo de reconocimiento o *extended model* de letra gótica que sea aplicable a cualquier texto impreso con esta tipografía a lo largo del Quinientos. La creación de este modelo permitiría a los usuarios en un futuro transcribir cualquier tipo de texto en gótica de este periodo con un reducido margen de error a la hora de lanzar el reconocimiento automático. Para ello, se partió de la selección de un pequeño corpus de textos que presentasen cierta homogeneidad en su conformación tipográfica, cierta extensión y que poseyesen estándares de calidad en su digitalización en formato imagen. Además, como rasgo definitorio, estos debían pertenecer a impresores y lugares diferentes, estar estampados en tipografía gótica y datarse en un periodo temporal amplio que finalmente ha abarcado desde 1526 hasta 1563. El listado de obras con el que empezaron a trabajar fue el siguiente (Bazzaco, 2020, 549):

[18] <https://readcoop.eu/> (cons. 8/10/2021). Se celebran encuentros anuales donde investigadores ponen en común sus resultados en el empleo de la herramienta, ver por ejemplo el celebrado en 2020: <https://readcoop.eu/transkribus-user-conference-2020> (cons. 29/10/2021).

Título	Autor	Impresor(es)	Localización
Lisuarte de Grecia	Juan Díaz	Jacobo y Juan Cromberger Sevilla, 1526	BNE R/71
Florando de Inglaterra	Anónimo	German Gallarde Lisboa, 1545	BL C.62.h.14.
Silves de la Selva	Pedro de Luján	Dominico de Robertis Sevilla, 1549	BNE R/865
Leandro el Bel	Pedro de Luján	Miguel Ferrer Toledo, 1563	BNE R/9030

Tabla 1. Obras que constituyen el *dataset* del trabajo realizado por el Progetto Mambrino

De esta manera, se inició el proceso de aprendizaje automático o *Machine Learning* (Carbonell, Michalski *et al.*, 2013) mediante la creación de unos algoritmos que sirvieran de base (o *groundtruth*) que buscasen una correlación entre los signos presentes en la imagen y sus respectivas transcripciones que permitiese llegar a la creación de dicho modelo extendido (Jander, 2016). Los resultados iniciales arrojaron buenos augurios, lo que llevó a la necesidad de plantear una aproximación colaborativa y un enriquecimiento de los materiales de base usados para el entrenamiento, pues solo mediante el incremento de transcripciones posteriores podía llegarse a implementar un reconocimiento más fiable de los textos sobre los que se aplicaba el modelo (Bazzaco, 2020, 552).

Es en esta etapa donde COMEDIC se suma a la iniciativa, añadiendo para la creación del modelo extendido o *extended model* de gótica los siguientes títulos a modo de textos de entrada (Tabla 2).

La peculiaridad de los textos propuestos por nuestra base de datos, además de mostrar una gran variedad de tipos pertenecientes a diferentes impresores del Quinientos, radica fundamentalmente en una *dispositio textus* a línea tirada, frente a la doble columna de los textos caballerescos sugeridos por Progetto Mambrino para el modelo de gótica. Se añaden además otras particularidades de puesta en página, como las que presentan la *Tragicomedia de Calisto y Melibea* (Roma, Marcellus Silber, s.a.) o el *Retablo de la vida de Cristo* de Juan de Padilla (Sevilla, Juan Cromberger, 1510)[19], con el objetivo de enriquecer el *Layout Analysis* (análisis de la *mise en page*).

[19] Remito para ello a la contribución de Bazzaco *et al.* (pp. 67-125) en este mismo volumen.

Título	Autor	Lugar de edición y fecha
Historia de la linda Magalona	Anónimo	Sevilla, Jacobo Cromberger, 1519
Historia de la reina Sebilla	Anónimo	Burgos, Felipe de Junta, 1551
Historia del rey Canamor	Anónimo	Valencia, Jorge Costilla, 1527
Libro del conde Partinuplés	Anónimo	Sevilla, Jacobo Cromberger, 1519
Libro del conde Partinuplés	Anónimo	Burgos, Herederos de Juan de Junta 1558
Libro del conde Partinuplés	Anónimo	Burgos, Felipe de Junta, 1563
Doctrinal de los Caballeros	Alonso de Cartagena	Burgos, Fadrique Biel de Basilea, 1487
La Fiameta	Juan Boccaccio	Salamanca, [Impresor de la Gramática de Nebrija], 1497
Crónica del Rey Don Rodrigo (Crónica Sarracina)	Pedro de Corral	[Sevilla], [Meinardo Ungut y Estanislao Polono], 1499
Tragicomedia de Calisto y Melibea	Fernando de Rojas	Roma, Marcellus Silber, a. 1515
Retablo de la Vida de Cristo	Juan de Padilla	Sevilla, Jacobo Cromberger, 1505

Tabla 2. Obras que constituyen el *dataset* del trabajo realizado por los colaboradores de COMEDIC

En relación con los criterios de transcripción, la propuesta de futuro pasaba por trabajar sobre dos modelos distintos de reconocimiento, uno basado en transcripciones más conservadoras, y otro basado en criterios de transcripción más basado en criterios de transcripción con un mayor grado de modernización (Bazzaco, 2020, 561). Finalmente, nos hemos inclinado por el primero, que facilita un patrón de reconocimiento con un índice de error algo más reducido. Sin embargo, se ha procedido a desarrollar las abreviaturas, que se han indicado mediante un sistema de etiquetado que proporciona la propia plataforma Transkribus. El resultado final ha sido presentado en el congreso *Humanidades Digitales y estudios literarios hispánicos. De los impresos de la Edad Moderna a las ediciones académicas digitales,* celebrado en Verona del 22 al 23 de junio de 2021 bajo la coordinación académica de Stefano Bazzaco.

La integración de Transkribus en el trabajo de COMEDIC

Actualmente nos situamos en lo que ha venido a denominarse como «segunda textualidad», coincidente con el momento en que la escritura se transforma tras surgir en Grecia en el siglo VIII a.C tal y como la conocemos en la cultura occidental, que permite conjugar características de la primera textualidad con algunas características de la oralidad y la introducción del nuevo soporte traído por la informática. Su modelo representativo será el texto digital (Lucía Megías, 2012b, 263-165), definido como un texto en cuyo proceso de difusión interviene una codificación de la información por parte de lenguajes artificiales, y que se presenta materialmente como información lingüística codificada matemáticamente y representada bajo una forma legible, interpretable y reconocible por el ser humano. Por ello, en su composición confluyen dos capas de información: una capa humana continuadora de la tecnología de la escritura, y que supone codificar y descodificar de manera lógica una serie de signos, y una capa de información matemática que entraña el procedimiento que transforma esa escritura en algo legible por la tecnología informática, el conocido como código binario, que le permite nuevos usos y nuevos modelos de relación (Paisão de Sousa, 2009).

A la hora de tratar la concepción del llamado texto digital, que incluye a su vez las posibilidades que ofrece la Web 2.0, entre ellas la interacción con los propios receptores, son imprescindibles los conceptos de «digitalización» o «texto digitalizado», que supone el traslado de una imagen con texto escrito como imagen fija al medio digital y «digitación» o «texto digitado», que incluye necesariamente la intervención de la codificación estándar ASCIII y Unicode propia de un procesador (Bazzaco, 2020, 537). Lucía Megías (2012a, 115-116) ha establecido una división en tres estadios distintos que permiten llegar a este concepto y que suponen una gradación entre los tres aspectos de la digitalización textual teniendo en cuenta la finalidad, la tecnología y la relación con los medios de transmisión analógica, en la que los conceptos anteriores entroncan a la perfección:

1) La reproducción digital de un manuscrito o un impreso por medio de la fotografía digital y el escaneado, la cual supone una fase de

acumulación de información que desarrollan, principalmente, las bibliotecas digitales virtuales. El resultado es un objeto textual «digitalizado».

2) La creación o digitalización de textos con la pretensión de ser difundidos fuera del ambiente y los medios de transmisión digitales, especialmente en el medio impreso. Son los textos generados por los procesadores de textos que cierran el texto en una única imagen y que han sufrido un proceso de «digitación».

3) Texto digital que utiliza procesos de codificación pensados para poder ser visualizados en la pantalla del ordenador aprovechando la hipertextualidad y la posibilidad de establecer relaciones entre el nivel estructural y semántico, y para el que cobran especial importancia los lenguajes de marcado HTML, XML, XHTML o XML-TEI. En esta etapa no se intenta emular en el medio digital los modelos textuales del mundo analógico, sino que lo que se pretende es profundizar en modelos que ofrezcan experiencias variadas a la unión del creador, del lector y del propio medio que ofrece este modelo.

Actualmente el proyecto se encuentra conjugando las dos primeras fases. La recopilación en los indicadores de las fichas de los ejemplares de cada edición que se encuentran digitalizados permite esa acumulación de información que facilita el análisis de este objeto libro casi analógico para proceder a su descripción, al mismo tiempo que ahorra desplazamientos de los investigadores, algo que se han convertido en indispensable desgraciadamente en los tiempos actuales. Mediante el campo «ediciones modernas», consignamos esas ediciones en las que el texto ha pasado un proceso de «digitación» para ver luego la luz de nuevo en formato analógico, si bien cada vez más se ofrece la posibilidad de su consulta en línea en formatos como epub, el propio de los libros electrónicos. El análisis de las ediciones de obras concretas también está propiciando la creación de nuevas ediciones críticas mediante tesis doctorales que van a ver la luz de nuevo también en formato cercano al analógico, si bien los repositorios digitales facilitan su consulta en la red. Todo ello inserta a COMEDIC dentro de las llamadas «bibliotecas de investigación» —«aquellas que intentan aunar el rigor filológico con el temático agrupando normalmente textos de una o varias áreas de conocimiento»

(Canet, 2005, 151)—, pertenecientes al grupo de las que solo incorporan las reproducciones facsímiles y cuyo trabajo está encabezado por un profesor universitario.

Mediante la incorporación de Transkribus pretendemos dar un paso adelante e integrarnos en un grado intermedio entre las fases 2 y 3. En primer lugar, nos centraremos en el reconocimiento de obras de extensión breve que posean un número reducido de ediciones, tomando como punto de referencia la *princeps*, o que tan solo se hallen conservadas en una. Esta transcripción será paleográfica, de acuerdo con los modelos de entrenamiento que en la actualidad hemos desarrollado para el modelo de HTR *SpanishGothic_XV-XVI_extended*, y figurará mediante un hipervínculo en el apartado «ediciones modernas». El hipervínculo conducirá a la transcripción, alojada en la nube, en dos formatos, el pdf en la fase 2 y HTML, ya en la fase 3, mediante la introducción de un lenguaje de marcado capaz de ser reconocido y visualizado por un ordenador y que permite abrir nuevas vías dentro de la edición digital. COMEDIC evolucionaría así hacia una nueva configuración de la «biblioteca de investigación» a través de la publicación del texto y el facsímil, y se gestaría como una hibridación entre una base de datos y una biblioteca digital «que no solo permite realizar búsquedas complejas, sino que también da acceso a textos codificados en HTML y/o formato pdf» (Rojas Castro, 2013, 43). De acuerdo con los criterios que ya rigen la base, cada edición contará con un responsable y un revisor que examine el resultado.

Nuevas vías de futuro que se abren

Editar en la actualidad supone ser consciente del cambio que la sociedad y el mundo académico está viviendo dentro del paradigma digital, y esto implica que el editor digital debe ahora conocer y dominar gran parte de las herramientas informáticas que tiene a su disposición para fijar el texto de manera científica y difundirlo en el nuevo formato (Lucía Megías, 2019, 98). Actualmente nos encontramos en el «incunable del hipertexto», concepto con el que Lucía Megías (2007) pretende establecer una analogía entre estos momentos iniciales del texto digital y los

momentos iniciales del texto impreso. Mostrar una transcripción paleográfica del texto de una obra que no ha sido editada o que no cuenta con ediciones modernas puede servir de punto de partida para elaborar nuevas ediciones críticas analógicas mediante la adaptación del texto a los criterios de edición que el editor quiera aplicar. Además, puede contribuir a la legibilidad de documentos que se han visto dañados o cuya tinta se ha desdibujado, teniendo en cuenta siempre que hay que prestar atención a informaciones a las que no se podrá acceder, como las características físicas del documento, la calidad del papel o las marcas de agua (Ogilvie, 2017, 82).

No obstante, la posibilidad de ofrecer nuestros textos transcritos en lenguaje de marcado HTML abre nuevas vías que encaminan nuestro trabajo a la consecución de los objetivos del texto digital. En primer lugar, facilita la comparación entre la imagen digitalizada y el texto transcrito, pero además permite implementar esta transcripción aplicándole un sistema de marcado semántico XML-TEI potenciando sus posibilidades en la red. Esto permitiría también avanzar hacia la posible creación de una edición sinóptica integral, aquella que permite poder acceder simultáneamente a las transcripciones de todos los testimonios conservados de una obra, así como a sus variantes a partir de la lección de uno de ellos. En segundo lugar, facilita asentar los cimientos que abran el camino a la edición crítica hipertextual. Esta parte de la presentación analógica, pero se sirve del nuevo medio de transmisión al mismo tiempo que conserva el modo habitual de trabajar del filólogo. Mediante el añadido de la hipertextualidad y la hipermedialidad[20], se permitiría el acceso a la obra desde varias aproximaciones: la paleográfica, el análisis lingüístico o el estudio de la evolución textual. Supondría el establecimiento de una relación emergente de transformación donde el texto empieza a modificarse para adaptarse a las características y posibilidades del nuevo medio conjugando las tres morfologías de la información (texto, imagen y sonido) (Lucía Megías, 2009, 12). De esta forma, desembocaría en un entramado hipertextual que pondría en

[20] La hipertextualidad, la hipermedialidad y la intertextualidad han sido definidos como los tres principales rasgos definitorios del ámbito digital y, por lo tanto, deberían estar presentes en una edición digital.

relación las digitalizaciones de los facsímiles conservados, las transcripciones paleográficas, las ediciones críticas del texto, la bibliografía y cualquier información adicional que ofreciera posibilidades de acercamiento a la obra (Lucía Megías, 2019, 109).

Sea como fuere, el texto siempre estará configurado como el elemento central sobre el que se podrá acceder al resto de materiales, que tendrán que estar estrechamente vinculados aprovechando las herramientas informáticas que han aparecido en los últimos años. COMEDIC pretende sembrar la semilla para que una de estas nuevas realidades de estudio y análisis del objeto literario sea posible.

§

Bibliografía citada

Alvite Díez, María Luisa y Nieves Pena Sueiro, «Colecciones digitales patrimoniales especializadas. Estudio de la Red ARACNE», en *Actas del IV Congreso ISKO España-Portugal 2019. XIV Congreso ISKO España (Barcelona, 11 y 12 de julio de 2019)*, eds. Jesús Tramullas, Piedad Garrido-Picazo y Gonzalo Marco Cuenca, Zaragoza, Sociedad Internacional para la Organización del Conocimiento (ISKO)-Capítulo Ibérico, 2020, pp. 185-195.

Arrigoni, Eleonora y Eduardo Rodríguez López, «La red de investigación de "Humanidades Digitales y Letras Hispánicas": avance de la Red-Aracne», en *Visibilidad y divulgación de la investigación desde las Humanidades Digitales: experiencias y proyectos,* coord. Álvaro Baraibar Echeverría, Pamplona, Universidad de Navarra/GRISO (Grupo de Investigación Siglo de Oro), 2014, pp. 243-251. URL: <https://dadun.unav.edu/handle/10171/35723> (cons. 28/10/2021).

Baranda Leturio, Consolación y Eduardo Rodríguez López, «Red ARACNE: retos y objetivos de un proyecto de coordinación en letras hispánicas digitales», en *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, eds. Sagrario López Poza y Nieves Pena Sueiro, *Janus*, Anexo 1 (2014), pp. 101-109. URL: <https://www.janusdigital.es/anexos/contribucion.htm?id=10> (cons. 28/10/2021).

Barbuti Nicola y Tommaso Caldarola, «An Innovative Character Recognition for Ancient Book and Archival Materials: A Segmentation and Self-learning Based Approach», en *Digital Libraries and Archives. IRCDL 2012. Communications in Computer and Information Science,* eds. Maristella Agosti, Floriana Esposito, Stefano Ferilli y Nicola Ferro, Berlin-Heidelberg, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 261-270. DOI: <https://doi.org/10.1007/978-3-642-35834-0_26> (cons. 28/10/2021).

Bazzaco, Stefano, «El Progetto Mambrino y las tecnologías OCR: estado de la cuestión», *Historias fingidas*, 6 (2018), pp. 257-272. DOI: <https://doi.org/10.13136/2284-2667/89> (cons. 15/10/2021).

—, «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. URL: <https://www.janusdigital.es/articulo.htm?id=160> (cons. 15/10/2021).

Berg-Kirkpatrick, Taylor y Dan Klein, «Improved Typesetting Models for Historical OCR», en *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore-Maryland, Association for Computational Linguistics, 2014, pp. 118-123. URL: <https://aclanthology.org/P14-2020.pdf> (cons. 28/10/2021).

Cacho Blecua, Juan Manuel, «Hacia un catálogo de los textos medievales impresos (COMEDIC): el ejemplo de la *Crónica popular del Cid*», en *Texto, edición y público lector en los albores de la imprenta*, eds. Marta Haro Cortés y José Luis Canet, València, PUV, 2014, pp. 29-52.

Canet, José Luis, «Bibliotecas digitales españolas a texto completo», *Syntagma. Revista del Instituto de Historia del Libro y de la Lectura,* 1 (2005), pp. 149-159.

—, «Reflexiones sobre las Humanidades Digitales», en *Janus*, Anexo 1: *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, eds. Sagrario López Poza y Nieves Pena Sueiro, (2014), pp. 11-20. URL: <https://www.janusdigital.es/anexos/contribucion.htm?id=4> (cons. 15/10/2021).

Carbonell, Jamie G, Ryszard S. Michalski, Tom M. Mitchell, «An overview of machine learning», en *Machine Learning: An Artificial Intelligence Approach*, eds. Jamie G. Carbonell, Ryszard S. Michalski y Tom M. Mitchell, Berlin-Heidelberg, Springer-Verlag, 2013, pp. 3-23.

Faulhaber, Charles, y Francisco Marcos Marín, «La conservación y utilización de textos en el futuro inmediato: ADMYTE, el archivo digital de manuscritos y textos españoles», *Hispania* 75/4 (1992), pp. 1010-1023.

Genette, Gérard, *Umbrales,* Madrid, Siglo XXI, 2001.

González-Blanco García, Elena, «Actualidad de las Humanidades Digitales y un ejemplo de ensamblaje poético en la red», *Cuadernos hispanoamericanos,* 761 (2013), pp. 53-67.

Hernández Lorenzo, Laura, «Humanidades Digitales y Literatura española: 50 años de repaso histórico y panorámica de proyectos representativos», *Janus. Estudios sobre el Siglo de Oro*, 9 (2020), pp. 562-595. URL: <https://www.janusdigital.es/articulo.htm?id=154> (cons. 04/11/2021).

Jander, Melina, «Handwritten Text Recognition–Transkribus: A User Report», en *The electronic Text Reuse Acquisition Project (eTRAP)*, Göttingen, Institute of Computer Science, University of Göttingen, 2016. URL: <http://www.etrap.eu/transkribus-a-user-report/> (cons. 05/11/2021).

Karlsson, Lina, y Linda Malm, «Revolution or Remediation? A Study of Electronic Scholary Editions on the Web», *HUMAN IT,* 7/1 (2004), pp. 1-46. URL: <https://humanit.hb.se/article/view/135> (cons. 11/11/2021).

Lacarra, María Jesús, «COMEDIC: un "Catálogo de obras medievales impresas en castellano" en construcción», en *En Doiro antr'o Porto e Gaia. Estudos de Literatura Medieval Ibérica*, ed. J. C. Ribeiro Miranda, Porto, Estratégias criativas, 2017, pp. 599-610.

—, «Comedic», *Historias fingidas,* 7 (2019), pp. 419-42. DOI: < https://doi.org/10.13136/2284-2667/137> (cons. 18/10/2021).

López Poza, Sagrario, «Humanidades Digitales y literaturas hispánicas: presente y futuro», *Ínsula. Revista de letras y ciencias humanas,* 833 (2015), pp. 3-5.

—, «Humanistas y Humanidades Digitales. Trayectoria y proyección en la Filología española», en *Humanidades y humanismo. Homenaje a María Pilar Cuartero*, eds. Aurora Egido, José Enrique Laplana y Luis Sánchez Laílla, Zaragoza, Institución Fernando el Católico, 2019, pp. 125-160.

López Poza, Sagrario y Nieves Pena Sueiro (eds.), *Janus*, Anexo 1: *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, (2014). URL: <https://www.janusdigital.es/anexo.htm?id=5> (cons. 03/11/2021).

Lucía Megías, José Manuel, «La edición crítica hipertextual. Hacia la superación del incunable del hipertexto», en *Lecturas y textos en el siglo XXI. Nuevos caminos en la edición textual*, coords. Cristina Castillo Martínez y José Luis Ramírez Luengo, Lugo, Editorial Axac, 2009, pp. 11-74.

—, «Los nuevos filólogos del siglo XXI: La literatura medieval hispánica en la Red», en *Actas del XIII Congreso de la Asociación Hispánica de Literatura Medieval. In memoriam Alan Deyermond (Valladolid, 15 a 19 de septiembre de 2009)*, eds. José Manuel Fradejas Rueda, Déborah Dietrick Smithbauer, Demetrio Martín Sanz, María Jesús Díez Garretas, Valladolid, Universidad de Valladolid, 2010, pp. 1233-1254.

—, *Elogio del texto digital. Claves para entender el nuevo paradigma,* Madrid, Fórcola Ediciones, 2012a.

—, «El escritor en la era digital (un elogio a la segunda textualidad)», *Revista de Humanidades y Ciencias Sociales,* 2 (2012b), 255-269. URL: <http://rabida.uhu.es/dspace/bitstream/handle/10272/6332/El_escritor_en_la_era_digital.pdf?sequence=2> (cons. 11/11/2021).

—, «El editor de textos ante el reto digital: elogio de la edición 2.0», *Revista de Humanidades Digitales,* 4 (2019), 93-114. URL: <http://revistas.uned.es/index.php/RHD/article/view/25188> (cons. 13/11/2021).

Mancinelli, Tiziana, «Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work», *Historias Fingidas*, 4 (2016), pp. 255-260. DOI: <https://doi.org/10.13136/2284-2667/65> (cons. 13/05/2022).

Memon, Jamshed, Maira Samid, Rizwan Ahmed Khan y Mueen Uddin, «Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)», *IEEE Access*, 8 (2020). URL: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9151144> (cons. 05/11/2021).

Morràs, María y Antonio Rojas Castro (coords.), *Humanidades Digitales y Literaturas hispánicas, Ínsula*, 822 (2015).

Ogilvie, Brian, «Scientific Archives in the Age of Digitization», *Isis*, 107/1 (2016), pp. 77-85. URL: <https://www.journals.uchicago.edu/doi/full/10.1086/686075> (cons. 03/11/2021).

Paisão de Sousa, Maria Clara, «Conceito material de "Texto digital": um ensaio», *Revista Texto Digital*, 5/2 (2009), pp. 159-187. URL: <http://www.textodigital.ufsc.br/> (cons. 11/11/2021).

Pena Sueiro, Nieves y Ángeles Saavedra Places, «Aracne. Red de Humanidades Digitales y Letras Hispánicas», *Historias Fingidas,* 7 (2019), pp. 407-412. DOI: <https://doi.org/10.13136/2284-2667/149> (cons. 03/11/2021).

Rojas Castro, Antonio, «El mapa y el territorio: una aproximación histórico-bibliográfica a la emergencia de las Humanidades Digitales en España», *Caracteres. Estudios culturales y críticos de la esfera digital*, 2/2 (2013), pp. 10-53.

—, *Editar las soledades de Góngora en la era digital: texto crítico y propuesta de codificación XML-TEI,* tesis doctoral, Barcelona, Universitat Pompeu Fabra, 2015.

—, «La edición crítica digital y la codificación TEI. Preliminares para una nueva edición de las *Soledades* de Luis de Góngora», *Revista de Humanidades Digitales,* 1 (2017), pp. 4-19. URL: <http://revistas.uned.es/index.php/RHD/article/view/16379> (cons. 03/11/2021).

Santonocito, Daniela, «Reescrituras y relecturas: hacia un catalogo de obras medievales impresas en castellano hasta 1600 (COMEDIC)», *Le forme e la storia*, VI/1 (2013), pp. 175-187.

Simón Díaz, José, «La literatura medieval castellana y sus ediciones españolas de 1501 a 1560», en *El libro antiguo español. Actas del Primer Coloquio Internacional (Madrid, 18 al 20 de diciembre de 1986)*, eds. María Luisa López-Vidriero y Pedro M. Cátedra, Universidad de Salamanca/Biblioteca Nacional de Madrid/Sociedad Española de Historia del Libro, Salamanca, 1988, pp. 371-396.

Spence, Paul, «Centros y fronteras: el panorama internacional», en Sagrario López Poza y Nieves Pena Sueiro (eds.), *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, *Janus*, Anexo 1 (2014), pp. 37-61. URL: <https://www.janusdigital.es/anexos/contribucion.htm?id=6> (cons. 03/11/2021).

Spence, Paul y Elena González-Blanco García, «A Historical Perspective to Digital Humanities in Spain», *H-Soz-Kult* (2014). URL: <https://www.hsozkult.de/debate/id/diskussionen-2449> (cons. 04/11/2021).

Springmann, Uwe y Anke Lüdeling, «OCR of historical printings with an application to building diachronic corpora», *Digital Humanities Quarterly*, 7/2 (2017). URL: <http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html#> (cons. 24/10/2021).

Toscano, Maurizio, Aroa Rabadán, Salvador Ros y Elena González-Blanco García (2020), «Digital humanities in Spain: Historical perspective and current scenario», *Profesional de la Información*, 29/6 (2020).

Whinnom, Keith, «Spanish Literary Historiography: Three Forms of Distortion», en *An Inaugural Lecture Delivered in the University of Exeter on 8 December 1967,* 1967.

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

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

Manuel Ayuso García

(Universidad Nacional de Educación a Distancia)*

Abstract

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

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

§

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

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

Introducción

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Corpus de trabajo

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

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

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

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

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

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

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

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

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

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

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

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

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

Preparación de la transcripción

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

Ejecución de la transcripción

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

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

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

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

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

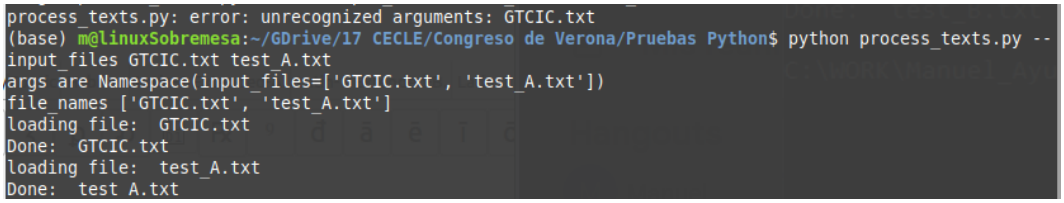


Fig. 1. Terminal en la ejecución del guión de Python

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

Tabla 1. Resultado del guión ejecutado sobre una transcripción

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

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

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

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

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

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

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

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

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

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

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

---

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

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

Entrenamiento de los sistemas de reconocimiento y creación de los modelos

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

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

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

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

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

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

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

Tabla 3a. Reconocimiento con Transkribus (READ Coop)

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

Tabla 3b. Reconocimiento con OCR4all

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

transcripción diplomática, OCR4all logra mejores resultados.

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

Reconocimiento de texto

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

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

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

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

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

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

Anexo 1: Tabla de ediciones

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

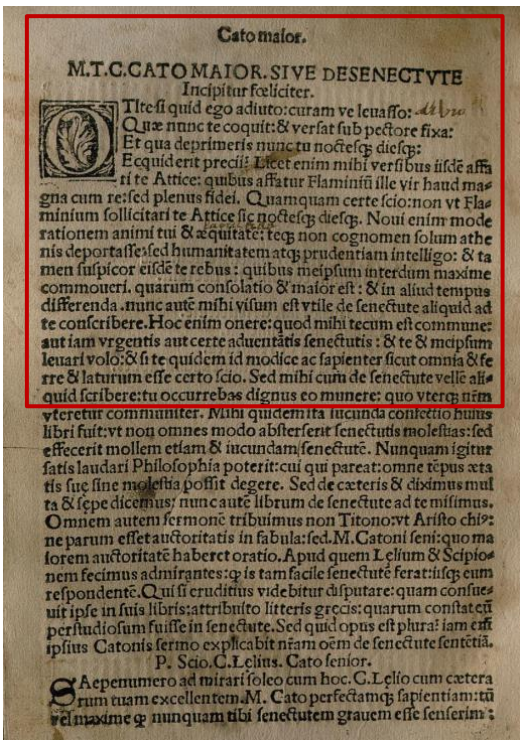
Aclaro el contenido de algunas columnas de esta tabla: «Título» aparece el título normalizado de la obra latina. «D.(isposición) de la pág.(ina)» se informa de la disposición en línea tirada (LT), dos columnas (2 col.) o dos textos, uno rodeando al otro (2 text.), que además puede tener *marginalia*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

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

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

Anexo 2: Imágenes de las ediciones y sus transcripciones

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



Cato maior.

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

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

P. Scio.C.Lelius. Cato ſenior.

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

Cato malor.

M.T. CATO MAIOR SIVE DE SENECTVTE

Incipitur fOeliciter.

Tite ſi quid ego adiuto: curam ve leuaſſo:

Quae nunc te coquit: & verſat ſub pectore fixa:

Et qua deprimeris nunc tu noceſq: die ſq3s

Ecquid erit preci. eēet enim mihi rerſibus iſcē aff

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

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

minium ſolicitari te Attice ſiq gocteſq3 dieſq3s. Noui enim

mode

rationem animi tui & aqualitate: teq3s non cognomen ſolum

athe

nis deportaſſeſed humanitatem atq3s prudentilam intelligo: &

ta

men ſuſpicor eiſē te rebus qubus meipſum interdum mexime

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

differenda tunc autē mſhi ſum eſt rtile de ſenectute aliquid ad

te conſcribere. Hoc eriim onere: quod mihi tecum eſt

commume:

aut iam rgentis aut certe aduentaātis ſenectutis: & te &

mcipſum

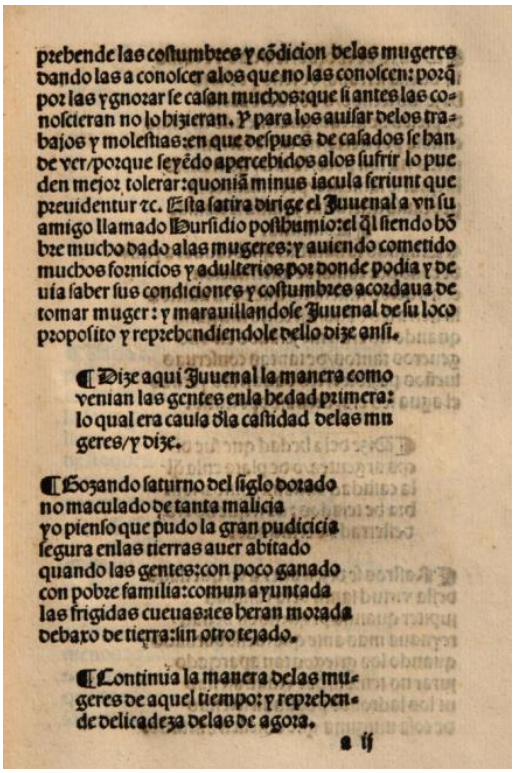
lerar' olo: & ſi tae quidem id modee ac fapiente ficut ommia &

fe

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

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

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



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

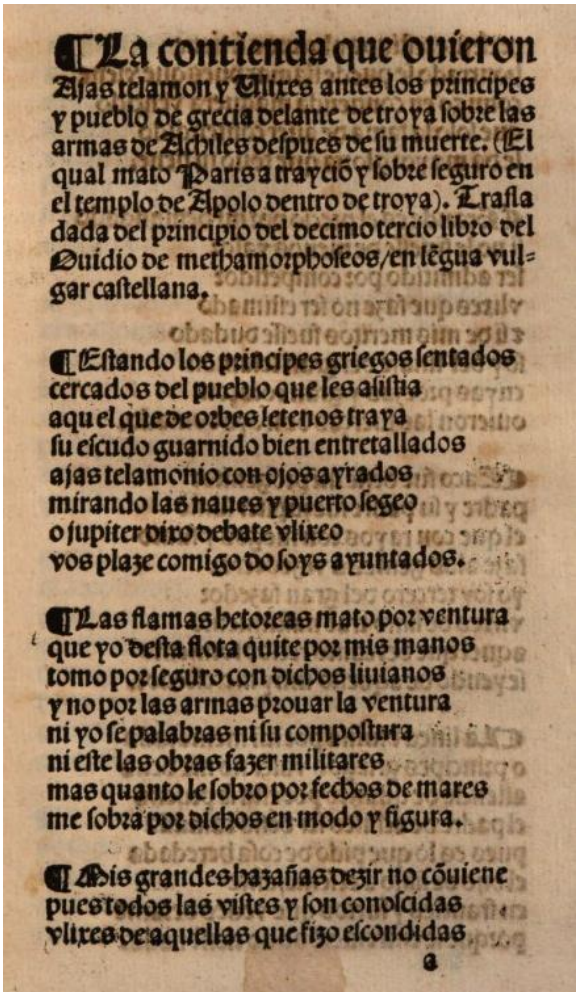
¶ Dize aqui Iuuenal la manera como
venian las gentes en la hedad primera:
lo qual era cauſa de la caſtidad de las mn
geres / y dize.

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

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

a ij

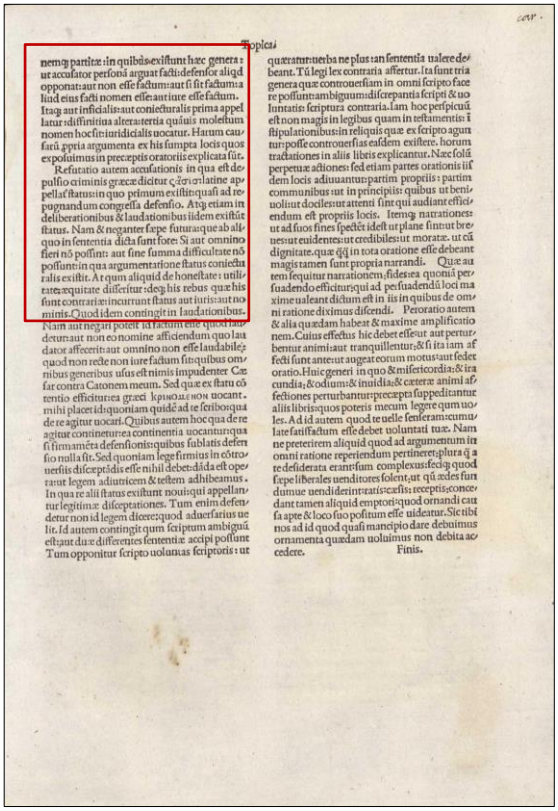
Ejemplo 2. Folio a2r del ejemplar BE.8.S.76 PS, Österreichische Nationalbibliothek, CECLE0155



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

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

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



Topicas

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

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

Bibliografía citada

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

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

Bazzaco, Stefano, «El Progetto Mambrino y las tecnologías OCR: estado de la cuestión», *Historias Fingidas*, 6 (2018), pp. 257-272. DOI: <https://doi.org/10.13136/2284-2667/89> (cons. 13/05/2022).

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

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

Mancinelli, Tiziana, «Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work», *Historias Fingidas*, 4 (2016), pp. 255-260. DOI: <https://doi.org/10.13136/2284-2667/65> (cons. 13/05/2022).

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

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

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

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

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

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

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

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

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

Reul, Christian, Uwe Springmann, Christoph Wick, Frank Puppe, «State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines» en *Proceedings of the DHd, 2019 Digital Humanities: Multimedial & Multimodal*, Mainz, 2019. URL: <https://arxiv.org/ftp/arxiv/papers/1810/1810.03436.pdf> (cons.18/05/2022).

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

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

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

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Los modelos de HTR *Silves1549_BNE* y *Spanish Gothic* como herramientas de la labor ecdótica

Giada Blasut

(Università di Verona)

Abstract

Esta contribución informa de la aplicación a textos áureos impresos en letra gótica de dos modelos de transcripción automatizada realizados dentro de la plataforma Transkribus. Serán objeto de estudio el modelo individual *Silves1549_BNE* y el modelo extendido *Spanish Gothic* ideados por Stefano Bazzaco en colaboración con diferentes miembros de las universidades de Lyon, Oporto, Verona y Zaragoza[1].
Palabras clave: libros de caballerías castellanos; *Silves de la Selva*; Transkribus; HTR (*Handwritten Text Recognition*); modelo HTR *Spanish Gothic*

This work reports the results of the application of two Transkribus print models based on Spanish books published in Gothic typeface between the 15th and the 16th centuries: the *Silves1549_BNE* and the extended *Spanish Gothic* model. The two models were created by a group of researchers from the Universities of Lyon, Oporto, Verona and Zaragoza.
Key words: Spanish Romances of Chivalry; *Silves de la Selva*; Transkribus; HTR (*Handwritten Text Recognition*); HTR model *Spanish Gothic*

§

[1] Han formado parte del grupo de investigación coordinado por Stefano Bazzaco los siguientes investigadores: Nuria Aranda García (École Normale Supérieure de Lyon), Ángela Torralba Ruberte (Universidad de Zaragoza), Ana-Milagros Jiménez (Universidad de Zaragoza), Pedro Monteiro (Universidade do Porto), Federica Zoppi (Università di Verona) y quien firma estas líneas.

Gaiada Blasut «Los modelos de HTR Silves1549_BNE y Spanish Gothic como herramientas de la labor ecdótica», *Historias Fingidas*, Número Especial 1 (2022) Humanidades Digitales y estudios literarios hispánicos, pp. 175-193.
DOI: https://doi.org/10.13136/2284-2667/1178 - ISSN: 2284-2667.

Hacia la primera edición crítica del *Silves de la Selva*

El libro de caballerías castellano *Silves de la Selva* es el duodécimo y último de la saga narrativa *Amadís de Gaula*[2]. Hoy en día, esta novela constituye la única publicación del ciclo amadisiano que carece todavía de una edición moderna del texto. Por esta y otras razones que he tenido ocasión de describir recientemente (Blasut, 2021), el *Silves de la Selva* es el objeto de estudio de mi tesis doctoral en la que aspiro a ofrecer lo siguiente: la biografía actualizada de su autor, Pedro de Luján; un estudio literario que informe de las estructuras narrativas de la novela; y la edición crítica del texto[3]. Como es de esperar, la realización del tercer objetivo, la edición de la novela, cuenta con las dos fases fundamentales de la labor filológica: la *recensio* y la *constitutio textus* (Blecua, 2001, 33).

La tradición textual del *Silves de la Selva* está constituida por dos ediciones publicadas en Sevilla en el taller de Dominico de Robertis en 1546 y 1549, respectivamente (Eisenberg y Marín Pina, 2000, 259-269). Aun así y como ha descubierto la comunidad científica, Dominico de Robertis ya había muerto cuando en 1549 salió la segunda publicación del *Silves de la Selva*. Las investigaciones revelan también que quien se estaba haciendo cargo de la gestión del taller era Pedro de Luján, sobrino político de Dominico de Robertis, al ser su madre Isabel Álvarez de Rivas, hermana de la mujer del impresor (Romero Tabares, 1998; Maillard Álvarez, 2007; Álvarez Márquez, 2009; Gestoso y Pérez, 1924, 118 y 118 n. 1; Hazañas y la Rúa, 1949, 90 y 102 n. 3)[4]. De modo que, con toda probabilidad, el autor del *Silves de la Selva*, Pedro de Luján, fue también el impresor de la segunda edición del texto; lo que lo convierte, junto con Andrea Pescioni, en uno de los dos tipógrafos sevillanos que «llevaron su

[2] En verdad, si se consideran como parte integrante del género de los libros de caballerías castellanos también las obras redactadas originariamente en otras lenguas europeas, habría que considerar el *Esferamundi* como última entrega del ciclo amadisiano. El libro fue escrito en italiano por Mambrino Roseo da Fabriano y publicado en seis partes en los talleres venecianos de Michele Tramezzino entre 1558 y 1565. Sobre la vida y la obra de Mambrino Roseo da Fabriano, consúltense Bognolo (2010) y Bognolo-Cara-Neri (2013).

[3] Para la biografía de Pedro de Luján véase: Romero Tabares (1998), Maillard Álvarez (2007), Álvarez Márquez (2009), Castillejo Benavente (2019) y Bazzaco (2020).

[4] Fue Francisco Escudero y Perroso quien, por primera vez, avanzó la hipótesis de que Pedro de Luján debía ser el sucesor de Dominico de Robertis (1894, 25 y 236).

destreza en el manejo de la pluma a la composición de obras de propia creación o a la traducción al castellano de otras ajenas» (Álvarez Márquez, 2007, 13-14)[5].

En un estudio reciente he tenido ocasión de indicar que las ediciones del *Silves de la Selva* no son idénticas, a pesar de que se publicaron en la misma casa tipográfica. En realidad, el cotejo revela que la segunda derivó de la primera –al compartir errores comunes conjuntivos–, pero que se diferencia notablemente de la *princeps* por las abundantes variantes textuales que transmite (Blasut, 2021). En la base del cotejo efectuado y consciente de que en 1549 Pedro de Luján estaba gestionando la tipografía que había fundado Robertis, he elegido la segunda publicación de la novela (Sevilla, 1549) como texto base de la edición que propongo en mi tesis.

Por consiguiente, necesitaba fuera la transcripción del texto de 1549 para proseguir en la labor filólogica. Sin embargo, el *Silves de la Selva* es, al igual que los demás libros de caballerías castellanos, un texto largo; su longitud ocupa 150 folios (recto y verso) escritos a doble columna, de unas 46 líneas cada una. Esto implica disponer de mucho tiempo para transcribirlo manualmente a través de un normal procesador de textos como, por ejemplo, Microsoft Word.

Por este motivo y a raíz de los avances que la aplicación de modelos de transcripción automatizada habían aportado en el campo ecdótico, decidí recurrir a una herramienta digital que permitiera reducir el tiempo de trabajo sin renunciar a la calidad del mismo. En virtud de ello, Stefano Bazzaco me aconsejó que utilizara la plataforma Transkribus; lo que me ofrecía también la ocasión de poder valorar la aportación de la plataforma a la tradicional labor filológica. ¿Puede Transkribus revelarse una herramienta no solo útil, sino también fundamental para realizar una edición crítica de un texto español del Siglo de Oro?; ¿qué ventajas ofrece respecto a una transcripción manual?; e incluso, ¿puede el filólogo utilizar Trankribus sin renunciar al contacto directo con el ejemplar que transmite el texto?

[5] El impresor, editor y mercader de libros Andrea Pescioni tradujo del francés al castellano la obra de Pedro Bovistau, Claudio Tesserant y Francisco Beleforest *Historias prodigiosas y maravillosas de diversos sucesos acaecidos en el mundo* (Francisco del Canto, 1586, Medina del Campo) (Álvarez Márquez, 2007, 14).

La necesidad de un modelo de transcripción para el *Silves de la Selva*

Cuando, en 2020, me dispuse a utilizar Transkribus para la transcripción automatizada del *Silves de la Selva*, todavía no estaba disponible un modelo aplicable a textos áureos impresos en letra gótica. El Progetto Mambrino (grupo de investigación de la Universidad de Verona coordinado por Anna Bognolo y Stefano Neri), ya había desarrollado un modelo de transcripción automatizada apto para la aplicación a libros de caballerías escritos en lengua italiana a imitación de los castellanos[6]. Sin embargo, por tratarse de impresos en letra cursiva, y no gótica como el *Silves de la Selva*, su modelo no podía aplicarse al texto objeto de mi interés (Bazzaco, 2018). Con todo, los excelentes resultados de la transcripción automatizada de las novelas italianas demostraban, de manera eficiente, la fiabilidad de Transkribus y las ventajas de su aplicación: decrecimiento del tiempo útil a la transcripción, fiabilidad de la plataforma Transkribus, reducción de las tareas del editor del texto (Bazzaco, 2020). Por eso, y a falta de un modelo aplicable a los textos castellanos impresos en letra gótica, se hacía necesario crear un modelo *ex profeso* para transcribir el *Silves de la Selva*; lo que realicé bajo la coordinación de Stefano Bazzaco.

Como es sabido, en la plataforma se pueden crear dos tipos de modelos: los individuales, cuya aplicación queda reservada a un solo texto; y los modelos extendidos aptos para textos diferentes, pero parecidos en su conformación tipográfica. Ambos tipos precisan de una fase inicial denominada *Golden Transcription* (transcripción manual) mediante la cual Transkribus aprende, sirviéndose de técnicas de *machine learning*, la correlación entre la representación visual de los caracteres de la imagen y

[6] En el siglo XVI, los libros de caballerías castellanos se tradujeron a las principales lenguas europeas (francés, italiano, alemán, inglés y holandés). Incluso, algunos libros del ciclo narrativo del *Amadís de Gaula* contaron también con traducción al hebreo. En Italia, Francia y Alemania se produjeron además continuaciones originales o *aggiunte* que fueron sucesivamente traducidas al castellano. La cuestión de la recepción de los libros de caballerías en Europa puede leerse en Bognolo (2020a; 2020b), Bognolo-Cara-Neri (2013), Neri (2008a; 2008b; 2013), Schlusemann y Wierzbicka-Trwoga (2020), Sánchez-Martí (2019) y Wilson (2014). Para la difusión en Italia consúltese Bognolo (1984; 2008; 2011) y Bognolo-Cara-Neri (2013).

su transcripción[7]. Aun así, la labor de transcripción manual difiere por la cantidad de textos transcritos y, por ende, por el número de páginas sobre las que se construyen. Si, por un lado, los modelos individuales se basan en un único texto, del que se transcribe tan solo una veintena de páginas; por el otro los modelos extendidos disponen de un corpus más vasto de textos y por consiguiente de un número más copioso de páginas transcritas manualmente. De tal manera que las principales diferencias que atañen la génesis de los modelos son dos. En primer lugar, el número de textos empleados: uno en el caso del modelo individual, mientras que varios en el caso del modelo extendido; y, en segundo, la cantidad de caracteres que el software tiene la oportunidad de aprender mediante la transcripción manual. En el caso del modelo extendido es mayor el número de páginas transcritas manualmente, por lo que aumenta la posibilidad de que la plataforma vea y aprenda nuevos caracteres (mayúsculas, minúsculas, signos de puntuación u otros especiales). Del mismo modo, la transcripción de un elevado número de páginas conlleva, evidentemente, la reiteración de muchos caracteres que la plataforma ve una y otra vez, aumentando así su capacidad mnemotécnica. Como tendré ocasión de profundizar a continuación, además de la diferenciación ya señalada, hay otro punto fundamental de divergencia entre los dos modelos y que se refleja incluso en su denominación: su finalidad. Frente al modelo individual que se puede aplicar al solo texto que ha sido empleado durante su creación, el modelo extendido aspira a transcribir automáticamente también textos externos a su corpus, es decir, a obras que no se han utilizado como *dataset* para su producción.

En lo que atañe a la creación del modelo individual del *Silves de la Selva*, he transcrito manualmente dentro de la plataforma las primeras veinte páginas de la novela siguiendo los criterios que se indican a continuación:

- u/v/i/j/y aparecen como en el texto;
- letras mayúsculas y minúsculas aparecen como en el texto;
- longitud de las líneas es fiel al texto. Un guion indica la división

[7] Ofrezco aquí solamente una descripción somera de las principales fases de creación del modelo individual del *Silves de la Selva*. Para una explicación detallada del flujo de trabajo que supone Transkribus véase Bazzaco (2020).

de las palabras cortadas al final de línea incluso cuando este no se halla en el ejemplar;
- la unión y la separación de las palabras refleja el uso actual, con la excepción de algunos vocablos como «toda via» y «tan/m bien» que se mantienen separados;
- toda abreviación se ha desarrollado y en ningún caso se han acentuado las palabras, ni se han introducido apóstrofes al fin de subrayar la fusión de dos vocales («dellos» y no «d'ellos»);
- los errores de imprenta no se han corregido;
- no se han introducido signos de puntuación que no estén en el texto.

En definitiva, se trata de criterios destinados a obtener un modelo de transcripción diplomática del *Silves de la Selva*, al que denominaré *Silves1549_BNE*.

Valoración del modelo individual del *Silves de la Selva*

La creación y la aplicación del modelo individual de transcripción automatizada del *Silves de la Selva* se han revelado fructuosas y de gran utilidad. Entre las ventajas más destacadas resalta, sin duda, la rapidez de ejecución de la plataforma. Transkribus necesita solamente de algunos minutos para transcribir la novela entera. Pero es más, la aplicación de Transkribus no solo implica un ahorro en términos de tiempo, sino también de deberes; el editor encomienda la transcripción del texto que necesite a la plataforma y se reserva para sí mismo solamente la fase sucesiva de su labor, es decir, la corrección de la transcripción. Esta puede realizarse de distintas maneras, bien dentro de la plataforma, bien fuera de ella. Si bien la primera se revela de gran utilidad, ya que permite visualizar simultáneamente tanto la imagen como su transcripción automatizada, son muchas las ventajas que aporta la corrección realizada fuera de la plataforma. En primer lugar, es preciso señalar que es posible exportar la transcripción en múltiples formatos (Transkribus Document, PDF, TEI, DOCX, Simple TXT, Excel e IOB), modificables, por lo general, por parte del editor. Este aspecto es una de las características más destacadas y

apreciadas de Transkribus, puesto que el editor puede emendar la transcripción automatizada directamente dentro del documento y, posteriormente, utilizar este mismo como esqueleto de la edición que pretende realizar. Lamentablemente, una vez revisada la transcripción, el texto no puede volver a subirse a la plataforma; lo que permitiría, en mi opinión, no solamente disponer de una transcripción correcta incluso dentro de la plataforma, sino también contribuir al *entrenamiento* de Transkribus.

En lo que a la valoración de la aplicación del modelo individual del *Silves de la Selva* se refiere, el estudio revela lo siguiente: se han contabilizado catorce errores por página, espacios incluidos. Cabe señalar que la mayoría de los fallos cometidos por la sistema de HTR concierne a la errónea unión y separación de palabras, cuando no su acentuación. Por lo tanto, la naturaleza de tales errores determina una mínima intervención por parte del editor, quien se limita a colocar o eliminar algún espacio en blanco o bien a añadir los acentos necesarios. La adición de un espacio en blanco se puede ejemplificar con el caso en el que la conjunción copulativa «y» y el nombre del personaje que lo sucede, «Anaxartes», vienen concebidos por la plataforma como si de un solo término se tratara «yanaxartes». De tal manera que en la fase de revisión se hace necesario separar los dos términos para restablecer la exacta lectura del texto. Al contrario, un poco más adelante, la plataforma sobreinterpreta la información separando la palabra «tierra» en dos partes sin significado alguno, «tier» y «ra». También en este caso, el quehacer del editor es rápido y limitado, si bien fundamental; solamente hace falta eliminar el espacio en blanco que Transkribus ha colocado entre los dos fragmentos. Cabe destacar que no se trata de errores excesivamente graves, ya que no afectan seriamente al significado del texto. El editor puede, además, subsanarlos fácilmente incluso sin recurrir a la imagen del texto.

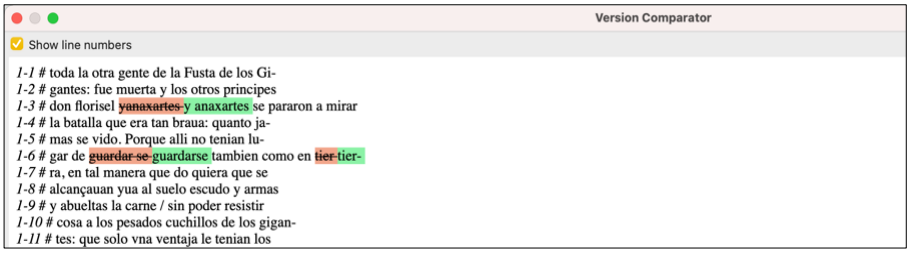


Fig. 1. Errores de transcripción cometidos por Transkribus

Dado que esta tipología de error representa la mayoría, el editor ya sabe dónde puede hallar tales incorrecciones. Como se puede apreciar (Fig. 1 y 2), dichas incongruencias suelen darse, por lo general, al finalizar una línea; lo que las convierte en errores muy previsibles y fácilmente subsanables.

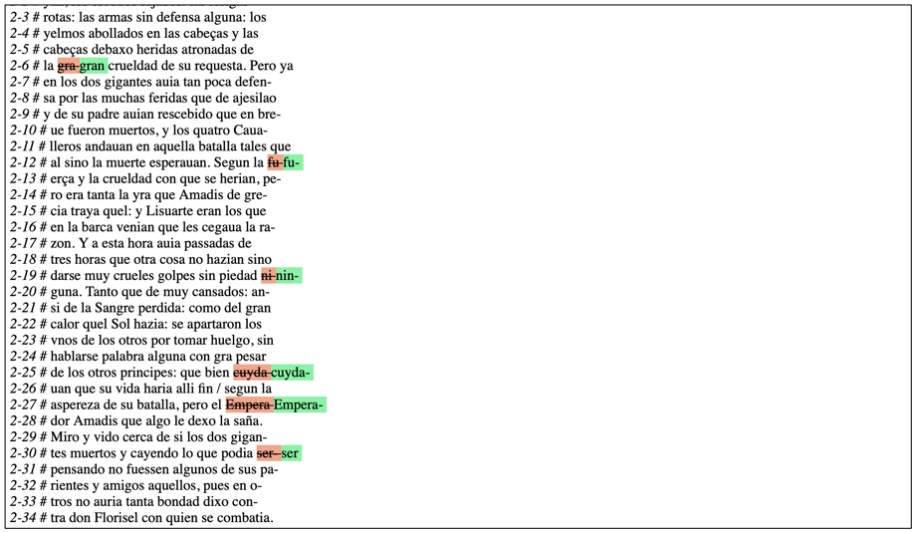


Fig. 2. Ejemplos de errores más frecuentes en una página transcrita

Se han localizado también situaciones en las que Transkribus se equivoca en el reconocimiento de un carácter que transcribe sustituyéndolo por otro, o más de uno, como ocurre, por ejemplo, con el término «cuñada» donde Transkribus convierte equivocadamente una «ñ» en una «u», o bien el caso del verbo «Hablo» cuya letra inicial viene por

error interpretada como una «N». Según se verá también analizando el modelo extendido, es bastante frecuente que la plataforma se equivoque con las letras mayúsculas. Esto se debe a que tales caracteres suelen escasear en los textos de la época, lo que implica que Transkribus tiene pocas ocasiones de verlos y aprenderlos durante la transcripción manual necesaria para su aprendizaje.

Valoración del modelo extendido *Spanish Gothic* (Siglos XV-XVI)

La cronología de publicación de los títulos elegidos para crear el modelo *Spanish Gothic* (nombre en Transkribus: *SpanishGothic_XV-XVI_extended*) abraza un marco temporal muy amplio que empieza en 1487 y finaliza en 1563, fechas en la que se publicaron las ediciones de *Doctrinal de caballeros* y *Libro del conde Partinuplés*, respectivamente, con las que se ha trabajado[8]. Mediante este corpus se aspira a realizar un modelo destinado a la transcripción de textos impresos en castellanos durante el Siglo de Oro español y sus albores. Análogamente, se ha querido cubrir también una vasta geografía que refleje la multitud de talleres activos en la publicación de libros en castellano cuya localización trasciende incluso, amén de los confines de los reinos españoles, las fronteras peninsulares como en el caso de la *Tragicomedia de Calisto y Melibea* impresa en 1515 en Roma, en el taller de Marcellus Silber. A pesar de haberse impreso en lugares y décadas diferentes, todas estas obras comparten unas características editoriales que las convierten en un corpus homogéneo de

[8] Las obras y las ediciones utilizadas para la creación del modelo *Spanish Gothic 15th-16th Century* son las siguientes: *Doctrinal de los Caballeros* de Alonso de Cartagena (Burgos, Fadrique Biel de Basilea, 1487), *La Fiameta* de Juan Boccaccio (Salamanca, [Impresor de la Gramática de Nebrija], 1497), *Crónica del Rey Don Rodrigo* (Crónica Sarracina) de Pedro de Corral ([Sevilla], [Meinardo Ungut y Estanislao Polono], 1499); *Retablo de la Vida de Cristo* de Juan de Padilla (Sevilla, Juan Cromberger, 1510); *Tragicomedia de Calisto y Melibea* (Roma, Marcelo Silber, 1515); *Historia de la linda Magalona* (Sevilla, Jacobo Cromberger, 1519); *Libro del conde Partinuplés* (Sevilla, Jacobo Cromberger, 1519); *Lisuarte de Grecia* de Juan Díaz (Sevilla, Jacobo y Juan Cromberger, 1526); *Historia del rey Canamor* (Valencia, Jorge Costilla, 1527); *Florando de Inglaterra* (Lisboa, Germán Gallarde, 1545); *Silves de la Selva* de Pedro de Luján (Sevilla, Herederos de Dominico de Robertis), 1549; *Lisuarte de Grecia* de Feliciano de Silva (Sevilla, Jácome Cromberger 1550); *Historia de la reina Sebilla* (Burgos, Felipe de Junta, 1551); *Libro del conde Partinuplés* (Burgos, Herederos de Juan de Junta, 1558); *Leandro el Bel* de Pedro de Luján (Toledo, Miguel Ferrer, 1563); *Libro del conde Partinuplés* (Burgos, Felipe de Junta, 1563).

textos apto para la creación de un modelo de transcripción automatizada. La *conditio sine qua non* es la tipología de tipos móviles utilizados para su impresión: la letra gótica. Al mismo tiempo, es de suma importancia la calidad de la reproducción digital de los textos, sea a color o en blanco y negro. Gracias a las peculiaridades del corpus que se han descrito anteriormente es posible aplicar el modelo a textos publicados no solo en diferentes décadas del Siglo de Oro español, sino también en lugares y talleres distintos. Considérese además que la transcripción manual de las obras antes citadas se ha realizado a partir de reproducciones digitales extraídas de varios portales y que las disparidades en el formato y en la calidad de la imagen han garantizado, contrariamente a lo que se podría pensar, un entrenamiento más objetivo de Transkribus. El software puede no solamente trabajar con reproducciones a color y en blanco y negro, sino también, como tendré ocasión de demostrar, con imágenes más o menos borrosas.

Para valorar la eficacia del modelo *Spanish Gothic* se han concebido dos tipos de pruebas. Primero, y como en el caso del modelo individual antes descrito, se ha aplicado el modelo a un texto utilizado para su creación y, en segundo lugar, se ha experimentado en obras ajenas que no pertenecen al corpus de textos transcritos manualmente para crear dicho modelo. Dado que entre los textos utilizados para la creación del modelo extendido figura también el *Silves de la Selva*, o sea, la misma obra para la cual se había realizado previamente un modelo individual de transcripción automatizada, se ha decidido aplicar también el modelo extendido a este texto a fin de evaluar cuál de los dos modelos, el individual creado *ex profeso* para esta novela, o el extendido –de cuyo corpus forma parte también el *Silves*– se revela mejor para su transcripción. De tal manera que la aplicación del modelo extendido al *Silves de la Selva* permite comparar directamente los dos modelos y realizar un análisis de su eficacia.

La aplicación del modelo extendido *Spanish Gothic* para la transcripción del *Silves de la Selva* ha dado resultados sorprendentes. La plataforma ha cometido solamente siete errores por página, es decir, la mitad exacta de los que había cometido sirviéndose del modelo individual. En la imagen que se propone a continuación es posible comparar la cantidad y la calidad de los errores (Fig. 3). Según se aprecia, el modelo

extendido no deja de cometer, si bien en raras ocasiones, los mismos errores producidos por el modelo individual, como ocurre, por ejemplo, con el verbo «cuidauan» que en ambos casos no se reconoce como una única palabra sino como dos, «cuyda» y «uan». Adviértase que en esta, como en otras situaciones, el error se produce al final de una línea, punto crítico de la transcripción debido a que sus cortes inducen evidentemente a error a la plataforma. Al contrario, por lo que concierne a la unión y separación de palabras colocadas en el interior y no al final de una línea, el *Spanish Gothic* resulta más idóneo para su transcripción. Tan solo por poner un ejemplo, considérese la palabra «guardarse» que cuando se aplica el modelo extendido se escribe como corresponde, mientras que se separa indebidamente en el individual. Sin embargo, en la transcripción del modelo extendido pueden apreciarse errores que no se dan en el modelo individual, como ocurre con la abreviatura de la palabra «do*n*» que no se desarrolla, pero se trata de raros casos.

Como conclusión del cotejo de las dos transcripciones puede verse que el modelo extendido aplicado a un texto utilizado para su creación es mucho más fiable que un modelo individual creado para el mismo texto. Gracias a la utilización del modelo *Spanish Gothic* los errores cometidos por la plataforma han disminuido notablemente determinando una exactitud de la transcripción automatizada del 99,78%; la plataforma se equivoca solamente 0,22 veces cada cien caracteres, espacios incluidos.

Así pues, falta por analizar la eficacia del modelo extendido cuando se aplica a los textos para los que ha sido creado: las obras que no formaban parte del corpus de textos utilizados para su creación.

Fig. 3. Errores de transcripción en unas páginas del *Silves*

Con el cometido de investigar más en profundidad las ventajas que el modelo extendido *Spanish Gothic* puede ofrecer a la comunidad científica y, sobre todo, a los editores de textos, se ha efectuado un segundo tipo de prueba. Esta última consiste en la aplicación del modelo a un corpus externo de textos que no hayan sido previamente utilizados para su creación. Se pretende simular con este examen la labor de un editor interesado en transcripción automatizada, a fin de ofrecer un análisis objetivo de la aplicación del modelo extendido y de su valoración. Las obras utilizadas en esta prueba se han seleccionado casualmente dentro del portal Biblioteca Digital Hispánica, donde están disponibles numerosas reproducciones digitales de diferente formato y calidad, en blanco y negro o a color. Los títulos seleccionados y que, evidentemente, cumplen con las condiciones necesarias para la aplicación del modelo –la tipología de carácter móvil, la letra gótica; la lengua, el castellano; y el periodo de publicación, los siglos XV-XVI– son los que se enumeran a continuación:

- *Felixmarte* (Valladolid, Francisco Fernández de Córdoba, 1556);
- *Reprobación de las supersticiones y hechicerías* (Salamanca, Pedro Tovans, 1540).

La primera de las obras que se ha sometido a esta prueba ha sido el libro de caballerías castellano *Felixmarte* (Valladolid, Francisco Fernández de Córdoba, 1556), cuyo aspecto tipográfico guarda un estrecho parecido con el *Silves de la Selva* al pertenecer ambas al mismo género literario y editorial. De igual manera, también su reproducción digital es, al igual que la anterior, de buena calidad y a color.

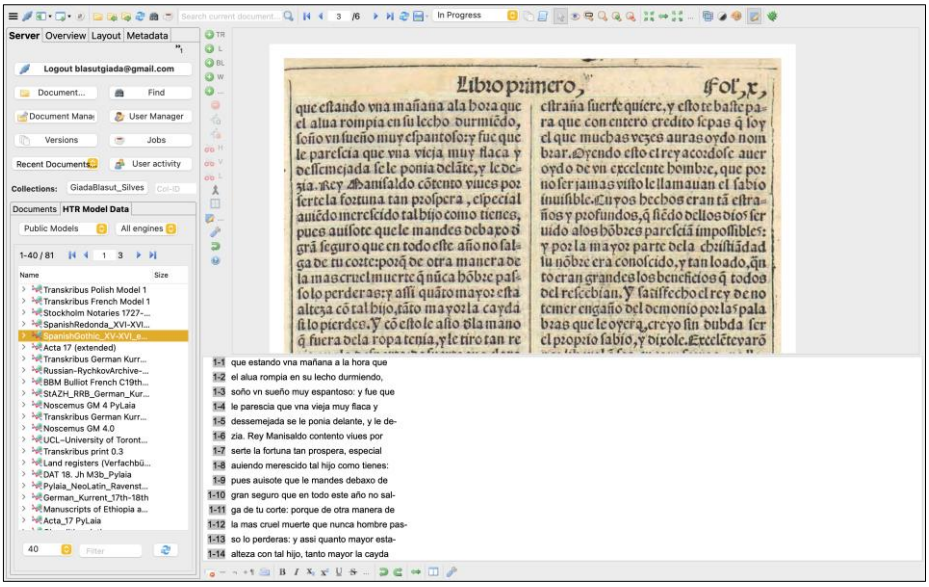


Fig. 5. Prueba 1: *Felixmarte de Hircania* (1556)

Cuando se aplica el modelo extendido al *Felixmarte*, los errores cometidos por la transcripción automatizada aumentan levemente, pasando de los siete del *Silves de la Selva* a los diez de este libro. La clasificación de los errores coincide con las precedentes, al mismo tiempo, destacan las omisiones y adiciones impropias de guiones al finalizar una línea. Por poner unos ejemplos, considérense la palabra «christiano» que la plataforma no reconoce como una única palabra, o, al contrario, los términos «esta alteza» que se unen por error mediante la añadidura de un guion. En esta misma estela se insertan también aquellos puntos donde sobran o faltan espacios en blanco entre dos palabras que aparecen así erróneamente unidas o separadas, como ocurre, por ejemplo, con la expresión «Jesuchristo» donde el software une dos palabras. Además, en unas pocas situaciones, aparece un carácter en lugar de otro, como la «m» en la lugar de la «n» en la palabra «con», o un punto en lugar de la coma en la frase «gran sobresalto puso al rey. y a todos los que la vieron». Incluso, en algunas ocasiones, se aprecia la omisión de la letra «n» cuando se trata de una abreviación colocada en el interior o al final de una palabra, como se da en las palabras «con» y «dilacion». A pesar de estos pequeños

y escasos errores, hay que reconocer que también la aplicación del modelo a un texto externo ofrece resultados satisfactorios; tan solo unos diez errores por página, lo que significa que Transkribus se equivoca solamente 0,29 veces cada cien caracteres o, lo que es lo mismo, que la traducción automatizada es correcta en el 99,71% de los casos. Se trata, de toda manera, de un éxito mayor del que puede obtenerse aplicando un determinado modelo individual. Estos datos demuestran, por un lado, la validez del modelo extendido *Spanish Gothic* y, por otro, su enorme utilidad y practicidad; ya no es necesario que el usuario desarrolle un modelo individual para cada obra que le interesa, dado que su aplicación no solo requeriría mucho más tiempo, sino que sería también menos eficaz que el modelo extendido.

El estudio de la aplicación del modelo a obras externas al corpus prosigue con el análisis de un libro didáctico de Pedro Ciruelo, la *Reprobación de las supersticiones y hechicerías* (Salamanca, Pedro Tovans, 1540) –más conocido como *Supersticiones*–, que no está impreso a doble columna como los demás que se han considerado hasta ahora, sino a línea tirada, y cuya reproducción digital está disponible en blanco y negro. Téngase además en cuenta la siguiente variante: la calidad de estas imágenes no es tan nítida como las que se han utilizado anteriormente. Con todo, los resultados de la aplicación del modelo a las *Supersticiones* continúan dando prueba de la validez del modelo extendido. El número de errores sigue siendo escaso, once errores por página, lo que implica que la transcripción es exacta al 99,39%, dato que adquiere aún más relevancia si se considera que este texto tiene una decena de líneas menos que los demás textos analizados.

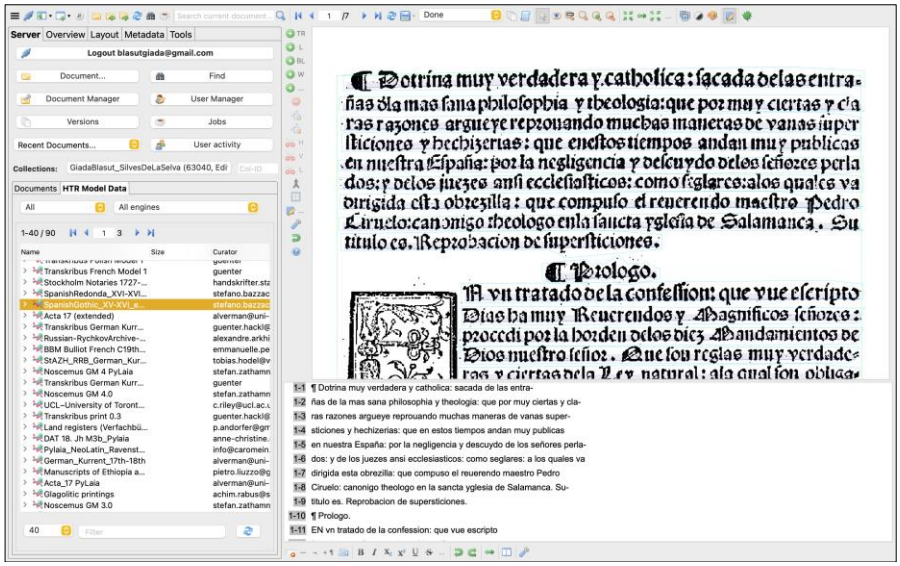


Fig. 5. Prueba 2: *Reprobación de las supersticiones y hechicerías* (1540)

Según el número de errores detectados, se infiere que el modelo *Spanish Gothic* trabaja mejor con las reproducciones a color y con imágenes de buena o calidad media. De este modo, el éxito del reconocimiento de los caracteres por parte del software está directamente relacionado con la calidad de la imagen. Asimismo, la transcripción de las *Supersticiones* realizada a partir de esta peculiar digitalización ofrece la posibilidad de destacar otra de las ventajas más exitosas de Transkribus. Realmente, incluso en los casos de difícil o laboriosa lectura, la plataforma proporciona siempre una interpretación del texto. Dicho de otro modo, aun cuando Transkribus no consigue leer debidamente algún carácter a causa, generalmente, de la mala calidad de la imagen, ofrece siempre una solución al editor, es decir, una posible lectura del texto. En tales casos, la plataforma interpreta los caracteres basándose en el parecido que tienen con otros que ya conoce o a los que está más acostumbrada. Esto hace de Transkribus y sobre todo del modelo *Spanish Gothic* una herramienta puntual y cuidada que ofrece incluso, como en el caso de las *Supersticiones,* una importante ayuda para el editor cuando la calidad de la reproducción digital no es la mejor y dificulta la exacta interpretación del texto..

Bibliografía citada

Álvarez Márquez, María del Carmen, *Impresores, libreros y mercaderes de libros en la Sevilla del quinientos*, Zaragoza, Libros Pórtico, 2009, vol. I.

—, *La impresión y el comercio de libros en la Sevilla del Quinientos*, Sevilla, Secretariado de Publicaciones de la Universidad, 2007.

Bazzaco, Stefano, «El Progetto Mambrino y las tecnologías OCR: estado de la cuestión», *Historias fingidas*, 6 (2018), pp. 257-272. DOI: <https://doi.org/10.13136/2284-2667/89> (cons. 15/10/2021).

—, «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. URL: <https://www.janusdigital.es/articulo.htm?id=160> (cons. 15/10/2021).

—, «Introducción», en *Leandro el Bel*, ed. Stefano Bazzaco, Alcalá de Henares Editorial, Universidad de Alcalá, 2020.

Blecua, Alberto, *Manual de crítica textual*, Editorial Castalia, Madrid, 2001.

Bognolo, Anna, «La prima traduzione italiana dell'*Amadís de Gaula*: Venezia 1546», *Annali della Facoltà di Lingue e Letterature Straniere di Ca' Foscari*, Bulzoni, XXIII, 1, 1984, pp. 1-29.

—, «Libros de caballerías en Italia», en *«Amadís de Gaula», 1508: quinientos años de libros de caballerías*, ed. José Manuel Lucía Megías, Madrid, Biblioteca Nacional de España, Sociedad Española de Conmemoraciones Culturales, 2008, pp. 333-341.

—, «Vida y obra de Mambrino Roseo da Fabriano, autor de libros de caballerías», *eHumanista. Journal of Iberian Studies*, 16 (2010), pp. 77-98.

—, «Il romanzo cavalleresco spagnolo in Italia e la collezione di *Amadís* della Biblioteca Civica di Verona», en *L'età di Carlo V. La Spagna e l'Europa. Sesto quaderno del Dottorato in Letterature Straniere e Scienze della Letteratura Università di Verona*, ed. Silvia Monti, Verona, Edizioni Fiorini, 2011, pp. 125-145.

—, «Il romanzo spagnolo del Rinascimento», en *Forme di romanzi dall'Antico alle soglie del Moderno. Le forme e la storia*, ed. Antonio Pioletti, XIII/2, Soveria Mannelli, Rubbettino, 2020a, pp. 163-186.

—, «Libros de caballerías: ficciones españolas para el Renacimiento europeo», en *En la villa y corte. Trigesima Aurea. Actas del XI Congreso de la Asociación Internacional Siglo de Oro (Madrid, 10-14 de julio de 2017)*, eds. Ana Martínez Pereira, María Dolores Martos Pérez, Esther Borrego Gutiérrez y Inmaculada Osuna Rodríguez, Madrid, UNED-Universidad Complutense, 2020b, pp. 53-82.

Bognolo, Anna y Bazzaco, Stefano, «Tra Spagna e Italia: per un'edizione digitale del Progetto Mambrino», *eHumanista/IVITRA*, 16 (2019), pp. 20-36.

Bognolo, Anna, Giovanni Cara y Stefano Neri, *Repertorio delle continuazioni italiane ai romanzi cavallereschi spagnoli. Ciclo di Amadis di Gaula*, Roma, Bulzoni, 2013.

Castillejo Benavente, Arcadio y Cipriano López Lorenzo, *La imprenta en Sevilla en el siglo XVI (1500-1600),* Córdoba, Sevilla, Editorial Universidad de Córdoba, Editorial Universidad de Sevilla, 2019.

Eisenberg, Daniel y Mª del Carmen Marín Pina, *Bibliografía de los libros de caballerías castellanos*, Zaragoza, Prensas Universitarias de Zaragoza, 2000.

Escudero y Perosso, Francisco, *Tipografía hispalense: anales bibliográficos de la ciudad de Sevilla desde el establecimiento de la imprenta hasta fines del siglo XVIII*, Madrid, Establecimiento tipográfico «Sucesores de Rivadeneyra», 1894.

Hazañas y la Rúa, Joaquín, *La imprenta en Sevilla: Noticias inéditas de sus impresores desde la introducción del arte tipográfico en esta ciudad hasta el siglo XIX*, Sevilla, Diputación Provincial, 1945-1949, 2 vols.

Maillard Álvarez, Natalia, *Difusión y circulación de la cultura escrita en Sevilla. 1550-1600*, Tesis doctoral, dir. Carlos Alberto González Sánchez, Sevilla, Universidad de Sevilla, 2007.

Neri, Stefano, «Cuadro europeo de la difusión del ciclo del Amadís de Gaula (siglos XVI-XVII)», en *Amadís de Gaula: quinientos años después. Estudios en homenaje a Juan Manuel Cacho Blecua*, eds. José Manuel Lucía Megías y María Carmen Marín Pina, Alcalá de Henares, Centro de Estudios Cervantinos, 2008a, pp. 565-591.

—, «El Progetto Mambrino. Estado de la cuestión», en *"Tus obras los rincones de la tierra descubren". Actas del VI Congreso Internacional de la Asociación de Cervantintas. Alcalá de Henares, 13 al 16 de diciembre de 2006*, eds. Alexia Dotras Bravo, José Manuel Lucía Megías, Elisabet Magro García y José Montero Reguero, Madrid, Asociación de Cervantistas; Centro de Estudios Cervantinos, 2008b, pp. 577-589.

—, «Cuadro de la difusión europea del ciclo palmeriniano (siglos XVI-XVII)», en *Palmerín y sus libros: 500 años*, eds. Aurelio González, Axayácatl Campos García Rojas, Karla Xiomara Luna Mariscal, Carlos Rubio Pacho, México D.F., El Colegio de México - Centro de Estudios Lingüísticos y Literarios, 2013, pp. 285-314.

Romero Tabares, Isabel, *La mujer casada y la amazona. Un modelo femenino renacentista en la obra de Pedro de Luján*, Sevilla, Universidad de Sevilla, 1998.

Sánchez-Martí, Jordi, «The Printed Popularization of The Iberian Books of Chivalry Across Sixteenth-Century Europe», en *Crossing Borders, Crossing Cultures*, eds. Massimo Rospocher, Jeroen Salman and Hannu Salmi, Berlin, Boston, De Gruyter Oldenbourg, 2019, pp. 159-180.

Schlusemann, Rita y Krystyna Wierzbicka-Trwoga, «Narrative Fiction in Early Modern Europe», *Quaerendo*, 51/1-2 (2021), pp. 160-188.

Wilson, Louise, «The Publication of Iberian Romance in Early Modern Europe», en *Translation and the Book Trade in Early Modern Europe*, eds. J. Pérez Fernández y E. Wilson-Lee, Cambridge, Cambridge University Press, 2014, pp. 201-216.

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

La publicación de ediciones digitales académicas y el caso de las *Soledades* de Luis de Góngora

Antonio Rojas Castro

(Berlin-Brandenburgische Akademie der Wissenschaften)

Abstract

Este artículo tiene por objetivo caracterizar el proceso de publicación de ediciones digitales académicas mediante herramientas informáticas que transforman documentos XML/TEI a HTML. Tras explicar las diferencias existentes entre las webs dinámicas y las webs estáticas en el primer apartado, se exponen algunos desafíos que supone el uso de herramientas de publicación y los efectos derivados de la estandarización y se analiza la publicación web de las *Soledades* de Luis de Góngora realizada con la herramienta Edition Visualization Technology.
Palabras clave: Edición digital Académica; EVT; Luis de Góngora; publicación; *Soledades*

This article aims to describe the process of publishing of digital scholarly editions using tools that transform XML/TEI documents into HTML. After introducing the main differences between dynamic and static webs in the first section, this article discusses some challenges involved in the use of publishing tools and the effects derived from standardization, and analyzes the web publication of Luis de Góngora's *Soledades* built with Edition Visualization Technology.
Keywords: Digital Scholarly Editions; EVT; Luis de Góngora; publication; *Soledades*

§

Introducción

Cualquier persona que tenga un poco de experiencia en la producción de contenidos textuales (desde artículos de revista, pasando por libros electrónicos, hasta la publicación de blogs o incluso comentarios en foros), sabe que la labor editorial ocupa hoy en día una posición ambivalente. Se habla, a menudo, de «revolución» o de «cambio de paradigma» para referirse a la transformación en la forma de trabajar en instituciones culturales y académicas, tal y como la describió Thomas

S. Kuhn (2013). Con todo, y pese al optimismo de los tecnólogos, la labor editorial sigue siendo deudora del paradigma impreso, pues consiste en una serie de tareas conocidas desde antaño como la selección, lectura, análisis, revisión, transcripción, corrección, normalización, anotación, representación y presentación[1]. Por supuesto, algunas de estas tareas pueden agilizarse o incluso automatizarse gracias a los avances informáticos, pero el núcleo de la labor editorial sigue siendo el mismo: dar al lector actual el mejor texto posible[2].

Ahora bien, con la digitalización ha habido dos desafíos principales en gran parte inéditos hasta entonces. Nos enfrentamos ante dos problemas importantes desde décadas: por un lado, la obsolescencia de los formatos, que dificulta la preservación de los datos, cuando no provoca pérdidas, sobre todo si se utilizan formatos propietarios; por el otro lado, la casi prácticamente ausencia de interoperabilidad, es decir, de comunicación entre programas informáticos sin mediación humana. La emergencia a finales de los ochenta del siglo pasado de las *Directrices* TEI y del lenguaje de marcado XML a finales de los noventa se explica, en gran parte, debido a estos dos desafíos (Hockey, 2000).

Es por esta razón que el editor académico que pretenda llevar a cabo una edición digital académica (EDA)[3] considera a menudo que su principal cometido es identificar con metadatos la fuente de la que deriva el texto y representar el texto y las intervenciones editoriales siguiendo las prácticas más o menos consensuadas por la comunidad internacional en un formato estandarizado y no propietario que sea fácilmente convertible a otros formatos. Ahora bien, tras la representación de los datos en formato electrónico, sigue la publicación digital, la última fase del flujo de trabajo. Cuál es el papel del editor académico aquí sigue siendo, en gran parte, una pregunta difícil de contestar. Por lo general, la figura del editor desaparece

[1] Sobre si la edición digital académica supone una revolución verdaderamente, ver Karlsson y Malm (2004) y Bordalejo (2018). En adelante y para acortar, llamaré EDA a la edición digital académica.

[2] Debo esta idea a mi director de tesis, José María Micó, cabal defensor de las «razones de la filología» (Micó, 1999). Sobre la situación de la crítica textual en la era digital, véase Allés-Torrent (2020).

[3] Sigo aquí la ya clásica definición de Patrick Sahle, según la cual las EDAs son «ediciones académicas que se guían por un paradigma digital en su teoría, método y práctica». Es decir, una edición *digitalizada* no cumple con los requisitos definidos y una edición digital académica verdadera no podría imprimirse sin una pérdida significativa de contenido y funcionalidades (2017). Para una revisión del concepto de EDA, véase Alvite-Díez y Rojas Castro (2022).

en fondo cuando se empieza a hablar de desarrollo web, programación o bases de datos. Ciertamente, en la publicación de libros en formato impreso, las actividades de maquetación, impresión y distribución son realizadas por las casas editoriales en colaboración con impresores y librerías. ¿Por qué entonces debería el editor formar parte del proceso de publicación digital?

Además, la publicación de una EDA es, sin duda, el fruto del trabajo en equipo (editores, programadores, bibliotecarios, diseñadores gráficos, etc.) que toman decisiones basadas en distintos factores, experiencias pasadas, competencias y preferencias. La misma elección de las tecnologías con las que se desarrollan las aplicaciones web depende en gran medida del repertorio actual de los lenguajes informáticos, formatos y programas disponibles en el momento de publicación, así como de la infraestructura institucional de la universidad o de la biblioteca en la que se desarrolla la EDA.

En proyectos editoriales pequeños a menudo la edición del texto, el desarrollo web y el diseño gráfico son realizados por una sola persona y, por tanto, para aligerar la carga de trabajo, se pueden adoptar herramientas de publicación que transforman de manera más o menos estandarizada el documento XML a formato HTML y añaden estilo, formato e interactividad. El resultado obtenido suele ser una interfaz gráfica de usuario que puede adaptarse y personalizarse a las necesidades de cada proyecto editorial.

Tras más de tres décadas de experiencia navegando por la web, sabemos que las interfaces deben ser fáciles de utilizar por todo el mundo, con independencia del *hardware*, *software* o las habilidades de los usuarios. Por desgracia, esto no siempre ocurre de manera óptima al analizar las EDAs porque las herramientas de publicación existentes actualmente pueden condicionar la apariencia, las funcionalidades, el uso, el acceso y la preservación a largo plazo. El objetivo de este artículo es precisamente contribuir al análisis crítico de las herramientas de publicación de EDAs existentes actualmente.

Este artículo se divide en tres apartados principales: en el primero se expone una panorámica de la publicación web poniendo el énfasis en los aspectos tecnológicos y en las diferencias existentes entre las webs

dinámicas y las webs estáticas; en el segundo apartado se reflexiona sobre algunos desafíos que supone el uso de herramientas de publicación y los efectos derivados de la estandarización de tecnologías; finalmente, en el último apartado, se analiza la publicación web de las *Soledades* de Luis de Góngora con la herramienta Edition Visualization Technology poniendo especial atención en el acceso, la identificación y la exploración.

La publicación web como proceso tecnológico

Aunque el papel del editor se puede limitar al proceso de establecimiento del texto y de modelado o representación, a menudo también debe decidir sobre la publicación en sentido amplio: ¿cómo se quiere presentar el texto? ¿Cómo va a interactuar el usuario con la información a través del navegador web? ¿En qué dispositivos se puede leer? ¿Es necesario proporcionar otros formatos de descarga como ePUB y PDF? Este tipo de preguntas no son baladíes y no se deberían delegar solamente en informáticos, pues, en realidad, las interfaces gráficas de usuario condicionan el acceso al texto (Bleier *et al.*, 2018).

Como defiende Pierazzo (2015) las decisiones editoriales dependen de muchos factores, de entre los cuales podemos contar el estado de las tecnologías, el tiempo y el presupuesto disponibles. Cuando hablamos de publicación web, los programas informáticos reparten las tareas entre los proveedores de los recursos o servicios, llamados servidores, y los demandantes, llamados clientes. El funcionamiento es fácil de entender si se considera como un diálogo: un cliente realiza peticiones a otro programa, el servidor, que le da respuesta siguiendo protocolos estandarizados de intercambio de información como HTTP y el lenguaje de marcado HTML.

Desde inicios del siglo XXI, la mayoría de las páginas webs a las que accedemos habitualmente son «dinámicas», es decir, son sitios que se construyen principalmente con un programa alojado en la parte del servidor (por ejemplo, una base de datos) en combinación con otros procesos que se ejecutan por el lado del cliente. Por este motivo, permiten un nivel de interactividad elevado; el contenido (texto, imágenes, etc.) se

modifica en función del contexto y las condiciones, poniendo a dialogar tres capas de información: el *front-end* (interfaz), *middle tier* (intermediario o intérprete) y el *back-end* (almacenamiento) (Birnbaum *et al.*, 2019).

El *front-end* (es decir, la presentación o interfaz gráfica de usuario) corresponde a la presentación del contenido en el navegador web y suele componerse mediante las tecnologías ya mencionadas (HTML, CSS y JavaScript); el intermediario o intérprete controla la interacción entre el usuario y los datos, y corresponde con lenguajes de programación que aplican lógica como PHP, Ruby o Python; por último, el *back end* o almacenamiento suele ser una base de datos, que indexa la información y recupera partes de los documentos, entre otras tareas relacionadas con el filtrado y la administración de usuarios. Así, pues, en términos generales, el navegador web manda peticiones a la capa intermedia y esta responde ofreciendo servicios valiéndose de consultas y actualizaciones a la base de datos y proporcionando una interfaz de usuario.

Aunque las bases de datos relacionales dominan el sector profesional, en los últimos años, el uso de bases de datos XML como eXistDB se ha extendido en la comunidad TEI. eXistDB es un programa de código abierto creado en 2000 por Wolfgang Meier centrado en la gestión de documentos XML que, a diferencia de la mayoría de las bases de datos existentes, utiliza XQuery y XSLT como lenguajes de programación (Meier, 2003). La principal ventaja de las bases de datos XML nativas es que el desarrollo web se limita al uso de XQuery para transformar la información de un formato a otro en lugar de usar varias tecnologías (Birnbaum *et al.*, 2019). Por ejemplo, es bastante común en la mayoría de sitios webs utilizar una base de datos relacional construida con MySQL y utilizar el lenguaje de programación PHP para interrogar la información y procesar peticiones entre el cliente y el servidor. Sin embargo, la desventaja principal es que este flujo de trabajo presupone el soporte técnico para instalar y mantener la base de datos, algo que no siempre es posible en proyectos individuales o en instituciones académicas de tamaño pequeño o mediano.

Cuando la interactividad de una web no tiene lugar en el servidor web mediante un programa informático, sino en la parte del cliente, en el navegador web, interpretando las instrucciones basadas en HTML, CSS y

JavaScript, estamos ante lo que se conoce como una «web estática», es decir, una web que se visualiza en el navegador tal y como se almacena en el servidor.

Como es lógico, esta alternativa tiene como objetivo principal dar acceso a uno o varios textos, más que facilitar búsquedas complejas o la exploración del texto de manera dinámica en la línea de lo que Pierazzo ha llamado ediciones *prêt-à-porter* (Pierazzo, 2019). En algunos escenarios, como cuando no se dispone de grandes recursos para contratar a una persona encargada del desarrollo web (por ejemplo, piénsese en jóvenes investigadores sin financiación) o bien de una conexión rápida que procese de manera óptima muchas peticiones, una web estática puede ser una solución adecuada. Además, tiene la ventaja de que no requiere grandes costes de mantenimiento a largo plazo.

Así, por ejemplo, muchas de las «ediciones mínimas» publicadas en los últimos cinco años siguiendo la filosofía *minimal computing* pueden considerarse webs estáticas (Gil y Ortega, 2016; Risam, 2019). Por lo común, se han llevado a cabo dos formas de publicación: por un lado, se puede transformar el *input* en XML a formato Markdown y luego Jekyll transforma el documento a HTML y añade CSS para darle formato al texto y JavaScript para ejecutar acciones (por ejemplo, filtrar el contenido)[4]; por el otro, se puede emplear algún mecanismo que indica a los navegadores web cómo interpretar los elementos TEI como si fuera HTML, por ejemplo, utilizando CETEIcean, es decir, sin necesidad de transformar en fichero XML a otro formato de publicación. Estas ediciones, además, tienen la particularidad de almacenar todos los ficheros no en un servidor web alojado en una institución particular sino en un repositorio de GitHub o GitLab, que suele ser gratuito y abierto.

Tanto si se desarrolla una web estática como una web dinámica (con base de datos relacional o XML nativa), lo que está claro para el editor que participe en el proceso de publicación debe distinguir entre formatos estándares de preservación (XML) y formatos estándares de publicación (HTML, ePUB, PDF). Además, como la interfaz gráfica de las ediciones digitales en gran parte condiciona la recepción del texto, el editor también

[4] En español, véase el trabajo de Rio Riande (2019).

debe formar parte de la toma de decisiones en colaboración con otros especialistas.

Herramientas de publicación

Cuando no se tienen conocimientos informáticos avanzados o el proyecto editorial cuenta con pocos investigadores o presupuesto, las herramientas genéricas pueden facilitar y agilizar el proceso de publicación. En la comunidad TEI existe el consenso de que uno de los principales obstáculos que frena la creación de ediciones digitales es la ausencia de tecnologías fáciles de utilizar y adaptadas al proceso de edición, análisis y publicación (Burghart y Rehbein, 2012). Esta situación ha sido descrita como una «paradoja» porque el desarrollo de *software* para la publicación de EDAs parece mitigado por el número reducido de usuarios y viceversa:

> Esta situación parece estar definida por lo que podría llamarse la «paradoja de las herramientas»: la necesidad de tales herramientas es bastante urgente, y hay varias herramientas disponibles; sin embargo, el número de implementaciones reales de las herramientas por parte de los editores académicos es todavía relativamente bajo. ¿Son las herramientas disponibles demasiado complicadas de utilizar o no son adecuadas para las ediciones académicas digitales, y qué se puede hacer al respecto? ¿Basta con desarrollar herramientas que hagan el trabajo para el que fueron concebidas o es igualmente importante crear una base sólida de usuarios? ¿Existe un círculo vicioso, una profecía autocumplida? (Pape, Schöch y Wegner, 2012, mi traducción).

Los desafíos que supone construir una herramienta de publicación son difíciles de sortear. Para empezar, deben lograr cierto equilibrio entre lo particular y lo genérico; por ejemplo, deben satisfacer las necesidades genéricas de un tipo de edición (edición crítica y/o diplomática-modernizada), pero, al mismo tiempo, permitir una mayor adaptación para que se adecúen a las particularidades de los proyectos individuales (Pape, Schöch y Wegner, 2012).

En este sentido, parece que en los últimos cinco años se ha

consolidado la tendencia a construir herramientas que hacen solo una tarea específica en lugar de intentar crear y mantener toda una estación filológica que cubra todo el flujo de trabajo editorial. Es lo que Pierazzo ha descrito, más recientemente, como una aproximación más modular y quizás menos ambiciosa, como si las herramientas si fueran los ladrillos con los que se levanta una casa a medida de sus huéspedes:

> La idea que subyace a este planteamiento es bastante sencilla: si bien es imposible desarrollar un marco único que satisfaga todos los posibles casos de uso que caracterizan el abigarrado panorama de la erudición textual, parece relativamente más fácil destacar las «microtareas» que se llevan a cabo en la mayoría de las empresas editoriales y construir herramientas que las soporten; luego será el usuario quien deba combinar y personalizar las herramientas de forma que se adapten al flujo de trabajo específico (Pierazzo, 2015, mi traducción).

Siguiendo, pues, esta tendencia muchas de las herramientas utilizadas para publicar ediciones digitales académicas realizan una tarea principal: transformar los documentos TEI a formato HTML para visualizar el texto con un navegador web y añadir funcionalidades en la interfaz gráfica de usuario. Sin embargo, el modo en que llevan a cabo dicha transformación suele ser distinto: algunas herramientas como TEIViewer y Versioning Machine utilizan XSLT para transformar el documento XML a HTML por la parte del cliente y añaden estilo con CSS e interactividad con JavaScript; en cambio, otras herramientas como TEICHI utilizan una base de datos instalada en el servidor web para almacenar e indexar los documentos TEI y un intermediario para generar el *front-end*. Tal es el caso de TEICHI (Pape, Schöch, y Wegner, 2012) y de otra herramienta aparecida más recientemente llamada TEIPublisher.

En cuanto a las funcionalidades añadidas, es manifiesto un efecto derivado del uso de herramientas: las interfaces gráficas de usuarios son cada vez más parecidas y cumplen con unos requerimientos mínimos como la navegación, el uso de hipertextos, la manipulación de imágenes facsimilares (por ejemplo, mediante *zoom-in*), la búsqueda de palabras, etc. Esta convergencia ha hecho pensar a algunos investigadores como Pierazzo (2015) que estamos experimentando un proceso de estandarización no solamente de los datos sino también de su presentación

en formato digital.

Por último, con independencia de si son genéricas o muy específicas, del modo de transformación de los documentos XML y de las funcionalidades añadidas, las herramientas de publicación también deben ser sostenibles a largo plazo, lo cual depende de las tecnologías elegidas (si requieren poco mantenimiento y actualizaciones) y de la financiación disponible para contratar el personal y pagar otros costes[5]. Pero, como señala Roberto Rosselli del Turco (2019), los desafíos no son solo técnicos; también hay una carencia de reflexión teórica sobre las limitaciones que imponen las herramientas al modelado, a la codificación, la visualización, la usabilidad, la accesibilidad, etc.

Edition Visualization Technology y la publicación de las *Soledades*

Actualmente, existen varios criterios de evaluación de EDAs publicados por revistas como RIDE (Sahle y Vogeler, 2016) u organizaciones como la MLA (2016). Aunque muy elogiables en sus intenciones, estos criterios son excesivamente extensos y detallados porque están pensados para reseñar o bien son demasiado abstractos o difíciles de probar ya que la evaluación depende en gran parte de la competencia del usuario. Además, ambos han sido desarrollados en paralelo a la herramienta Edition Visualization Technology (EVT). Por último, los criterios de evaluación son el fruto de un trabajo teórico, en gran parte idealista, que se quiere objetivo y descriptivo, pero que acaba siendo prescriptivo, y que no tiene en cuenta factores como el contexto, los recursos disponibles, la infraestructura o la experiencia previa del editor, por lo que si se aplican de manera categórica e inflexible pueden hacer poca justicia a una herramienta innovadora y gratuita como EVT o a un proyecto editorial con fines didácticos –fruto de un doctorado en Humanidades y sin ningún tipo de financiación o soporte técnico– como la EDA analizada aquí como estudio de caso.

Por estos motivos se prefiere proponer una especie de itinerario de

[5] Esta dificultad puede ejemplificarse con la herramienta de publicación TEICHI y el portal TAPAS, que llevan tiempo sin actualizarse y no se sabe a ciencia cierta si siguen operativos o bajo mínimos.

lectura basado en las tres acciones principales –acceso, identificación y exploración– que todo usuario competente realiza cuando visita una EDA: en primer lugar, acceder a la EDA y al texto editado; en segundo lugar, identificar el recurso, cómo se ha llevado a cabo y cuáles son sus fines; por último, leer, navegar, consultar o explorar el contenido de manera interactiva. Aunque se presentan aquí de manera sucesiva, las tres acciones forman parte de un proceso en bucle difícil de compartimentar y que puede variar según los intereses, las expectativas y la competencia técnica de cada usuario.

La EDA analizada en este artículo se titula *«Soledades» de Luis de Góngora. Edición crítica digital* y fue llevada a cabo por quien escribe estas líneas en dos fases: la primera, entre 2011 y 2015, tuvo lugar en la Universitat Pompeu Fabra de Barcelona (España), como parte de mi tesis de doctorado en Humanidades y bajo la dirección del profesor José María Micó; se estudió la transmisión textual del poema gongorino, se cotejó manualmente el manuscrito Chacón (utilizado como texto base) con una veintena de manuscritos e impresos, se afilió genealógicamente algunas de las fuentes y se estableció el texto crítico tras analizar las variantes encontradas y normalizar la ortografía, la separación de palabras y el uso de mayúsculas. El texto resultante fue representado con lenguaje de marcado XML siguiendo las *Directrices* TEI.

Entre 2016 y 2017, tras defender mi tesis doctoral y empezar a trabajar en la Universidad de Colonia, el acceso a la infraestructura del Cologne Center for e-Humanities y a los servidores de la institución posibilitó la publicación de la edición digital en la web. A fin de agilizar el proceso y dado que no se disponía de presupuesto alguno, se utilizó la herramienta Edition Visualization Technology (EVT). Actualmente, la edición digital sigue hospedada en los servidores de la Universidad de Colonia y es accesible en línea, pese a que mi afiliación ha cambiado posteriormente[6].

Desde un punto de vista filológico, editar las *Soledades* supone varios retos: en primer lugar, la datación del poema sigue poco certera, sobre todo, si distinguimos entre creación y difusión manuscrita y tenemos en

[6] La URL actual es: <http://soledades.uni-koeln.de/#/critical?d=doc_1&e=critical> (cons. 31/05/2022).

cuenta que las dos partes del poema pudieron tener distintas cronologías, con algunos años de diferencia; en segundo lugar, el poema plantea un problema radical sobre el proceso de escritura, los estadios del texto y la intervención de juicios ajenos al autor, como el de Pedro de Valencia, poniendo de relieve cómo la escritura no fue tanto un acto solitario e inspirado por parte de un genio sino un proceso dilatado en el tiempo que permitía al poeta compartir avances a fin de obtener el «parecer» de su círculo de confianza; por último, dado que no disponemos de manuscrito autógrafos de Góngora, solo podemos tener acceso a los distintos estadios del texto a partir de copia manuscritas y a impresos tardíos o póstumos, que están dispersos por varias bibliotecas españolas y europeas.

Profundizando en esta última cuestión, el estudio del manuscrito 2056 de la Biblioteca Nacional de Catalunya arrojó nueva luz sobre el proceso de escritura del poema ya que nos permitió encontrar variantes de autor inéditas hasta la publicación de la edición digital. Como se ha expuesto en trabajos anteriores (Rojas Castro, 2018), el manuscrito 2056 contiene correcciones autógrafas del mismo Góngora y transmite seis pasajes de la *Soledad segunda* que varían respecto al texto transmitido por el manuscrito Chacón, más tardío y que sirve de texto base de todas las ediciones modernas del poema. Los seis pasajes obedecen, sin duda, al arranque poético anterior a los consejos del humanista Pedro de Valencia y responden a la voluntad de perfeccionar el poema intensificando una idea o bien creando paralelismos con otros pasajes de la primera parte del poema.

Durante la primera fase señalada arriba, tras estudiar la transmisión textual, cotejar los testimonios y establecer el texto, se creó un fichero XML y se representó la siguiente información siguiendo las *Directrices* TEI en el encabezado: en primer lugar, se identificó el recurso digital mediante metadatos descriptivos cubriendo aspectos como el título, el autor, el editor, la financiación, la institución responsable de la publicación, la fecha y lugar de publicación, y la licencia; en segundo lugar, se identificó de manera clara la fuente principal de la que deriva el texto, el manuscrito Chacón, proporcionando su signatura (Res/45, 1, Biblioteca Nacional de España) y se describió de manera estructurada los 21 testimonios cotejados; en tercer lugar, se proporcionó información sobre los fines de

la edición digital y los criterios editoriales y de codificación en el mismo fichero XML/TEI como parte del encabezado TEI; en cuarto lugar, se documentaron las principales fases de creación del poema en forma de metadatos.

La codificación del texto de las *Soledades* puede analizarse en dos partes: por un lado, se marcó con etiquetado TEI las tres partes principales del texto y sus títulos correspondientes (Dedicatoria, Soledad primera y Soledad segunda), los grupos de versos (aunque la silva no es una forma estrófica *per se*, en el poema gongorino se pueden identificar grupo de versos) y los versos. Por el otro lado, se etiquetó la variación textual mediante un aparato crítico empleando el método de segmentación paralela (*parallel segmentation method*), insertando las variantes en los lugares del texto base en que se encuentra una variante y a fin de distinguir entre variantes de copista y variantes de autor[7].

Como se ha dicho más arriba, durante la segunda fase del proyecto, se utilizó la herramienta Edition Visualization Technology (EVT), versión 2 (BETA 1)[8], para publicar la edición académica digital de las *Soledades*. Esta tecnología ha sido desarrollada bajo la dirección de Rosselli del Turco y se caracteriza porque transforma a HTML un documento TEI que contiene un aparato crítico visualizando las variantes en una ventana emergente, imitando la disposición, más o menos clásica, de una edición impresa o bien cada una de las versiones en ventanas paralelas tras sustituir el lema por la variante en cuestión. A diferencia de EVT1, la transformación de XML a HTML se lleva a cabo mediante JavaScript para transformar el *input* en JSON y se emplea el *framework* Angular para crear el *front-end* añadiendo estilos e interactividad. En ambas versiones, la transformación del documento TEI tiene lugar por la parte del cliente y, por tanto, no es necesario instalar una base de datos en el servidor web. Publicada con una licencia abierta (General Public Licence v3.0), la herramienta EVT2 ha sido utilizada con éxito en varios proyectos

[7] Véase Rojas Castro (2017) para una descripción más completa del método de codificación TEI del aparato crítico. Se puede consultar el fichero XML/TEI en Github: <https://github.com/arojascastro/soledades/blob/master/edicion/critica/soledades_critica.xml> (cons. 31/05/2022).

[8] Actualmente existe una versión BETA 2 publicada en julio de 2020. Cabe señalar, pues, que esta herramienta aún se encuentra en fase de desarrollo y que aún no ha alcanzado una versión 1.0, más estable.

editoriales como, por ejemplo, el proyecto Garrett Online, entre otros[9].

El resultado obtenido es una EDA que sigue el paradigma digital porque el contenido no se podría imprimir en papel sin una pérdida sustancial (Sahle, 2017). Esto se debe a que la publicación tiene un componente interactivo muy alto que permite al usuario seleccionar no solamente cómo visualizar el aparato de variantes mediante una ventana emergente sino también comparar el texto crítico con el testimonio seleccionado. Gracias al método de codificación TEI, resulta relativamente fácil para EVT2 sustituir el lema por la variante y así generar los textos que varían respecto al manuscrito Chacón. En otras palabras, se trata de una «edición digital paradigmática» (Pierazzo, 2015) porque a partir de un único fichero XML/TEI (*input*) es posible generar varios textos (*outputs*).

Al visitar las *Soledades* publicadas con EVT2, el usuario no tiene que navegar por el contenido ni seleccionar o filtrar parámetros porque el texto crítico es lo primero que aparece.

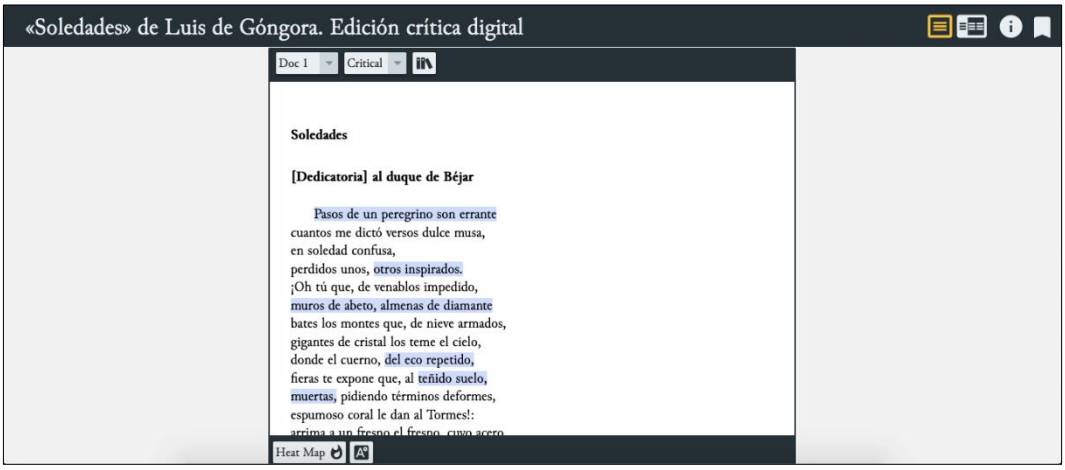


Fig. 1. Página de inicio de *«Soledades» de Luis de Góngora. Edición crítica digital»*

Como se puede ver en la figura 1, el texto crítico aparece en el centro en una caja con el fondo en blanco debajo del título de la publicación –

[9] <https://garrettonline.romanceiro.pt/romanceiro/livro-i/7-o-anjo-e-a-princesa/> (cons. 30/05/2022).

Soledades» de Luis de Góngora. Edición crítica digital–; la apariencia es convencional e imita, en gran medida, la página del libro impreso: los títulos de apartado están destacados en negrita, el verso inicial está ligeramente sangrado hacia la derecha, etc. La única novedad evidente es que el usuario debe desplazarse hacia abajo o hacia arriba para leer el texto en lugar de pasar páginas.

Tras el acceso al texto, siguen la identificación y la valoración de la EDA. La información necesaria para realizar estas tareas se encuentra clicando en el botón redondo que contienen una *I* mayúscula (de «Information») que aparece en la parte superior derecha. Al clicar, como se aprecia en la figura 2, aparece una ventana emergente que se sitúa en el centro de la pantalla titulada «Project Info» y que recupera la información codificada en el encabezado del documento TEI en tres secciones distintas: en primer lugar, en la sección «File Description» se identifican el autor del texto, el editor, la financiación, la fecha de la edición, la entidad responsable de la publicación, la licencia de uso y la fuente de la que deriva el texto; en segundo lugar, en la sección «Encoding Description» se proporciona información sobre la finalidad del proyecto, los criterios editoriales y de codificación y el método de codificación TEI del aparato de variantes; por último, en la sección «Text Profile» se documenta información sobre la fecha de creación y sobre el idioma del texto.

Las ediciones digitales académicas publicadas con EVT2 también son fácilmente citables porque la interfaz gráfica de usuario proporciona por defecto un botón en forma de marcador de página situado en el extremo derecho del menú. Como se puede apreciar en la figura 3, al clicar sobre el botón vuelve a emerger una ventana situada en el centro de la pantalla con una sugerencia de citación. En el caso de la edición de las *Soledades*, la citación sugerida contiene información el autor del texto, el título, la entidad responsable de la publicación y la URL. Sin embargo, no proporciona otros datos esperables como el nombre del editor o la fecha de publicación. En este sentido, la EDA analizada requeriría de algunas modificaciones para que la citación sugerida fuera realmente útil para el usuario. Con todo, cabe decir que, como se ha visto, la fecha de publicación y el editor aparecen en el apartado «Project Info» y, por tanto, el usuario interesado podría completar fácilmente los datos que faltan.

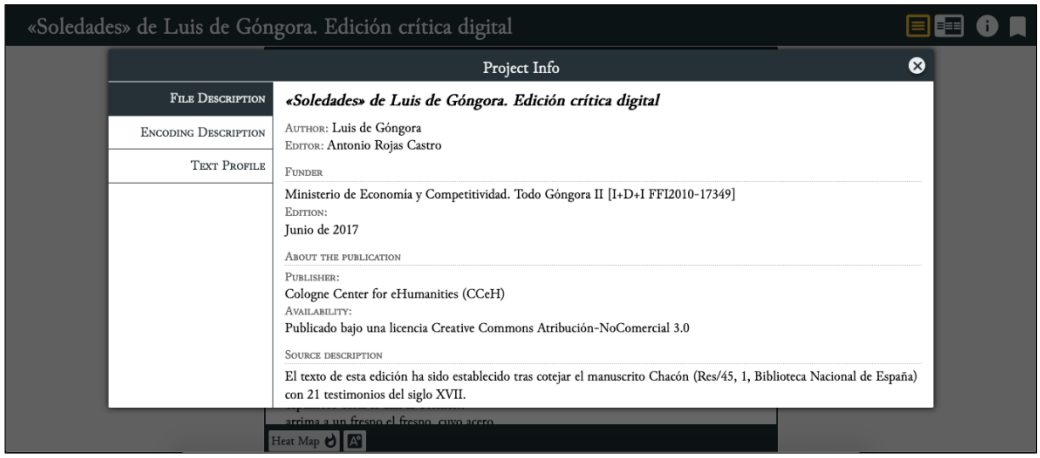


Fig. 2. Sección «Project Info» con los metadatos del proyecto editorial

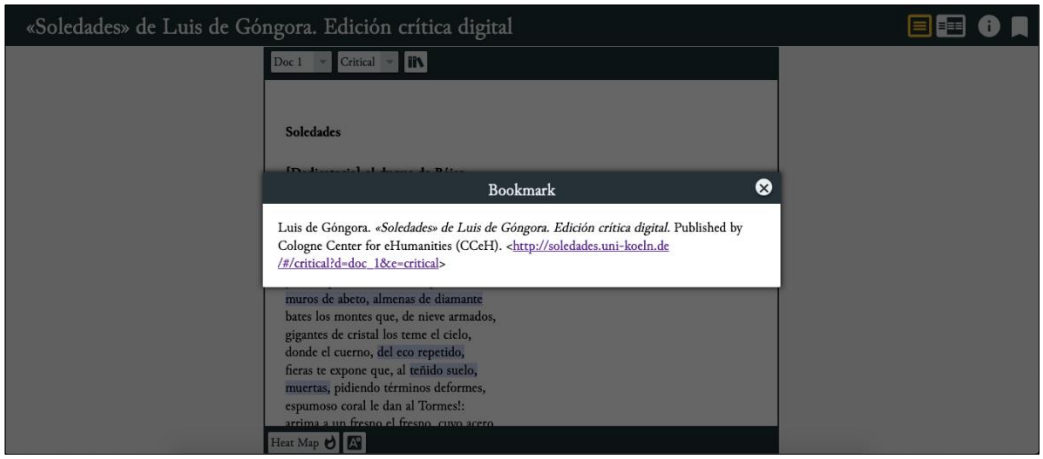


Fig. 3. Citación sugerida o «Bookmark»

Una vez ha conseguido identificar la EDA, el usuario puede volver a leer el texto y a explorar el contenido cambiando su presentación de múltiples maneras. En el caso de las EDAs publicadas con EVT2, lo primero que llama la atención es que algunos versos están destacados en colores (en azul por defecto). Al clicar sobre uno de estos versos, se visualiza una ventana emergente con el aparato de variantes y se puede

activar opcionalmente la vista del código XML, tal y como se percibe en la figura 4. En el caso de las *Soledades*, si se clica sobre el botón «Heat Map», situado en la parte inferior de la interfaz, los versos pasan a visualizarse en distintos tonos de color verde. Aunque sea difícil de comprender en un primer momento, basta explorar el contenido del aparato de variantes para que el usuario se percate de que la intensidad depende del número de entradas en el aparato de variantes. Es decir, a mayor número de variantes por aparato, más intenso es el color. Si bien este aspecto visual puede ocasionar problemas de accesibilidad, creemos que resulta importante para el usuario porque posibilita el «escaneo» del texto por zonas.

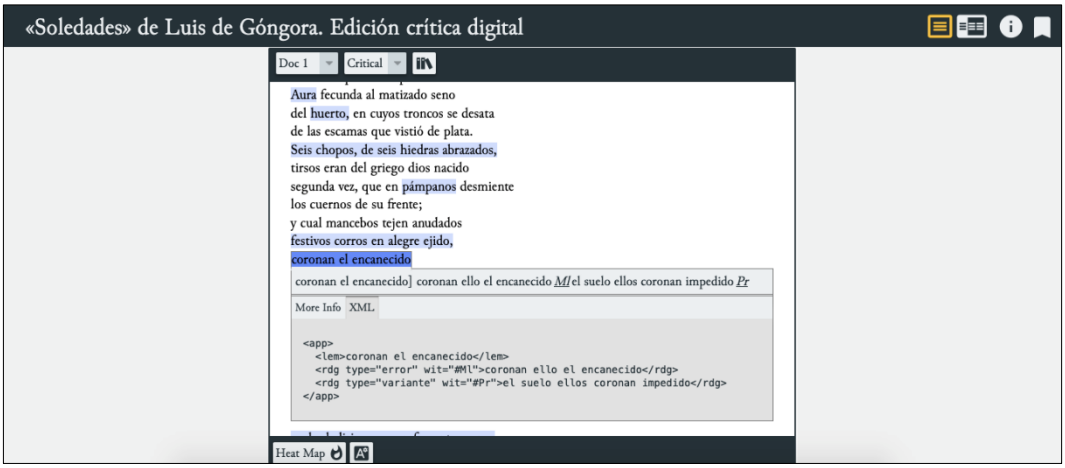


Fig. 4. Ejemplo de visualización del aparato de variantes

Pero EVT2 no solo permite visualizar el aparato en una ventana emergente, más o menos como si fuera el pie de página de una edición impresa, sino que si se clica en el segundo botón de la parte superior derecha –un cuadrado divido en dos partes– es posible dividir la interfaz gráfica en varias ventanas y seleccionar un testimonio. En la edición de las *Soledades* analizada aquí, el texto crítico se visualiza en la ventana izquierda mientras que en la ventana situada en la derecha de la pantalla se activa la opción de seleccionar uno de los 22 testimonios disponibles a partir de la sigla y del país en que se encuentra el documento. En la figura 5 se puede percibir cómo se visualiza el texto crítico en la ventana derecha con los

versos que transmiten variación textual destacados en azul y el texto transmitido por el testimonio *Pr*. Este testimonio corresponde al manuscrito 2056 de la Biblioteca de Catalunya y contiene, como ya se dijo más arriba, seis variantes de autor en la *Soledad segunda*. Si el usuario activa la función «Filters» y selecciona el tipo de variante que desea visualizar, se destacan en color verde los versos que transmiten una variante de autor y en color rojo los que transmiten un error de copia.

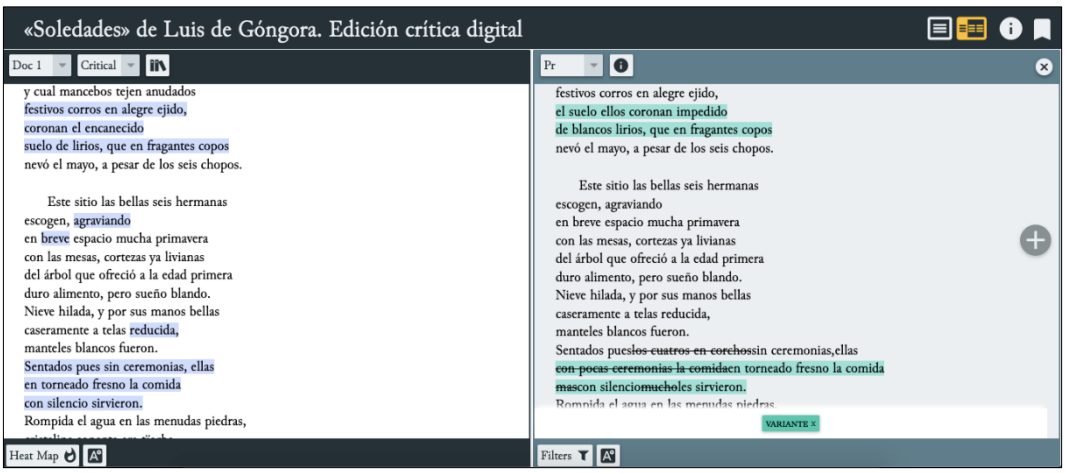


Fig. 5. Reconstrucción del texto transmitido por el testimonio *Pr*

Es posible añadir otros testimonios de modo que la interfaz gráfica se divida de manera sucesiva en ventanas cada vez más estrechas en función del tamaño de la pantalla utilizada; se presupone, por tanto, que se accede a la EDA a través de un ordenador portátil o de sobremesa y no a través de un teléfono móvil de dimensiones reducidas. Asimismo, debido a varias causas, algunas de las cuales son específicas del texto codificado (extensión del poema y número elevado de variantes) y otras a la tecnología empleada (recuérdese que EVT no almacena ni transforma los ficheros XML a HTML utilizando una base de datos), la visualización tarda entre 5 y 10 segundos y, por tanto, el usuario debe esperar impaciente a obtener el resultado de la transformación consistente en la sustitución del lema por las variantes transmitidas en el testimonio seleccionado.

Por último, como se muestra en la figura 6, es posible clicar en el

botón redondo con una *i* minúscula para acceder a la descripción del testimonio *Pr*. De esta manera, ocultando el texto y visualizando los metadatos contenidos en el encabezado TEI, el usuario obtiene más información sobre el repositorio que preserva el documento, su signatura, localización, y una lista de notas con algunas particularidades sobre la variación textual como, por ejemplo, lugares tachados e ilegibles.

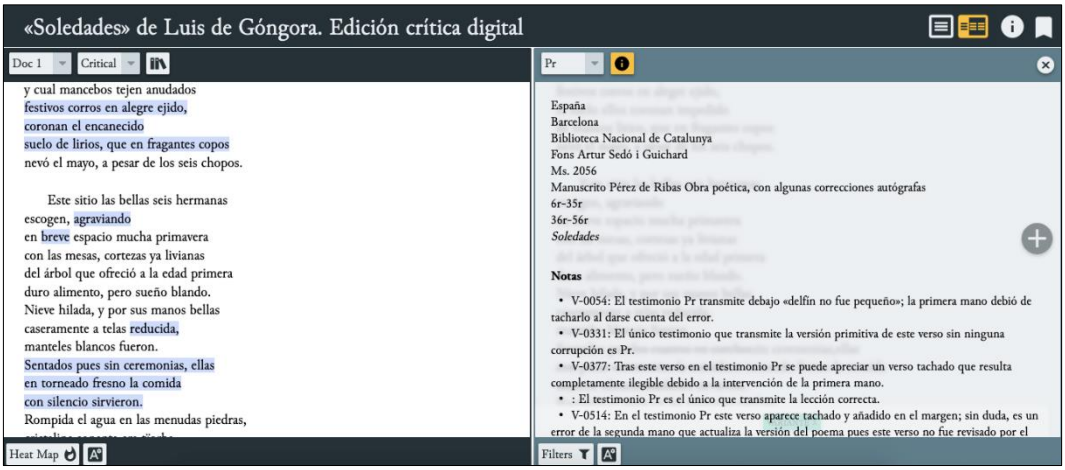


Fig. 6. Descripción bibliográfica del testimonio *Pr*

Tras este itinerario de lectura basado en el acceso, la identificación y la exploración, el usuario puede ir y volver sobre el material textual, los paratextos y las funcionalidades como los filtros a fin de comparar el texto crítico con los testimonios cotejados. Se trata, por tanto, de una recepción menos lineal en comparación con la lectura atenta propiciada por la edición impresa y más transparente porque el usuario puede evaluar con exhaustividad las decisiones editoriales relativas al establecimiento del texto crítico.

A pesar de que algunos aspectos son claramente mejorables, como la accesibilidad (uso de una pantalla de dimensiones grandes para acceder al contenido y de colores con valor semántico), la citabilidad o la rapidez de respuesta durante el proceso de transformación, la EDA publicada con EVT2 cumple con sus objetivos: da acceso al texto crítico, permite identificar el recurso y las fuentes utilizadas, y posibilita una lectura

interactiva en la que el usuario puede explorar el contenido mediante distintas visualizaciones y enjuiciar las decisiones editoriales. Y todo ello de manera gratuita, utilizando tecnologías de código abierto y beneficiándose del alto grado de interoperabilidad que goza el formato XML/TEI para representar ediciones críticas.

Conclusiones

Con este artículo hemos defendido que el rol del editor no debería limitarse al establecimiento del texto y su representación con lenguaje de marcado: la presentación del texto y la publicación web también forman parte de la labor editorial, sobre todo se si emplean herramientas de publicación. Seleccionar textos, leerlos y analizarlos, revisarlos, transcribirlos, corregirlos, normalizarlos, anotarlos, representarlos y darles formato de manera adecuada son actividades enraizadas en la filología desde los primeros destellos del Humanismo italiano, pero que deben complementarse con programas informáticos para agilizar o automatizar algunos procesos y con el conocimiento de nuevos formatos de preservación, representación y publicación.

Actualmente, poseer nociones básicas de desarrollo web, programación y bases de datos es un imperativo para toda persona que colabore en la construcción de ediciones digitales académicas. La publicación de textos en línea es un proceso tecnológico que afecta no solamente a la presentación sino también al acceso, uso e interacción. No solamente es necesario conocer la diferencia entre un servidor y un cliente sino también qué es el *front-end* y el *back-end* de un sitio web o en qué se diferencia una web dinámica de otra considerada «estática». Esto es especialmente relevante porque las EDAs han sido construidas tradicionalmente, sobre todo en Europa y Estados Unidos, utilizando bases de datos como eXistDB. Sin embargo, en los últimos cinco años, existe una tendencia –la filosofía *minimal computing*– en el campo de las Humanidades Digitales consistente en no utilizar bases de datos de ningún tipo y en su lugar publicar las EDAs como webs estáticas, ya que son más fáciles de mantener y preservar a largo plazo, si bien requieren otro

conjunto de conocimientos y habilidades informáticas.

Esta problemática afecta de manera directa a los proyectos editoriales con poca financiación o bien individuales como pueden ser las EDAs que se desarrollan con fines didácticos o como parte de un programa de doctorado. En tales circunstancias, dado que no se suele disponer de recursos para contratar a un especialista en desarrollo web, es común el uso de herramientas de publicación como las analizadas en el apartado tercero de este artículo. Los retos existentes en este ámbito también son numerosos: el tamaño reducido de la comunidad de usuarios no propicia la inversión en su desarrollo, el nivel de personalización permitido a veces requiere mayor conocimiento técnico y la falta de financiación a largo plazo desemboca en problemas de mantenimiento. Pese a ello, parece que hay futuro para el desarrollo de herramientas de publicación siempre y cuando sea modular y se centre en una o pocas tareas, como la transformación de documentos XML/TEI a HTML. El resultado suele ser una interfaz gráfica de usuario con una funcionalidad más o menos estándar que permite navegar por la web, acceder al texto, explorarlo, compararlo con otras visualizaciones y buscar palabras.

A modo de ilustración, en este artículo se ha analizado la publicación web de las *Soledades* de Luis de Góngora realizada con la herramienta EVT, en concreto, la versión 2 (BETA 1). Tras exponer las dos fases en las que se estudió la transmisión textual del poema, se cotejó una veintena de testimonios, se estableció el texto crítico y se representó toda la información en un documento XML/TEI, se ha defendido que el resultado obtenido con EVT2 es una EDA que sigue el paradigma digital porque si se imprimiera en papel perdería gran parte del contenido y de las funcionalidades. Para demostrar esto, se ha llevado a cabo un itinerario de lectura centrado en el acceso, la identificación y la exploración –al fin y al cabo, en la *experiencia de usuario* intencionada–. Pese a que algunos aspectos son mejorables como el tiempo de respuesta, la EDA analizada aquí permite acceder al texto crítico de manera sencilla y directa, proporciona una sección específica para identificar el recurso y las fuentes de la que deriva el texto, entre otros metadatos, y, sobre todo, permite una exploración del contenido verdaderamente interactiva ya que el usuario puede decidir cómo visualizar la información de distintas maneras

potenciando la transparencia de las decisiones editoriales.

§

Bibliografía citada

Allés-Torrent, Susanna, «Crítica textual y edición digital o ¿dónde está la crítica en las ediciones digitales?» *Studia Aurea* 14 *Veinte años de Imprenta y Crítica textual en el Siglo de Oro* (2020), 63-98. DOI: < https://doi.org/10.5565/rev/studiaaurea.395> (cons. 31/05/2022).

Alvite-Díez, María-Luisa y Antonio Rojas-Castro, «Ediciones digitales académicas: concepto, estándares de calidad y software de publicación», *Profesional de la información*, 31/2 (2022). DOI: <https://doi.org/10.3145/epi.2022.mar.16> (cons. 31/05/2022).

Birnbaum, David J., Hugh Cayless, Emmanuelle Morlock, Leif-Jöran Olsson y Joseph Wicentowski, «The integration of XML databases and content management systems in digital editions: Understanding eXist-db through Reese's Peanut Butter Cups», en *Balisage: The Markup Conference*, Washington DC, 2019. DOI: <https://doi.org/10.4242/BalisageVol23.Birnbaum01> (cons. 31/05/2022).

Bleier, Roman, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, y Gerlinde Schneider (eds.), *Digital Scholarly Editions as Interfaces*, Schriften Des Instituts Für Dokumentologie Und Editorik 12, Norderstedt: Books on Demand, 2018.

Bordalejo, Barbara, «Digital versus Analogue Textual Scholarship or The Revolution Is Just in the Title», *Digital Philology: A Journal of Medieval Cultures*, 7/1 (2018), pp. 7-28. DOI: <https://doi.org/10.1353/dph.2018.0001> (cons. 31/05/2022).

Burghart, Marjorie y Malte Rehbein, «The Present and Future of the TEI Community for Manuscript Encoding», *Journal of the Text Encoding Initiative*, 2 (2011). DOI: <https://doi.org/10.4000/jtei.372> (cons. 31/05/2022).

Gil, Alex y Élika Ortega, «Global outlooks in digital humanities. Multilingual Practices and Minimal Computing», en *Doing Digital Humanities. Practice, Training, Research*, eds. Constance Crompton, Richard J. Lane y Ray Siemens, London- New York, Routledge 2016, pp. 22-33.

Hockey, Susan, *Electronic texts in the humanities: principles and practice*, Oxford, Oxford University Press, 2000.

Karlsson, Lina y Linda Malm, «Revolution or Remediation? A Study of Electronic Scholarly Editions on the Web», *Human IT*, 7 (2014), pp. 1-46. URL: <http://etjanst.hb.se/bhs/ith/1-7/lklm.pdf> (cons. 31/05/2022).

Kuhn, Thomas S., *La estructura de las revoluciones científicas*, (trad.) Carlos Solís, México, D.F.: Fondo de Cultura Económica, 2013.

Meier, Wolfgang, «eXist: An Open Source Native XML Database», en *Web, Web-Services, and Database Systems*, eds. Akmal B. Chaudhri, Mario Jeckle, Erhard Rahm, y Rainer Unland, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg vol. 2593, 2003, pp.169-83. DOI: <https://doi.org/10.1007/3-540-36560-5_13> (cons. 31/05/2022).

Micó, José María, «Un verso de Góngora y las razones de la filología», *Criticón* 75 (1999), pp. 49-68.

MLA, «MLA Statement on the Scholarly Edition in the Digital Age», Modern Language Association of America, 2016. URL: <https://www.mla.org/Resources/Research/Surveys-Reports-and-Other-Documents/Other/Reports-from-the-MLA-Committee-on-Scholarly-Editions/MLA-Statement-on-the-Scholarly-Edition-in-the-Digital-Age> (cons. 31/05/2022).

Pape, Sebastian, Christof Schöch y Lutz Wegner, «TEICHI and the Tools Paradox: Developing a Publishing Framework for Digital Editions», *Journal of the Text Encoding Initiative*, 2 (2012). DOI: <https://doi.org/10.4000/jtei.432> (cons. 31/05/2022).

Pierazzo, Elena, *Digital Scholarly Editing: Theories, Models and Methods*, London, New York: Routledge, Taylor & Francis Group, 2015.

—, «What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter», *International Journal of Digital Humanities*, 1/2 (2019), pp. 209-220. DOI: <https://doi.org/10.1007/s42803-019-00019-3> (cons. 31/05/2022).

Rio Riande, Gimena del, «Humanidades Digitales, infraestructuras visibles e invisibles», *HD CAICYT LAB*, 2019. <https://doi.org/10.5565/rev/studiaaurea.395> (cons. 31/05/2022).

Risam, Roopika, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*, Evanston-Illinois, Northwestern University Press, 2019.

Rojas Castro, Antonio, «La edición crítica digital y la codificación TEI. Preliminares para una nueva edición de las Soledades de Luis de Góngora». *Revista de Humanidades Digitales* 1 (2019), pp. 4-19. DOI: <https://doi.org/10/gmb66p> (cons. 31/05/2022).

—, «Las Soledades de Luis de Góngora en el manuscrito 2056 de la Biblioteca de Catalunya: estudio bibliográfico y nuevas variantes de autor», *Rilce. Revista de Filología Hispánica*, 34/1 (2018), pp. 69-99. DOI: <https://doi.org/10.15581/008.34.1.69-99> (cons. 31/05/2022).

Rosselli del Turco, Roberto, «Designing an advanced software tool for Digital Scholarly Editions», *Textual Cultures*, 12/2 (2019), pp. 91-111 DOI: <https://doi.org/10.14434/textual.v12i2.27690> (cons. 31/05/2022).

Sahle, Patrick, «2. What Is a Scholarly Digital Edition?», en *Digital Scholarly Editing: Theories and Practices*, eds. Matthew James Driscoll y Elena Pierazzo, Digital Humanities Series, Cambridge, Open Book Publishers, 2017, pp. 19-39. URL: <http://books.openedition.org/obp/3397> (cons. 31/05/2022).

Sahle, Patrick y Georg Vogeler, «Criterios para la reseña de ediciones digitales académicas (EDA), versión 1.1», Institut für Dokumentologie und Editorik, 2016. URL: <https://www.i-de.de/publikationen/weitereschriften/criterios-version-1-1/> (cons. 31/05/2022).

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Avances en la creación de *BIDISO TEXTOS*. Edición académica digital de relaciones de sucesos

Carlota Fernández Travieso (Universidade da Coruña)

Manuel Garrobo Peral (Università di Verona / Universidade da Coruña)*

Abstract

Se informa de los avances en la configuración de un repositorio de textos asociado al proyecto *Biblioteca Digital Siglo de Oro* (BIDISO) –denominado *BIDISO Textos*–, en particular, en lo relativo a la colección de relaciones de sucesos. Se establecen aspectos clave de nuestro modelo de Edición Académica Digital para este género editorial informativo de la Edad Moderna relacionados con la delimitación de nuestro corpus, los destinatarios de las ediciones, los objetivos, los mecanismos para la obtención de los textos y la codificación, basada en el esquema TEI. Con ello, se realizan progresos para subsanar la carencia de textos de este género susceptibles de análisis sistemáticos precisos para realizar estudios estadísticos y cuantitativos rigurosos que permitan avanzar en el conocimiento de las relaciones de sucesos mediante la aplicación de nuevas tecnologías.
Palabras clave: Relaciones de sucesos; Edición académica digital; XML-TEI; Modelo de codificación; Siglos XVI-XVIII

Progress is reported on the configuration of a text repository associated with the *Biblioteca Digital Siglo de Oro* (BIDISO) project –called *BIDISO Textos*– related to the collection of *relaciones de sucesos*. Key aspects of our model of Digital Academic Edition of this informative publishing genre of the Modern Age are established. Such aspects are related to the delimitation of our corpus, the recipients of the editions, the objectives, the mechanisms for obtaining the texts and the encoding, based on the TEI scheme. We move forward to remedy the lack of texts of this genre that are susceptible to precise systematic analysis to carry out rigorous statistical and quantitative studies that allow us to advance in the knowledge of the *relaciones de sucesos* through the application of new technologies.
Keywords: News pamphlets; Digital Academic Edition; XML-TEI; Markup model; XVIth-XVIIIth Centuries

§

* Esta publicación es parte del proyecto de I+D+i Biblioteca Digital Siglo de Oro 6 (BiDISO 6), referencia: PID2019-105673GB-I00 financiado por MCIN/AEI/10.13039/501100011033. Sus autores se integran en el grupo de investigación HISPANIA (G000208) de la Universidade da Coruña.

1. Introducción

El proyecto *Biblioteca Digital Siglo de Oro* (BIDISO), a través de su portal web <https://www.bidiso.es> (cons. 13/05/2022, Fig. 1), ofrece acceso a nueve bases de datos (ocho de fuentes primarias impresas en la Edad Moderna y una que almacena referencias bibliográficas de estudios, ediciones y repertorios) sobre literatura emblemática, relaciones de sucesos, divisas o empresas históricas, polianteas e inventarios de bibliotecas y lecturas de los siglos XVI y XVII. En ellas, los usuarios, además de datos bibliográficos y la localización actual de ejemplares de obras de estos géneros, pueden encontrar imágenes o enlaces a ediciones facsimilares de utilidad para los interesados en la Literatura, la Historia, la Historia del libro y bibliotecas y la Historia del Arte de la época. Por ejemplo, el *Catálogo y biblioteca digital de relaciones de sucesos* (CBDRS) nos brinda datos sobre 6.259 ediciones de este tipo de documentos y la posibilidad de acceder a la digitalización facsimilar de 2.170 de ellas[1]. Con su labor, el proyecto BIDISO facilita el acceso a estas ediciones digitales codificadas en formato imagen a través de un punto de consulta único para las obras del género, permitiendo a los usuarios ahorrar mucho tiempo en búsquedas por distintos catálogos y eventuales gestiones o desplazamientos para observar, leer y obtener reproducciones de los ejemplares custodiados en diversas instituciones[2].

[1] Los datos procedentes de este catálogo en constante crecimiento se han consultado por última vez el 29/10/2021.

[2] Para informarse sobre la trayectoria del portal BIDISO –que tiene origen en el año 1993– y su evolución a lo largo del tiempo remitimos a Pena Sueiro (2017).

Fig. 1. Interfaz del portal BIDISO

La posibilidad de dar acceso a través de nuestro portal a ediciones facsimilares aumenta en paralelo a los progresos en la digitalización de las fuentes primarias y puesta a disposición en acceso abierto a través de Internet de los materiales procedentes de proyectos llevados a cabo, principalmente, por bibliotecas patrimoniales. Sin embargo, con el tipo de ediciones mayoritariamente disponibles, muchas de las necesidades actuales que se presentan en la investigación humanística no quedan

resueltas. La mayor parte de los impresos antiguos que encontramos en la red han sido codificados únicamente en formato imagen, con la adición de algunos metadatos, en su mayoría de carácter identificativo y descriptivo, quedando las posibilidades de búsqueda restringidas a estos y sin poder llevar a cabo análisis textuales apoyados en herramientas informáticas. Al proceder con la búsqueda de los documentos a través de nuestras bases de datos, las posibilidades de exploración aumentan, pues para completar los registros se tienen en cuenta categorías de análisis de interés para el género que nos ocupa; por ejemplo, el mote o el epigrama en el caso de las bases de datos dedicadas a literatura emblemática. Con todo, al no transcribirse los textos íntegramente, persisten limitaciones similares a las mencionadas.

Además de ediciones de textos de la Edad Moderna en formato imagen, también encontramos en Internet ediciones en las que los textos son procesables por ordenador. Sin embargo, por lo general, estos textos no reúnen las condiciones de fiabilidad y calidad suficientes para garantizar unos resultados adecuados para los estudios científicos. El rigor filológico es una característica con frecuencia ausente en las ediciones de textos que encontramos en Internet: en muchos casos no se comparte qué fuente o fuentes concretas se toman como base ni con qué criterios se ha establecido el texto; otras veces ni tan siquiera se menciona a los editores o a los responsables de esa transformación del impreso en texto digital (Peiró *et al.*, 2015, 344). Igualmente, cuando para obtener el texto se utilizan únicamente procesos automáticos basados en tecnología OCR, sin articular procesos de revisión en los que intervenga un agente humano para corregir errores, no se dispone de archivos que sirvan para análisis estadísticos y cuantitativos rigurosos, dado que la calidad de esos softwares con la tipografía de la época de la imprenta manual es limitada[3].

Así, con el afán de satisfacer la demanda de textos susceptibles de ser analizados sistemáticamente que los especialistas en los géneros de los que

[3] Este es el caso, por ejemplo, de los archivos ofrecidos en el ámbito de *Google Books*. Si bien este proyecto tiene el enorme mérito de mejorar la recuperación de una ingente cantidad de obras de la Edad Moderna a partir de palabras contenidas en el texto, la tasa de errores de los textos digitalizados propicia un carácter asistemático para el procesamiento automático de los datos. Suscribimos lo indicado sobre este proyecto en Mancinelli y Pierazzo (2020, 14-16).

se ocupa BIDISO precisan como base para sus estudios[4], hemos proyectado la creación de un repositorio de acceso libre y gratuito –*BIDISO Textos*–, para complementar nuestras herramientas y permitir, no solo la lectura y obtención de ejemplares en formato imagen *jpg*, sino también de diferentes versiones de las obras en las que el texto queda accesible para búsquedas y análisis automatizados en diversos formatos, con los que se favorece las posibilidades de reutilización (PDF, RTF, XML o TXT). Al crear este banco de textos, para garantizar la calidad de la investigación científica, consideramos imprescindible que en la transformación del impreso antiguo en edición digital se haya aplicado el rigor filológico, acudiendo a fuentes fiables y estableciendo el texto con criterios fundamentados de acuerdo con métodos y teorías ecdóticas que utilizamos con frecuencia en las ediciones que publicamos en papel. La creación de *BIDISO Textos* supone, además, una oportunidad para reflexionar sobre las nuevas funcionalidades que queremos brindar a nuestros usuarios aprovechando el enorme potencial de la digitalización, diferenciando nuestros productos finales de las ediciones cartáceas convencionales (Sahle, 2016, 19-39). De esta manera, *BIDISO Textos* pretende contribuir a la experimentación y reflexión sobre cómo ha de ser una edición académica digital, ofreciendo un posible modelo de referencia, que creemos que puede ser de utilidad considerando que en España, en comparación con otros países, todavía son pocas las iniciativas de este tipo que podemos encontrar en la red[5].

[4] Por ejemplo, Baena Sánchez *et al.* (2014) insisten en la necesidad de construir *corpora* marcados digitalmente y susceptibles de ser analizados de forma sistemática para avanzar en el estudio las primeras manifestaciones del periodismo europeo, entre las que interesan las relaciones de sucesos.

[5] Como afirma Allés Torrent (2017), «A diferencia de los modelos impresos, no existe todavía una idea clara de cómo debe ser una edición digital», lo que revela ausencia de modelos y prácticas consolidadas en la edición digital académica. González-Blanco (2017, 242) urge a llevar a la práctica más ediciones digitales en España, para romper la «brecha» que nos separa con otros países que presentan un panorama más rico de ediciones y la red. Como indica Allés Torrent (2017), existen dos grandes catálogos de obligada referencia y consulta internacional: el *Catalogue of Digital Scholarly Editions* de Patrick Sahle (Universität zu Köln) y el *Catalogue of Digital Editions* de Greta Franzini (GeorgAugust-Universität Göttingen). En ellos, apenas hay referencias a Literatura en español. En el primer catálogo, en su última actualización del 28/09/2020, sólo 14 de las 714 ediciones académicas digitales son de textos en español. De estas, sólo 3 fueron realizadas íntegramente en España: *Hypertexto del Orlando Furioso*, dirigido por María de las Nieves Muñiz Muñiz (Universitat de Barcelona, 2006); *CHARTA: Corpus hispánico y americano en la red: textos antiguos*, coordinado por Pedro Sánchez-Prieto Borja (Universidad de Alcalá de Henares), y *HESPERIA: Banco de datos de lenguas paleohispánicas*, dirigido por Javier de Hoz (Universidad

En este artículo, nos proponemos dar cuenta de los progresos en la conformación de *BIDISO Textos,* en concreto, en lo relativo a la colección de relaciones de sucesos, en la que se han producido avances trabajando en un modelo que recoge el conjunto de principios y prácticas que aplicaremos a los textos de este género. Comenzaremos delimitando el objeto de este proyecto de edición, nuestros destinatarios y objetivos prioritarios e identificando los materiales de partida disponibles. A continuación, haremos referencia a los procedimientos para la obtención de textos y luego centrarnos en los aspectos más destacables del modelo de codificación implementado. Para ejemplificar estos aspectos, recurriremos, siempre que sea posible, a la codificación de la que ha sido objeto la relación *Fiesta que hizo en Aranjuez a los años del rey nuestro señor don Felipe IIII*…, escrita por Antonio de Mendoza y editada por Garrobo Peral (2020, Fig. 2)[6].

Complutense de Madrid, 1997). Además, existe otro registro, *Lope de Vega - La Dama Boba – Edición crítica y archivo digital*, que dirigió Marco Presotto y surge de la colaboración entre el grupo PROLOPE (Universitat autónoma de Barcelona) y la Università di Bologna. En el caso del catálogo de Greta Franzini, en su última actualización de junio de 2021, filtrando por trabajos que sean considerados «scholarly», «editions» y «digital» aparecen 257 ediciones sobre un total de 316 registros; de estos, sólo 12 son textos en español, 6 de ellos ya presentes en el catálogo anterior. Entre estos, el único registro que menciona un proyecto realizado en España aún no mencionado es el *Cancionero digital de Gómez Manrique*, impulsado por la Real Biblioteca. En los últimos años, sabemos de otras iniciativas surgidas en España que realizan ediciones digitales que no aparecen en estos catálogos, quizá por no coincidir plenamente con el concepto de edición académica digital que aplican, por tratarse de proyectos en marcha o, tal vez, por haber pasado inadvertidos para sus recopiladores. Refiriéndonos al ámbito de la literatura del Siglo de Oro, este es el caso, por ejemplo, de las ediciones de paratextos de obras de autoras de BIESES, la Biblioteca Digital incluida dentro de ARTELOPE, las *Soledades* de Luis de Góngora de Antonio Rojas Castro o las iniciativas que prevén impulsar PROLOPE o PRONAPOLI (mencionadas en el encuentro de investigadores sobre Humanidades Digitales del Congreso de la AISO 2020 por Dolores Martos). Con todo, en la actualidad, la nómina de ediciones en España continúa siendo limitada.

[6] Garrobo Peral, que realizó una edición modernizada y anotada de este texto para su trabajo de fin de máster, toma como base la única edición conocida de la obra, publicada en Madrid por Juan de la Cuesta en 1623. El trabajo incluye, además, un capítulo introductorio al lenguaje XML-TEI y su aplicación en la relación de sucesos seleccionada.

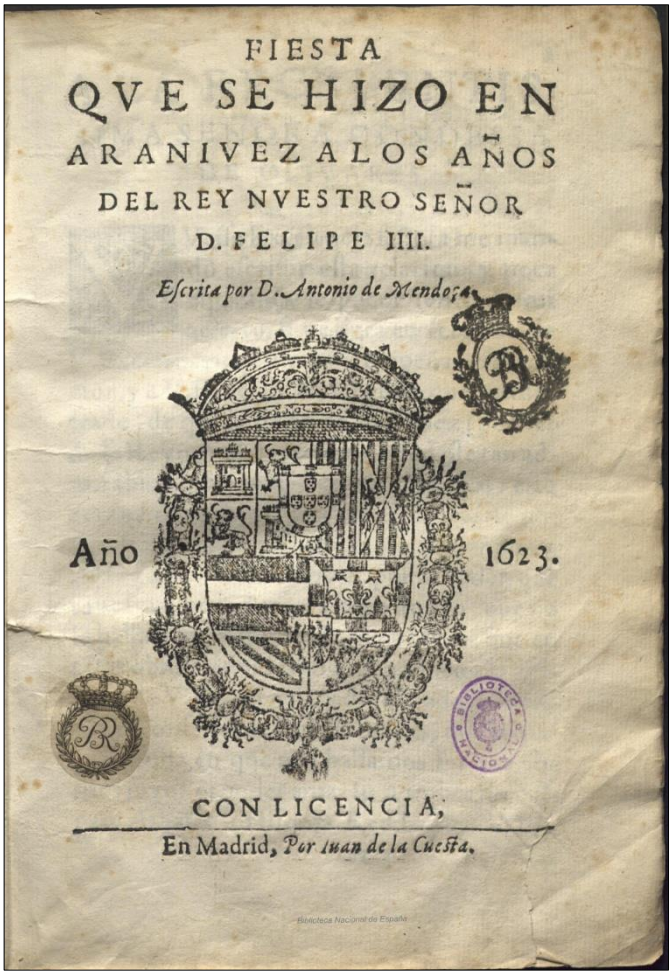


Fig. 2. Portada del ejemplar R/015515 de la Biblioteca Nacional de España (CBDRS: 0007066A)

2. Objeto del proyecto de edición, destinatarios, objetivos prioritarios e identificación de materiales disponibles

Las relaciones de sucesos son documentos publicados de manera no periódica a lo largo de la Edad Moderna para dar noticia –habitualmente desde una perspectiva subjetiva– de los más diversos tipos de eventos: guerras; autos de fe; recibimientos, bodas o exequias reales; canonizaciones, beatificaciones o procesiones; catástrofes naturales;

calamidades personales o viajes, entre otros. Se considera, centrándose en sus características tipográficas y materiales, que las relaciones «prototípicas» presentan una serie de rasgos distintivos que permiten hablar de un género editorial (Infantes, 1996; Pena Sueiro, 2001; Ruiz Astiz y Pena Sueiro, 2019). De acuerdo con esta concepción de las relaciones y en correspondencia con el trabajo realizado en CBDRS, limitamos nuestro corpus a los textos impresos de los siglos XVI a XVIII.

Además de servir a una comunidad investigadora que se interesa por las relaciones de sucesos desde diferentes perspectivas en correspondencia con una amplia variedad de áreas de estudio –como la Historia de las mentalidades o la del periodismo; la Antropología o la Bibliografía, entre otras–, tratamos de no dejar de lado al público general, pues las obras del género, que reflejan múltiples aspectos de la cultura de la Edad Moderna europea, pueden suscitar el interés de muchos curiosos. Así pues, entre nuestros usuarios, puede haber diferentes grados de habilidades lectoras al enfrentarse a textos antiguos, por lo que consideramos que, cuando sea factible, debemos acercarlos al lector contemporáneo ofreciendo grafías modernizadas y notas que aclaren aspectos contextuales.

La diversidad de áreas de estudio desde las que se abordan las relaciones de sucesos redunda en la variedad y riqueza de puntos de vista de los trabajos producidos en torno al género, pero establecer un etiquetado en correspondencia con todas las posibles categorías de análisis de interés conllevaría que el corpus creciese a un ritmo muy lento. Ante esta situación y con el deseo de realizar una contribución que pueda servir de manera global a todos los que investigan sobre el género, hemos decidido conceder prioridad (al menos en una primera fase) a dos propósitos fundamentales: difundir el mayor número de textos posible garantizando su fiabilidad y potenciar las posibilidades de recuperación de los mismos considerando su contenido y las relaciones intertextuales que se establecen entre las diversas obras y ediciones que incluya la colección[7].

La larga trayectoria en el estudio de las relaciones de sucesos de BIDISO supone una oportunidad para hacer crecer esta nueva colección.

[7] Algunas de las posibilidades que Fernández Travieso (2013) ilustraba, quedan ahora circunscritas a los objetivos de esta fase, derivando en una simplificación el marcado. Esos planteamientos serán susceptibles de desarrollo en fases posteriores del proyecto.

En CBDRS, hemos reunido información sobre la disponibilidad en la red de reproducciones facsimilares digitales de impresos producidos durante los siglos XVI a XVIII. Algunas de las instituciones que producen estas digitalizaciones atribuyen a sus imágenes licencias que permiten su uso. Este es el caso, por ejemplo, de la Biblioteca Nacional de España, de la que en CBDRS hay registradas 921 ediciones de relaciones[8]. También, por ejemplo, de la Biblioteca Universitaria de Sevilla, de la que en CBDRS hay 446 ediciones[9]. En otros casos, pueden establecerse acuerdos puntuales o convenios para conseguir reproducciones y/o el permiso para la producción de obras derivadas y la redistribución de las imágenes.

Por otra parte, con los debidos permisos de los autores, el corpus puede incluir un buen número de ediciones de relaciones que fueron tesinas de licenciatura, trabajos de fin de grado o máster y contribuciones de los miembros del equipo a volúmenes colectivos, como los que se realizaron para el volumen *Malas noticias, noticias falsas. Relaciones de sucesos en los siglos XVI y XVII* (2019), editado por Nider y Pena Sueiro, o el recién publicado *Buenas noticias. Relaciones de sucesos en los siglos XVI y XVII* (2021), editado por Andrés Renales y Peñasco González[10]. Además, se puede llegar a acuerdos con otros estudiosos ajenos al SIELAE que hayan producido ediciones de acuerdo con criterios filológicos rigurosos.

Asimismo, no queremos dejar de lado la posibilidad de recurrir a ediciones de relaciones contemporáneas que, por su fecha de publicación, hayan pasado a ser de dominio público. A modo de ejemplo, podríamos mencionar la *Relación verdadera del rebato que dieron cuatrocientos y cincuenta turcos en la almadraba de Zahara...*, editada por Amalio Huarte Echenique, bibliotecario en la Biblioteca Nacional de España (1941), de la que podemos también disponer de imágenes facsimilares de la edición de 1562 (BNE R/11907(6)). En este caso, al ofrecérsenos un texto con modernización de grafías, puntuación, acentuación..., una de las versiones

[8] Sobre los permisos de uso que otorga la Biblioteca Nacional de España, véase: <http://www.bne.es/es/Servicios/ReproduccionDocumentos/UsoReproducciones> (cons. 15/05/2022).

[9] Esta biblioteca ofrece imágenes de sus obras de su fondo antiguo a través del *Internet Archive*. Como se puede observar, por ejemplo, accediendo al enlace <https://archive.org/details/A109085079> (cons. 15/05/2022). las imágenes de esta relación cuentan con una licencia Creative Commons Public Domain 1.0.

[10] Esta tarea se acometerá una vez decaídos los eventuales derechos de exclusividad de editores.

de la obra que podemos incluir tomaría como base esta publicación, pues con ello se facilita la lectura a los receptores actuales, a ese público más amplio al que queremos llegar.

Si bien contar con todos estos materiales supone una ventaja a la hora de acrecentar la colección de manera rápida, topamos también dificultades. En nuestro modelo debemos dar cabida a diferentes procedimientos para llevar a cabo la transformación del texto analógico en digital dependiendo del material de partida, garantizar que en cada caso los usuarios tengan conocimiento de cómo se ha llevado esta transformación y proporcionar una homogeneidad suficiente a la colección, de manera que sea posible su visualización y su explotación como conjunto. Al establecer los mecanismos para la obtención de los textos y el modelo de codificación, hemos tenido en cuenta la necesidad de hacer frente a estas dificultades.

3. La obtención de textos

Tras comprobar la disponibilidad de imágenes de los impresos antiguos, es necesario convertir aquellos documentos, que no lo estén ya, a un formato digital con texto accesible. Para las ediciones realizadas recientemente, en su mayoría, disponemos de archivos digitales en los que la transformación del impreso antiguo se ha acometido de manera manual. Para textos con tipografía actual que únicamente disponemos en papel, contemplamos el uso de software OCR. Con todo, la mayor parte de las relaciones conocidas no ha sido objeto de atención por parte de editores contemporáneos, por lo que también consideramos conveniente explorar las posibilidades del uso de tecnologías HTR para nuestros fines. Así, surge la colaboración de BIDISO con el Progetto Mambrino y COMEDIC para lanzar un proyecto de equipo de transcripción automatizada de impresos de la Edad Moderna, del que tratan Stefano Bazzaco *et al.* (2022), Blasut (2022) y Aranda García (2022) en este mismo monográfico.

4. La codificación

A continuación, tratamos los textos obtenidos con lenguaje de marcación, en concreto, con el formato de datos XML-TEI. En nuestro caso, la principal razón para adoptar este estándar es que permite poner de relieve todas las características de las relaciones de sucesos que necesitamos marcar para dar cumplimiento a los objetivos de nuestro proyecto, registrándolas en un archivo XML, cuya visualización puede realizarse en diversos entornos digitales sin conllevar pérdidas de información. En segundo lugar, por sus eminentes ventajas a la hora de alimentar la preservación del trabajo, la reutilización o migración de los datos y la interoperabilidad con otros proyectos. En otras palabras, empleamos el esquema TEI como una suerte de formato de *input-output*: como formato de entrada, es una modalidad cómoda para modelizar nuestros textos y, como formato de salida, es un lenguaje muy difundido y bien conocido por la comunidad científica. Con todo, la adopción del etiquetado XML-TEI no condiciona la modelización (son las características textuales las que prevalecen) ni la futura re-producción de los textos en la red, para la que podremos estudiar diversas opciones, pudiendo, eventualmente, utilizarse otros formatos en el futuro.

En nuestro modelo, distinguimos grupos de etiquetas de aplicación general para todas las relaciones de sucesos que editemos y módulos que utilizamos solo con algunos tipos de materiales de partida concretos, por ejemplo, aquellas ediciones en las que se utiliza más de un testimonio para el establecimiento del texto.

4.1. Etiquetas de aplicación general

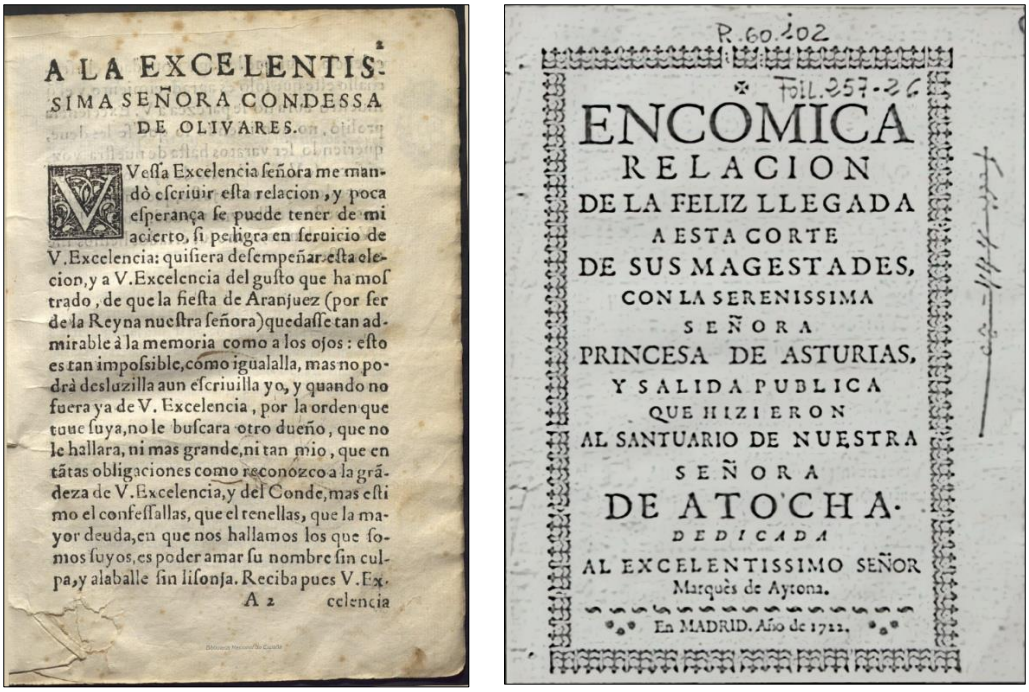
4.1.1. El elemento <text>

Desde un punto de vista estructural, la sistematización de las características del género no resulta sencilla, pues, como explican Ruiz Astiz y Pena Sueiro (2019), en las relaciones de sucesos, a menudo, se producen variaciones en la distribución textual, extensión, presencia o no

de paratextos legales o literarios etc. Esta variabilidad del género constituye uno de los retos que debemos afrontar a la hora de establecer un modelo de codificación, obligándonos a pensar en posibles formas de etiquetar elementos que eventualmente pueden aparecer en los textos que identificamos como pertenecientes al género.

Según las directrices de la TEI, el texto del documento editado, el <text>, se suele estructurar en tres subsecciones: <front>, que contiene los materiales paratextuales preliminares; <body>, que cuenta con el texto propiamente dicho y es la única subsección obligatoria de las mencionadas, y <back>, que abarca todo tipo de apéndice que sigue al cuerpo del texto.

Si ponemos la atención en el <front>, es preciso considerar que algunas relaciones tienen portada y la narración de los eventos comienza en otra página; pero, en otras, la narración comienza ya en la primera página, precedidos de una portadilla o solo de un encabezamiento. A la portada, pueden seguirle otros textos preliminares (aprobación, licencia, privilegio, dedicatoria, prólogo, poemas laudatorios…), en especial, si estamos ante relaciones que adquieren forma de libro, pues –aunque la mayoría de las relaciones son textos breves y se publican como pliegos sueltos– algunas pueden cobrar mayor extensión y complejidad.



A LA EXCELENTIS-
SIMA SEÑORA CONDESSA
DE OLIVARES.

VEssa Excelencia señora me mandò escriuir esta relacion, y poca esperança se puede tener de mi acierto, si peligra en seruicio de V. Excelencia: quisiera desempeñar esta eleccion, y a V. Excelencia del gusto que ha mostrado, de que la fiesta de Aranjuez (por ser de la Reyna nuestra señora) quedasse tan admirable à la memoria como a los ojos: esto es tan imposible, cómo igualalla, mas no podrà desluzilla aun escriuilla yo, y quando no fuera ya de V. Excelencia, por la orden que tuue suya, no le buscara otro dueño, que no le hallara, ni mas grande, ni tan mio, que en tãtas obligaciones como reconozco a la grãdeza de V. Excelencia, y del Conde, mas estimo el confessallas, que el tenellas, que la mayor deuda, en que nos hallamos los que somos suyos, es poder amar su nombre sin culpa, y alaballe sin lisonja. Reciba pues V. Ex-

A 2 celencia

R.60.102

Fol.257-26

ENCOMICA
RELACION
DE LA FELIZ LLEGADA
A ESTA CORTE
DE SUS MAGESTADES,
CON LA SERENISSIMA
SEÑORA
PRINCESA DE ASTURIAS,
Y SALIDA PUBLICA
QUE HIZIERON
AL SANTUARIO DE NUESTRA
SEÑORA
DE ATOCHA.
DEDICADA
AL EXCELENTISSIMO SEÑOR
Marquès de Aytona.

En MADRID. Año de 1722.

Fig. 3. Ejemplo de dedicatoria en página independiente (Antonio de Mendoza, *Fiesta que se hizo en Aranjuez*…, Madrid, 1623) y de dedicatoria inserta en el título de portada (*Encómica relación*…, Madrid, 1722, custodiada en la Biblioteca de la Universidad de Santiago de Compostela <http://hdl.handle.net/10347/6935> CBDRS: 0002334A)

En el caso de los impresos menores, es frecuente que los datos identificativos, dedicatorias o información sobre el cumplimiento de requisitos legales de la publicación se condensen en la portada o portadilla entremezclándose en el título –que suele ser largo– o en pie de imprenta[11]. Así, la mayoría de las relaciones tienen en el <front>, únicamente, un elemento <titlePage>, que aplicamos con cierta flexibilidad, tanto a portadas que ocupan la totalidad de la página, como a portadillas y enunciados. Sin embargo, también pueden aparecer otras secciones dentro de <front> identificadas con la etiqueta e indicando a través del valor atributo *type* si se trata de la aprobación, la tasa, la dedicatoria, el

[11] Una pragmática de 1627, pretendiendo evitar los fraudes y la publicación de muchos textos subversivos, obligaba a que los impresos menores publicados en la Monarquía Hispánica contasen con aprobación de los consejos territoriales e indicasen fecha, lugar de impresor, nombre el impresor y del autor (Reyes Gómez 1999, 328). Muchas veces, las relaciones de sucesos eludían estas leyes (Martín Molares, 2017). Sobre las características de los títulos de este tipo de textos véase Pena Sueiro (1999).

prólogo, etc.

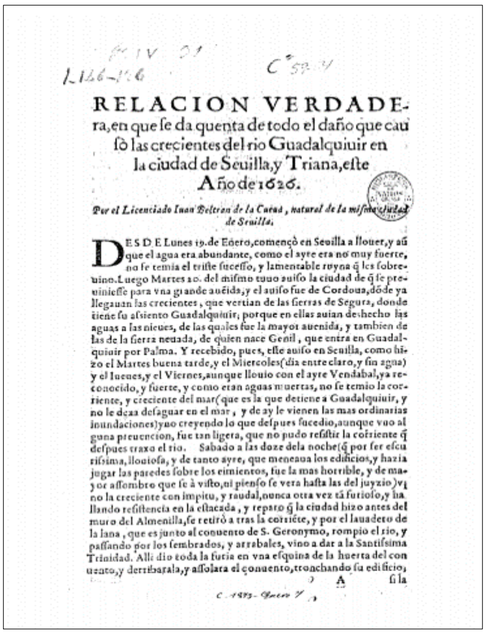
Dentro de <titlePage>, el elemento <docTitle>, que según el esquema TEI recoge el título, abarca de manera global informaciones que, en el caso de las relaciones, con frecuencia, aparecen tipográfica y gramaticalmente unidas a este, como el nombre del autor, la fecha de impresión, la dedicatoria… Estos datos se etiquetan con detalle en otro momento, al describir la fuente en los metadatos del <teiHeader>[12]:

```
<titlePage>
 <docTitle>
  <titlePart type="main">Fiesta que se hizo en Aranjuez a los años del rey
 nuestro señor Felipe IIII.</titlePart>
  <titlePart type="sub">Escrita por D. Antonio de Mendoza.</titlePart>
 </docTitle>
 <docDate>Año<figure><graphic
 url="0007066A_il/0007066A_il001"/></figure> 1623.</docDate>
 <docImprint>Con licencia. En Madrid. Por Juan de la Cuesta
 </docImprint>
</titlePage>
```

Centrándonos en el <body>, debemos tener en cuenta que las relaciones, a diferencia de otros géneros informativos de la Edad Moderna, como avisos o gacetas, suelen referir un único acontecimiento en vez de reportar varios, pero en algunas ocasiones nos encontramos más de una noticia en la misma publicación, con su título propio, como vemos, por ejemplo, en el caso de la figura 4. Para ello, hemos previsto que, opcionalmente, el cuerpo de las relaciones pueda estar compuesto de uno o más , etiqueta con la que marcaremos cada noticia.

Además, el cuerpo de la obra puede presentarse en prosa (normalmente a renglón tirado), en verso (y suele entonces presentarse en dos columnas) o combinando prosa y verso. Así, el modelo prevé y explica que se usarán etiquetas para marcar párrafos <p> o estrofas <lg> y versos <l>, según convenga (Fig. 5), y distinguir columnas, con <cb>.

[12] Obsérvese que el nombre del autor, Antonio de Mendoza, queda englobado en la etiqueta <docTitle>, en concreto, en <titlePart type="sub">.



RELACION VERDADE-
ra, en que se da quenta de todo el daño que cau
sò las crecientes del rio Guadalquiuir en
la ciudad de Seuilla, y Triana, este
Año de 1626.

Por el Licenciado Iuan Beltran de la Cueua, natural de la misma ciudad de Seuilla.

DESDE Lunes 19. de Enero, començò en Seuilla a llouer, y aũque el agua era abundante, como el ayre era no muy fuerte, no se temia el triste sucesso, y lamentable ruyna q̃ les sobreuino. Luego Martes 20. del mismo tuuo auiso la ciudad de q̃ se preuiniesse para vna grande auẽida, y el auiso fue de Cordoua, dõde ya llegauan las crecientes, que vertian de las sierras de Segura, donde tiene su asiento Guadalquiuir; porque en ellas auian deshecho las aguas a las nieues, de las quales fue la mayor auenida, y tambien de las de la sierra neuada, de quien nace Genil, que entra en Guadalquiuir por Palma. Y recebido, pues, este auiso en Seuilla, como hizo el Martes buena tarde, y el Miercoles (dia entre claro, y sin agua) y el Iueues, y el Viernes, aunque llouio con el ayre Vendabal, ya reconocido, y fuerte, y como eran aguas muertas, no se temio la corriente, y creciente del mar (que es la que detiene a Guadalquiuir, y no le dexa desaguar en el mar, y de ay le vienen las mas ordinarias inundaciones) y no creyendo lo que despues sucedio, aunque vuo alguna preuencion, fue tan ligera, que no pudo resistir la corriente q̃ despues traxo el rio. Sabado a las doze dela noche (q̃ por ser escurissima, llouiosa, y de tanto ayre, que meneaua los edificios, y hazia jugar las paredes sobre los cimientos, fue la mas horrible, y de mayor assombro que se à visto, ni pienso se vera hasta las del juyzio) vi no la creciente con impitu, y raudal, nunca otra vez tã furioso, y hallando resistencia en la estacada, y reparo q̃ la ciudad hizo antes del muro del Almenilla, se retirò a tras la corriẽte, y por el lauadero de la lana, que es junto al conuento de S. Geronymo, rompio el rio, y passando por los sembrados, y arrabales, vino a dar a la Santissima Trinidad. Alli dio toda la furia en vna esquina de la huerta del conuento, y derribarala, y assolara el conuento, tronchando su edificio,

A si la

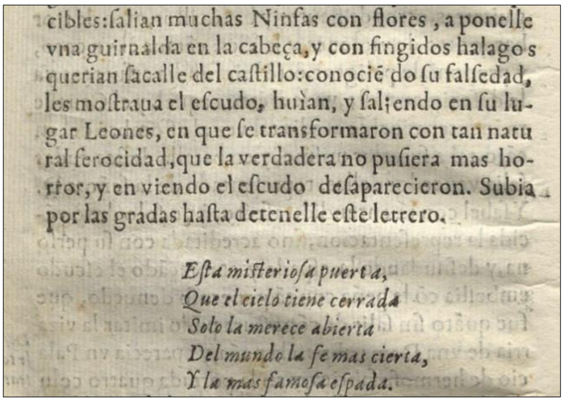
noche, el arroyo Tagarete ya tan crecido, q̃ se ahogaron de vna casa dos mugeres, y tres niños. A 10. del dicho, a las dos de la maña, rompio el agua por el lauadero dela lana, con tal furia, q̃ llegò junto a la puerta de Macarena, haziendo grãdissimo estrago en muchas casas, q̃ auian quedado de las demas auenidas despojadas, y se vá cayendo. Y assi para aplacar estos daños, todos los Conuẽtos, y Parroquias hazen vn nouenario en la Iglesia Mayor, al Santo Christo sin lançada, auiendoles señalado hora a todas las Religiones, y Cofradias, estãdo el [illegible] descubierto [illegible] Dios nos tenga de su mano, y nos de lo que nos conuiene para saluacion nuestra.

SEGVNDA Y BREVE RELACION DE LA GRAN auenida, y daños que causò el rio Tormes en la ciudad de Salamanca.

LVnes en la tarde, que se contaron 26. de Enero, deste presente año de 1626. auiendo precedido Sabado, y Domingo, grandes aguas, y furiosos vientos, tendio la noche su manto negro, como requeria para las exequias de la [illegible] [illegible]

Impresso en Lima; Por Geronymo de Contreras; Año de 1626.

Fig. 4. Portada y cuarta página de la *Relacion verdadera en que se da quenta de todo el daño que causò las crecientes del rio Guadalquiuir en la ciudad de Seuilla y Triana, este año de 1626* de Juan Beltrán de la Cueva (Sevilla, 1626), custodiada en la Biblioteca Nacional de España (CBDRS: 0007022A). Se observa que en la última página del pliego se incluye otra noticia, referente a la ciudad de Salamanca.



cibles: salian muchas Ninfas con flores, a ponelle vna guirnalda en la cabeça, y con fingidos halagos querian sacalle del castillo: conociẽdo su falsedad, les mostraua el escudo, huìan, y saliendo en su lugar Leones, en que se transformaron con tan natural ferocidad, que la verdadera no pusiera mas horror, y en viendo el escudo desaparecieron. Subia por las gradas hasta detenelle este letrero.

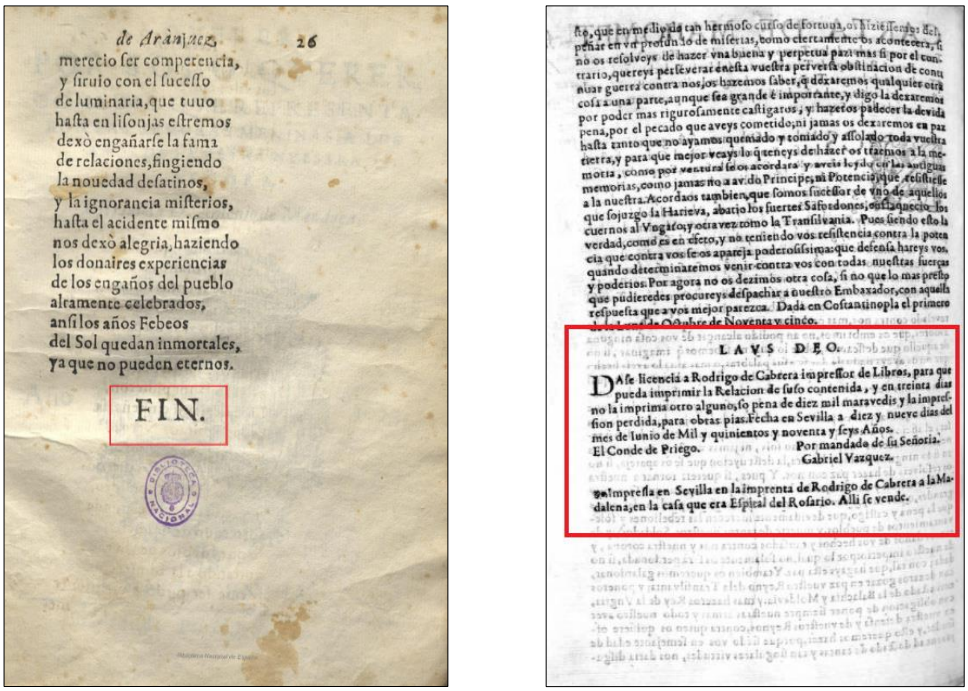
Esta misteriosa puerta,
Que el cielo tiene cerrada
Solo la merece abierta
Del mundo la fe mas cierta,
Y la mas famosa espada.

<p> [...] Salían muchas ninfas con flores a ponelle una guirnalda en la cabeza y con fingidos halagos querían sacalle del castillo. Conociendo su falsedad, les mostraba su escudo, huían y saliendo en su lugar leones en que se transformaron con tan natural ferocidad que la verdadera no pusiera más horror y, en viendo el escudo, desaparecieron. Subía por las gradas hasta detenelle este letrero:</p>

<lg>
<l>Esta misteriosa puerta,</l>
<l>que el cielo tiene cerrada</l>
<l>solo la merece abierta</l>
<l>del mundo la fe más cierta</l>
<l>y la más famosa espada.</l>
</lg>

Fig. 5. Página de *Fiesta que se hizo en Aranjuez…* de Antonio de Mendoza (Madrid, 1623) donde se alternan verso y prosa (fol. 10v)

En cuanto al <back>, algunas relaciones pueden presentar colofón y otras ir sin él. También, ocasionalmente, los datos en relación con la impresión, licencia o privilegio pueden aparecer al final. A veces, pueden aparecer otras informaciones como, por ejemplo, la licencia o avisos comerciales que, como en el caso de la figura 6, anuncian una continuación. Nuestro modelo propone, pues, el uso de ciertas etiquetas de manera opcional (como <trailer>, <docImprint>, y el propio <back>), para permitir que el codificador las use si es necesario.



de Aranjuez. 16

merecio ser comperencia,
y siruio con el suceſſo
de luminaria, que tuuo
hasta en liſonjas eſtremos
dexò engañarſe la fama
de relaciones, fingiendo
la nouedad deſatinos,
y la ignorancia miſterios,
hasta el acidente miſmo
nos dexò alegria, haziendo
los donaires experiencias
de los engaños del pueblo
altamente celebrados,
anſi los años Febeos
del Sol quedan inmortales,
ya que no pueden eternos.

FIN.

fio, que en medio vida tan hermoſo curſo de fortuna, os hizieſſe tener ſed,
penar en vn profundo de miſerias, como ciertamente os acontecera, ſi
no os reſolveys de hazer vna buena y perpetua paz: mas ſi por el con-
trario, quereys perſeverar en eſta vueſtra perverſa obſtinacion de conti-
nuar guerra contra nos, os hazemos ſaber, q dexaremos qualquier otra
coſa a una parte, aunque ſea grande e importante, y digo la dexaremos
por poder mas rigurosamente caſtigaros, y haze-os padecer la devida
pena, por el pecado que aveys cometido, ni jamas os dexaremos en paz
hasta tanto que no ayamos quemado y tomado y aſſolado toda vueſtra
tierra, y para que mejor veays lo que ueys de hazer os traemos a la me-
moria, como por ventura no os acordara y aveis leydo en las antigua
memorias, como jamas no a avido Principe, ni Potencia que reſiſtieſſe
a la nueſtra. Acordaos tambien, que somos ſucceſſor de vno de aquellos
que ſojuzgo la Harieva, abatio los fuertes Saſones, ocupando los
cuernos al Vngaro, y otra vez tomo la Tranſilvania. Pues ſiendo eſto la
verdad, como es en efecto, y no teniendo vos reſiſtencia contra la poten
cia que contra vos ſe os apareja poderoſiſsima que defenſa hareys vos,
quando determinaremos venir contra vos con todas nueſtras fuerças
y poderios. Por agora no os dezimos otra coſa, ſi no que lo mas preſto
que pudieredes procureys deſpachar a nueſtro Embaxador, con aquella
reſpuesta que a vos mejor parezca. Dada en Conſtantinopla el primero
de los [illegible] de Octubre de Noventa y cinco.

LAVS DEO.

DAſe licencia a Rodrigo de Cabrera impreſſor de Libros, para que
pueda imprimir la Relacion de ſuſo contenida, y en treinta dias
no la imprima otro alguno, ſo pena de diez mil maravedis y la impreſ-
ſion perdida, para obras pias. Fecha en Sevilla a diez y nueve dias del
mes de Iunio de Mil y quinientos y noventa y ſeys Años.
El Conde de Priego.
Por mandado de ſu Señoria.
Gabriel Vazquez.

Impreſſa en Sevilla en la imprenta de Rodrigo de Cabrera a la Ma-
dalena, en la caſa que era Eſpital del Roſario. Alli ſe vende.

Fig. 6. Ejemplo de un <back> incluyendo tan solo un colofón (*Fiesta que se hizo en Aranjuez*…, Madrid, 1623) y de otro con más información (*Carta de Mahomet*…, Sevilla, 1596, custodiada en la Biblioteca de la Universidad de Sevilla, CBDRS: 0004750A)

Otro aspecto variable, como hemos visto ya en los ejemplos, es que algunas relaciones incluyen ilustraciones, mientras que en otras aparece solo texto. El modelo prevé también esta circunstancia, proponiendo etiquetas opcionales para incluirlas.

4.1.2. El elemento <teiHeader>

En cuanto a los metadatos, trataremos de algunos aspectos clave relacionados con nuestros objetivos y las dificultades que hemos identificado previamente.

Para dejar constancia de cómo se ha llevado el proceso de transformación del impreso antiguo a la edición digital, en la sección de <encodingDesc> de nuestro <teiHeader>, tal y como determina el esquema TEI, se documenta la relación entre el documento electrónico y la fuente o fuentes de la que procede. En ella, en <appInfo>, dejamos constancia de los programas utilizados: Transkribus, ABBYY Fine Reader, Oxygen XML... En <editorialDecl>, incluimos los principios de editoriales utilizados, que pueden variar de unas relaciones a otras. En el ámbito de BIDISO, recomendamos el uso de criterios modernizadores; pero, al trabajar con Transkribus, producimos versiones de carácter más conservador; si se trata de ediciones realizadas fuera del ámbito del proyecto, consideramos los criterios que se proporcionen en cada caso.

Por otra parte, para potenciar las posibilidades de recuperación de los textos, se han introducido en el modelo medios que permitan en un futuro:

1) *Incorporar la posibilidad de buscar las relaciones por acontecimientos históricos*. La TEI prevé que este elemento, tenga una subsección que condense la información clasificatoria y contextual del texto: <profileDesc>. Dentro de esta subsección, podríamos encontrar, por ejemplo, los datos clasificatorios expresados de este modo:

```
<textClass>
        <keywords scheme="tematica">
                <term>Relaciones de ceremonias y festejos</term>
                <term>Fiestas monárquicas</term>
        </keywords>
        <keywords>
                <term type= "acontecimiento">Aniversario de Felipe IV</term>
        </keywords>
</textClass >
```

Los términos marcados con la etiqueta <term type= "acontecimiento"> ("Aniversario de Felipe IV", en el ejemplo que proponemos) complementan la taxonomía por géneros y subgéneros temáticos implementada en CBDRS que, como vemos más arriba, reflejamos también en el archivo XML-TEI («Relaciones de ceremonias y festejos» y «Fiestas monárquicas»). De esta manera, se propicia que los usuarios puedan recuperar los textos disponibles en nuestro corpus sobre un mismo evento. «Intercambio de princesas de 1615», «Terremoto de Lisboa de 1755», «Guerra de los treinta años», etc., son ejemplos de otros acontecimientos históricos que podrían dar lugar a interesantes agrupaciones de los documentos.

2) *Localizar en el texto a las personas que participan en el suceso*. Utilizamos la etiqueta <rs type="person"> para marcar en el cuerpo del texto las alusiones a las personas que intervienen en el acontecimiento. Según se explicita en el esquema TEI, esta etiqueta puede recoger un nombre de propósito general o una cadena de referencia; con ella, no solo marcaremos nombres propios, sino también a personas que puedan ser identificadas con un nombre común como un niño, una mujer, un bandolero, un soldado, el rey nuestro señor..., aun cuando no podamos identificarlos. También hemos seleccionado etiquetas para cuando sea posible identificar a las personas aludidas y normalizar sus nombres. Este podría ser un ejemplo de cómo se lleva a cabo:

```
<rs type="person">
  <choice>
        <orig> el rey nuestro señor</orig>
        <reg> <persName>Felipe III, Rey de España (1598- 1621)</persName>
        </reg>
  </choice>
</rs>[13]
```

Dentro de la etiqueta <rs>, puede incluirse, como en este ejemplo, la etiqueta <choice> para introducir dos alternativas posibles para un

[13] En la fase actual, la identificación y normalización la introducimos cuando utilizamos como materiales de partida ediciones críticas que facilitan esta información o cuando la asignación de identidad se puede comprobar de manera inmediata.

pasaje del texto: «el rey nuestro señor», marcada con<orig>, que es la lectura del original y la identificación y normalización de ese nombre de persona, marcado con <reg> y <persName>.

El conjunto de términos marcados como acontecimientos históricos y de nombres de persona que intervienen en la acción marcados en los textos pueden ser el germen de futuros tesauros que aumenten las posibilidades de búsqueda por parámetros no contemplados en CBDRS.

3) *Navegar por una secuencia de textos de relaciones de sucesos con diferentes tipos de vínculos bibliográficos*, como la que se establece entre diferentes ediciones que forman parte de una misma relación publicadas por partes o de manera seriada o, en caso de que puedan establecerse vínculos con otros investigadores internacionales que editasen relaciones, las distintas traducciones posibles de una misma relación.

PRIMERA PARTE
DE LA RELACION DE LAS
Reales disposiciones, y Magestuosos aparatos,
con que su Magestad (Dios le guarde) se ha ser-
uido hazer jornada a la Provincia de Guipuz-
coa, a entregar a la Serenissima señora Doña
Maria Teresa Bibiana de Austria, su hija,
al Christianissimo Luis Dezimo-
quarto de Francia, su Esposo.
Año de 1660.
CON LICENCIA.
En Sevilla, por Juan Gomez de Blas, Impressor mayor de dicha Ciudad. Año de 1660.

SEGUNDA PARTE
DE LA RELACION DIARIA
DEL ITINERARIO, QVE SV MAGESTAD
(que Dios guarde) ha seguido desde q̃ salió de Madrid,
hasta llegar a Fuenterabia y forma de los Desposorios,
que en aquella Ciudad se celebraron, entre la Sereniſ-
sima señora D. Maria Teresa, Infanta de España, y el
Christianissimo Luis Decimoquarto de Francia, y en su
nombre, y en virtud de su poder, el Excelentissimo
Don Luis Mendez de Haro y Guzman.
Año de 1660.
DESPVES que su Magestad salio de Ma-
drid para la Ciudad de Fuenterabia (que
fue Jueves 15. de Abril proximo passado)
llegó aquella noche a la Villa de Alcalá de
Henares, donde se venera el milagroso
cuerpo de San Diego, Religioso de la Or-
den N. Serafico Padre S. Francisco, con quien tienen particu-
lar deuocion los señores Reyes de España. Fue grande el re-
cebimiento, y fiestas que hizo aquella Villa a su Magestad, el
poco tiempo que gozo de su Real presencia.
Viernes 16. llegó a la Ciudad de Guadalaxara, cuyo Cabildo
manifestó con varias muestras de regozijo, el afecto que, como
tan leal vassallos, tienen a su legitimo Rey, y Señor.
Sabado 17. a la Villa de Hita.
Domingo 18. a Xadraque, Villa del Duque del Infantado,
y aunque ella, y la antecedente son cortas de vecindad, igua-
laron en afectos, y demostraciones sus fuerças, a las Ciudades
mayores y populosas de este Itinerario.
Lunes 19. a Arena, cuya Villa alegró con su luzimiento, entre
otros regocijos, unos fuegos artificiales.
B Mar.

TERCERA PARTE
DE LA RELACION DE LAS VISTAS
DE LOS DOS PODEROSOS MONARCAS
de la Europa, y entregas de la señora Reyna de Francia,
Presentes que le han hecho sus Magestades,
y llegada del Rey N. señor a Valladolid,
Año de 1660.
DESPVES que se celebraron los Desposorios en
la Iglesia Mayor de la Ciudad de Fuenterrabia
(como se dio noticia en la Relacion antece-
dente) se retiraron sus Magestades a Palacio,
y como la Christianissima en publico, assistien-
dola a la mesa muchos Monsiures, y Madamas
de quenta, y en particular la Duquesa de Montpensier, hija del
Duque de Orleans, la qual besó la mano a la Reyna, y su Magesta
dChristianissima la echó los braços al cuello, y la acarició, y re-
galó, mandandola sentar a su lado, y tratandola como a su prima.
Este dia hizo en su alojamiento el Excelentissimo D. Luis Men
dez de Haro, vn suntuoso banquete a los Illustrissimos Patriarca
de las Indias, y Obispo de Pamplona, y Breua, y a los Excelen-
tissimos Duques de Medina de las Torres, Terranova, y Veras-
gua: al Conde de Fuensaldaña, y a otros muchos Titulos, y Ca-
valleros, assi Españoles, como Franceses.
Dicho dia llegó a Fuenterrabia el Duque de Crequi, primero
Ministro de la Camera del Rey Christianissimo, acompañado de
30. Titulos, y Monsiures de quenta, con lucidissimas galas al vso
de Francia, adornadas de tan costosas, como vistosas Joyas de Pe-
rlas, y diamantes. Traxo dos Cartas del Rey Christianissimo
vna para el Rey nuestro señor, y otra para la señora R
Francia, su Esposa.

Fig. 7. Ejemplo de relación seriada que narra cómo se procedió a la entrega de María Teresa de Austria para su matrimonio con el rey Sol. Las imágenes proceden de la Biblioteca de la Universidad de Sevilla. Sus códigos en CBDRS son, respectivamente, 0003549A, 0002944A y 0003557B

En la definición de nuestro género, se menciona con frecuencia su carácter ocasional; sin embargo (aunque no se llega a establecer una periodicidad fija como la de las gacetas), encontramos también casos de continuaciones y relaciones seriadas.

CBDRS, yendo más allá de un OPAC bibliotecario convencional (que registra ediciones y sus ejemplares), vincula entre sí las ediciones del mismo texto, permitiéndonos ver qué ediciones hay en el catálogo además de la que estamos considerando. Sin embargo, en nuestro catálogo no queda traza de la conexión entre pliegos publicados de manera independiente, –por partes– que integran una misma relación, es decir, no tiene presente la existencia de documentos multiparte.

Así, al describir la fuente de la edición en los metadatos (en la etiqueta <souceDesc>), hemos pensado en cómo ampliar la posibilidad de expresar relaciones bibliográficas de manera que, en una fase futura, pueda diseñarse una aplicación web que permita a los usuarios navegar fácilmente por el conjunto de pliegos que hacen parte de una misma obra. Para ello, nos hemos inspirado en una propuesta realizada por Rojas Castro en el ámbito del proyecto *HallerNet* para describir las relaciones bibliográficas, que se basa en vínculos entre las entidades del modelo *Requisitos funcionales de registros bibliográficos* (FRBR) de la IFLA (1999) y el etiquetado TEI[14]. Gracias al elemento <relatedItem> podemos explicitar que dos «manifestaciones» diferentes (que es como denominaríamos en FRBR a cada pliego publicado por separado) son parte de la misma «expresión» (u obra global)[15].

[14] Sobre este proyecto nos habló Rojas Castro en el Congreso de la Asociación de Humanidades Digitales Hispánicas de 2019 en su intervención «Llevando TEI al límite: sobre el modelado de la plataforma HallerNet». A modo de propuesta todavía por aplicar, se ha compartido la presentación «Modeling FRBR entities and their relationships with TEI. A look at HallerNet bibliographic descriptions», accesible en <https://tinyurl.com/y2ggcug9> (cons. 13/05/2022). En Francisco Baena *et al.* (2014), se anunciaban ya opciones para marcar la serialidad a través del etiquetado TEI. El sistema que proponemos en el actual modelo permite contemplar también otras relaciones bibliográficas, como las traducciones.

[15] Sobre el funcionamiento modelo FRBR, podemos leer Tillett (2003).

Fig. 8. Utilización del elemento <relatedItem> vinculando la *Primera parte de las disposiciones reales… a entregar a la Serenísima señora Doña María Teresa Bibiana de Austria…* (manifestación) con la obra entendida como suma de sus partes (expresión)

Como se puede observar en este ejemplo, en el elemento <relatedItem>, remite con un código incluido en la etiqueta <ref target="“#expression_0003549”> a la expresión. Los pliegos que sean parte de una misma relación tendrán en común un mismo xml:id de expresión, pudiendo así un ordenador reconocer la existencia de este vínculo entre los dos archivos. A través del valor asignado al type de <relatedItem>, “isPartOf”, el ordenador reconocerá de qué tipo de vínculo se trata.

El sistema propuesto en nuestro modelo sirve también para recoger vínculos entre traducciones de una misma relación.

4.2. Etiquetas procedentes del módulo <TEXTCRIT>

De la mayor parte de relaciones de sucesos conocemos una única edición, pero, a través de CBDRS, se constata que existe un buen número de obras del género que fueron publicadas en más de una ocasión. En concreto, en este catálogo, se registran, al presente, hasta 488 ediciones para las que se ofrecen referencias a otras impresiones de la obra.

Fig. 9. Un ejemplo de una relación de sucesos con varias ediciones es esta *Copia de una carta que escrivio el muy R. P. F. Iacobo de Ambrosi*… Se pueden observar las portadas de tres de las cinco ediciones registradas en CBDRS de este texto. Estas fueron, respectivamente, impresas en Granada, Barcelona y Madrid en 1631; sus imágenes proceden de la Biblioteca de la Universidad de Sevilla, de la Biblioteca Nacional de España y de la Biblioteca de la Universidad de Barcelona y tienen los códigos 0002204E, 0002204B y 0002204C

En estos casos, podemos realizar o contar con trabajos de edición que tengan en consideración más de un testimonio, por lo que dentro de nuestro modelo, hemos previsto que, de manera optativa, puedan utilizarse etiquetas del módulo textcrit que nos permite implementar un aparato crítico. Dentro de <sourceDesc>, incluimos la lista de testimonios que se tienen en cuenta para la edición (<listWit>), describiendo cada uno de ellos e identificándolos con su sigla, que aparece como valor del xml:id de cada testimonio <witness>.

Cuando queremos dejar constancia de las variantes que se presentan

en un punto del texto, introducimos la etiqueta <app>, utilizada en TEI para las entradas del aparato crítico y, dentro de ella, <lem> y <rdg>, para marcar una lectura preferida por el editor y lecturas alternativas. Pueden, por ejemplo, utilizarse así:

> <app><lem wit= "#S2"> y los</lem> <rdg wit="#S1">y</rdg> <note>En el testimonio base se omite la conjunción «y» que el sentido de la oración exige y que sí se presenta en S2</note></app>.

5. Conclusiones

El establecimiento de un modelo supone una fase esencial de la labor de edición académica digital. A lo largo de estas páginas, hemos volcado nuestras reflexiones sobre el propio objeto de trabajo, señalando características esenciales y opcionales y desechando las no útiles, estableciendo qué etiquetas son precisas para marcar nuestros textos (Mancinelli y Pierazzo, 2020, 48). Con todo, somos conscientes de que este es un proyecto en curso y que aún quedan tareas por desarrollar para establecer un flujo de trabajo definitivo. Entre ellas, determinar el mecanismo para la visualización de archivos TEI que se integraría en *BIDISO Textos*. Así, nos proponemos experimentar con herramientas de acceso abierto como EVT (*Edition Visualization Technology*) o TEI Publisher, que permitirían a nuestros usuarios consultar en paralelo el facsímil digital, la edición paleográfica y la edición crítica del texto e interactuar con la aplicación para navegar, explorar y estudiar nuestras ediciones digitales de relaciones de sucesos. En definitiva, hemos emprendido el camino para resolver la necesidad de ediciones académicas digitales que hace surgir la idea de *BIDISO textos* y esperamos poder compartir pronto más progresos.

§

Bibliografía citada

Allés Torrent, Susana, «Tiempos hay de acometer y tiempos de retirar: literatura áurea y edición digital», *Studia Aurea*, 11 (2017), pp. 13-30. URL: <https://studiaaurea.com/article/download/v11-alles/261-pdf-es> (cons. 30/05/2022).

Baena Sánchez, Francisco; Fernández Travieso, Carlota; Espejo Cala, Carmen; Díaz Noci, Javier, «Codificación y representación cartográfica de noticias. Aplicación de las humanidades digitales al estudio del periodismo de la Edad moderna», *El profesional de la información*, 23/5 (2014), pp. 519-526.

BIDISO: Biblioteca Digital Siglo de Oro. URL: <http://www.bidiso.es/> (cons. 29/10/2021).

Fernández Travieso, Carlota, *Estudio de codificación XML/TEI para Relaciones de sucesos españolas*, A Coruña, SIELAE, 2013. URL: <https://www.bidiso.es/sielae/upload/estaticas/file/FTXMLTEIISBN2pr.pdf> (cons. 30/05/2022).

Franzini, Greta (2012), *Catalogue of Digital Editions*. URL: <https://dig-ed-cat.acdh.oeaw.ac.at> (cons. 29/10/2021).

Garrobo Peral, Manuel, *Fiesta que se hizo en Aranjuez por los años de Felipe IV en 1622. Hacia una edición digital académica de una relación de sucesos*, Trabajo de fin de master del Máster en Literatura Cultura y Diversidad, Universidade da Coruña, dirs. Nieves Pena Sueiro y Stefano Neri, defensa: julio 2020.

González Blanco, Elena, «La edición digital de textos literarios: planteamientos y perspectivas de futuro», *Rilce*, 33/1 (2014), pp. 239-58. URL: <https://revistas.unav.edu/index.php/rilce/article/view/137> (cons. 30/05/2022).

Huarte Echenique, Amalio, *Relación verdadera del rebato que dieron cuatrocientos y cincuenta turcos en la almadraba de Zahara*…, in *Relaciones de los reinados de Carlos V y Felipe II*, Madrid, Sociedad de Bibliófilos Españoles, vol. 1, 1941, pp. 161-166.

Infantes, Víctor, «¿Qué es una relación? (Divagaciones varias sobre una sola divagación)», en *Las relaciones de sucesos en España (1500-1750). Actas del primer coloquio internacional (Alcalá de Henares, 8, 9 y 10 de junio de 1995)*, coords. Henry Ettinghausen, Víctor Infantes de Miguel, Augustin Redondo, María Cruz García de Enterría, Alcalá de Henares, Universidad de Alcalá-Publications de la Sorbonne, 1996, pp. 203-216.

Mancinelli, Tiziana y Elena Pierazzo, *Che cos'è un'edizione scientifica digitale*, Roma, Carocci, 2020.

Martín Molares, Mónica, «Paratextos legales en las relaciones de sucesos impresas entre 1550 y 1650», en *Doce siglos de materialidad del libro: estudios sobre manuscritos e impresos entre los siglos VIII y XIX*, dir. Manuel José Pedraza Gracia, eds. Helena Carvajal González y Camino Sánchez Oliveira, Zaragoza, Universidad de Zaragoza, 2017, pp. 365-383.

Nider, Valentina y Nieves Pena Sueiro (eds.), *Malas noticias y noticias falsas. Relaciones de sucesos en los siglos XVI y XVII*, Trento, Università degli Studi di Trento-Dipartimento di Lettere e Filosofia, 2019.

Peiró Sempere, Julio, Mireya Fernández Merino, Elena Martínez Carro, «Edición digital y electrónica en España: un estado de la cuestión», *Texto Digital*, 11/1 (2015), pp. 339-354. DOI: <https://doi.org/10.5007/1807-9288.2015v11n1p339> (cons. 30/05/2022).

Pena Sueiro, Nieves, «El portal BIDISO: pasado, presente y futuro inmediato. Un ejemplo de evolución en aplicaciones de las HD», *Studia aurea*, 11 (2017), pp. 73-92. DOI: <https://doi.org/10.5565/rev/studiaaurea.264> (cons. 30/05/2022).

Pena Sueiro, Nieves, «El título de las *Relaciones de sucesos*», en *La fiesta: actas del II Seminario de Relaciones de Sucesos (A Coruña, 1998)*, eds. Sagrario López Poza y Nieves Pena Sueiro, Ferrol, Sociedad de Cultura Valle Inclán, 1999, pp. 293-302.

Pena Sueiro, Nieves, «Estado de la cuestión sobre el estudio de las Relaciones de sucesos», *Pliegos de Bibliofilia*, 13 (2001), pp. 43-66. URL: <https://www.bidiso.es/upload/estadocuestion.pdf> (cons. 30/05/2022).

Reyes Gómez, Fermín de los, «Los impresos menores en la legislación de imprenta (siglos XVI-XVIII)», en *La fiesta: actas del II Seminario de Relaciones de Sucesos (A Coruña, 1998)*, eds. Sagrario López Poza y Nieves Pena Sueiro, 1999, pp. 325-338.

Rojas Castro, Antonio, *Modeling FRBR entities and their relationships with TEI. Zenodo* [diapositivas] (2019). DOI: <https://doi.org/10.5281/zenodo.3446218> (cons. 29/10/2021).

Ruiz Astiz, Javier y Nieves Pena Sueiro, «Presentación. Las relaciones de sucesos: producto y género editorial en la Monarquía Hispánica», *Memoria y Civilización*, 22 (2019), pp. 371-380. DOI: <https://doi.org/10.15581/001.22.010> (cons. 30/05/2022).

Sahle, Patrick, (1994), *Digital Scholarly Editions*. URL: <https://v3.digitale-edition.de> (cons. 29/10/2021).

Sahle, Patrick, «What is a Scholarly Digital Edition?», en *Digital Scholarly Editing. Theories and Practices*, eds. Elena Pierazzo y Matthew James Driscoll, 2016, pp. 19-39.

Spedialeri, Graciela (2009), *FRBR: antecedentes, estructura e impacto* [diapositivas], 2009. URL: <https://www.loc.gov/catdir/cpso/frbryfrad/frbr-instructor_oct09.pdf> (cons. 13/05/2022).

Tillett, Barbara, *What is FRBR? A conceptual model for the bibliographic universe*, Washington D.C., Library of Congress, 2004. URL: <https://www.loc.gov/cds/downloads/FRBR.PDF> y <https://www.loc.gov/catdir/cpso/frbrspan.pdf> (cons. 13/05/2022).

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Reflexiones sobre la creación de una base de datos de motivos caballerescos: un desafío científico y digital

Federica Zoppi

(Università di Verona)*

Abstract

Se presentan unas reflexiones sobre las dificultades del trabajo de catalogación de motivos caballerescos, a partir del trabajo realizado por el grupo de investigación del Progetto Mambrino de la Università di Verona sobre el corpus de las traducciones y continuaciones italianas de los ciclos narrativos españoles. Se examinan algunos de los principales problemas metodológicos que el estudio de los motivos plantea, considerando las experiencias previas y los índices de motivos ya existentes, incluso en otras áreas científicas. Finalmente, se quieren analizar las posibilidades que el entorno digital ofrece en el ámbito de este intento de catalogación en una base de datos.
Palabras clave: libros de caballerías; motivos; Stith Thompson; Humanidades Digitales; base de datos

The article aims to give an overview of the issues encountered so far in the creation of a catalogue of chivalrous motifs, presenting the work carried out by the research group of Progetto Mambrino, from the Università di Verona, on the corpus of Italian translations and sequels of Spanish chivalric novels. We will examine some of the main methodological problems posed by the study of motifs, taking into account previous experiences and existing indexes of motifs, also in other scientific areas. Finally, we want to explore the possibilities that the digital environment offers in the composition of a database of motifs of chivalric novels.
Keywords: romances of chivalry; motifs; Stith Thompson; Digital Humanities; database

§

* Este trabajo se ha realizado en el marco del Proyecto de Investigación PRIN 2017 «Mapping Chivalry: Spanish Romances of Chivalry from Renaissance to 21th Century: a Digital approach», concedido por el MIUR y cofinanciado por la Università di Verona. Se inscribe en las tareas del Progetto di Eccellenza (2018-2022) del Dipartimento di Lingue e Letterature Straniere de la Università di Verona «Le Digital Humanities applicate alle lingue e letterature straniere».

El estudio de los libros de caballerías castellanos representa un ámbito de la investigación de la filología hispánica rico en desafíos y ambigüedades ya desde la época en la que este género nació y se difundió, convirtiéndose en un fenómeno cultural casi masivo que, con las oportunas objeciones que deben mediar este aserto, parece ser comparable a los *best-sellers* actuales (Díez-Borque, 1995, 79). A pesar de su difusión, es bien reconocido el juicio negativo que esta producción suscitó por parte de los intelectuales, que, a su vez, en la historia de la crítica, se ha convertido casi en un tema tópico, definido por Sarmati (1996: 23), de forma pertinente, como «*tópos* del biasimo». La recepción del género caballeresco representa, de hecho, un asunto repleto de contradicciones, que se establecen entre los polos opuestos del reproche formal y de la fascinación aparentemente irresistible que estos libros ejercieron en su público. En este sentido, es imposible no citar el *Quijote*, que representa perfectamente este tema, «novelizando» la tensión entre el juicio negativo de los bienpensantes y la capacidad de este corpus de conquistar a sus lectores para llevarlos a universos extraordinarios. La crítica cervantina a los libros de caballerías, con el escrutinio censor de las obras (I, 6) y los tajantes juicios del canónigo de Toledo (I, 48), representó durante siglos el eje en torno al cual se constituyó la aproximación crítica al género, resultando muchas veces en una «visión simplista y empobrecedora de su riqueza textual y narrativa» (Lucía Megías, 2019, 6), muy alejada de la propia perspectiva cervantina. De hecho, en el *Quijote*, si por una parte el canónigo de Toledo afirma que los libros de caballerías, «cuál más, cuál menos, todos ellos son una mesma cosa» (I, 47), por otra parte Sansón Carrasco le reprocha a los lectores de la época sus severas críticas a las incongruencias narrativas, reconociendo el valor que puede hallarse en los descuidos y formulando un auténtico «elogio de la imperfección» (Pini y Castillo Peña, 2013):

> Quisiera yo que los tales censuradores fueran más misericordiosos y menos escrupulosos, […] consideren lo mucho que estuvo despierto por dar la luz de su obra con la menos sombra que pudiese, y quizá podría ser que lo que a ellos les parece mal fuesen lunares, que a las veces acrecientan la hermosura del rostro que los tiene (DQ II, 3).

A la luz de todo esto, el estudio de los libros de caballerías a través de la identificación de motivos recurrentes se configura como una aproximación especialmente fructífera, precisamente porque nos restituye las dos caras del género, es decir, su tendencia a la uniformidad y a la repetición y, por otra parte, sus «lunares», aunque no propiamente en el sentido original de errores o incoherencias narrativas, sino más bien como variantes, desvíos del modelo principal que aportan una fecunda proliferación narrativa. En las palabras de Bueno Serrano:

> Lo común es el paso previo para discriminar lo diferente; son los desvíos los que marcarán la originalidad de la ficción aportando datos concretos sobre la evolución del género. Desde esta perspectiva, la prosa de ficción caballeresca castellana constituye un corpus dinámico, si bien esta capacidad de evolución y adaptación queda sometida a su inicial estabilidad que lo identifica como género (2007b, 88).

El objetivo central del Progetto Mambrino de la Università di Verona es trasladar esta aproximación metodológica a un entorno digital, con la creación de una base de datos de los motivos de los libros de caballerías italianos de derivación española (1540-1630); se quiere ofrecer este corpus textual (de unas 50 novelas, entre traducciones y continuaciones) en una biblioteca digital que reúna, en forma sinóptica, las reproducciones digitales facsimilares y la paralela transcripción de los textos; esta biblioteca digital se acompañará por unas bases de datos que permitan la recuperación de la información gracias a una plataforma digital interactiva que facilite un análisis semántico del corpus, en el ámbito de un sistema expandible. Se tomará como punto de partida el ciclo italiano de Amadís de Gaula, que comprende 25 novelas, 12 traducciones del español y 13 continuaciones originales.

Para llevar a cabo este proyecto, auténtica hazaña caballeresca, es necesario enfocar unos nudos teóricos y metodológicos en los que se asienta el mismo trabajo, aunque sin ninguna pretensión de desenredarlos; vamos entonces a proporcionar unas reflexiones sobre algunas dificultades que se están encontrando, tanto en un nivel científico preliminar, como en el plano de la realización digital.

Introducción: la catalogación de motivos y su evolución hacia lo digital

El estudio de los motivos representa un área de interés que se ha consolidado con numerosas aportaciones teóricas y metodológicas sobre su naturaleza y función. La clasificación de motivos se origina en el ámbito del folklore y de la recogida sistemática de este material (Cacho Blecua, 2020, 9); la propuesta pionera de la aplicación de esta metodología al estudio del folklore fue la de Antti Aarne con su sistema de clasificación de cuentos de hadas, en 1910 (*The Types of the Folktale: A Classification and Bibliography*), ampliado y traducido al inglés por Stith Thompson en 1928 y, posteriormente, en 1961, cuando se fijó una segunda revisión, con ulterior ampliación, que estableció definitivamente este sistema clasificatorio como el Aarne-Thompson (AT o AaTh). El mismo Thompson completó esta tarea con un índice propio entre 1932-1936, e incrementado en una segunda versión en 1955-1958 (*Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*); ya a partir de esta segunda edición se puede apreciar un rasgo esencial de esta metodología de trabajo que representa el eje central de su éxito, también en los estudios de crítica literaria y filológica, es decir, la posibilidad de implementaciones que conlleva. De hecho, el catálogo de Thompson se ha convertido en una obra de referencia de gran difusión, paradigma para la elaboración de otros índices fundados en la misma clasificación metodológica; algunos de ellos quedan incorporados en la segunda edición del *Index*, estableciendo así una interacción entre catálogos y, consecuentemente, subrayando la existencia de una tradición común de la que provienen o sacan inspiración varias expresiones culturales: la clasificación sistemática de motivos presenta una significativa capacidad de adaptación a ámbitos distintos y, al mismo tiempo, proporciona la posibilidad de revelar los vínculos existentes entre estos ámbitos.

La literatura es, evidentemente, el sector científico donde esta metodología de investigación se ha desarrollado de manera especialmente eficaz, tratándose también de un área que en muchos casos se relaciona,

de forma más o menos directa, con el folklore. Por lo tanto, se han elaborado, a lo largo de casi un siglo, varios índices que clasifican los motivos propios de un determinado género literario, tanto en una concreta literatura nacional como en la producción de un autor o en una obra, dando prueba de cómo la perspectiva práctica y la utilidad de este sistema ha superado las críticas –aunque legítimas– a su estructura:

Index of Spanish folktales [...], de Ralph. S. Boggs (1930);
Motif-Index of the Italian Novella in Prose, de Dominic Peter Rotunda (1942);
Motif-Index of Mediaeval Spanish Exempla, de John E. Keller (1949);
Motif-Index of Early Irish literature, de Tom Peete Cross (1952);
Motif-Index of the English Metrical Romances, de Gerald Bordman (1963);
Folk-Motifs in the Medieval Spanish Epic, de Alan D. Deyermond y Margaret Chaplin (1972);
Tales from Spanish Picaresque Novels: a Motif-Index, de James Wesley Childers (1977);
Motif-Index of the cuentos *of Juan Timoneda*, de James Wesley Childers (1980);
Index des motifs narratifs dans les romans arthuriens français en vers (XIIe-XIIIe siècles), de Anita Guerreau-Jalabert (1992);
Types and Motifs of the Judeo-Spanish Folktales, de Reginetta Haboucha (1992);
Motif-Index of Medieval Catalan Folktales, de Edward J. Neugaard (1993);
Folk Traditions of the Arab Worl: A Guide to Motif Classification, de Hasan M. El-Shamy;
Motif-Index of Medieval Spanish Folk Narratives, de Harriet Goldberg (1998);
Motif-Index of Folk Narratives in the Pan-Hispanic romancero, de Harriet Goldberg (2000);
The Types of International Folktales […], de Hans Jörg Uther (2004);
Motif-Index of German Secular Narratives from the Beginning to 1400, editado por la Austrian Academy of Sciences, bajo la dirección de Helmut Birkhan (2005-2010);

Archetypes and Motifs in Folklore and Literature: a Handbook, de Jane Garry y Hasan El-Shamy (2005);
A Motif Index of The thousand and one nights, de Hasan M. El-Shamy (2006).

La «expandibilidad» es precisamente uno de los rasgos de estos sistemas de catalogación que hacen esta metodología adaptable a un entorno digital y, en concreto, a una base de datos.

De hecho, se han ido desarrollando varios proyectos orientados a este objetivo en el ámbito de las Humanidades Digitales. En primer lugar, se han elaborado varias versiones digitales del *Index* de Thompson para facilitar su consulta y, a la vez, resolver implícitamente algunos de los problemas organizativos de su materia (Ardanuy Baró, 2016). Además de una versión en CD-Rom de la Indiana University Press, salida en 1993, la obra puede consultarse en su versión completa en línea en el enlace <http://www.ruthenia.ru/folklore/thompson> (cons. 10/05/2022), que forma parte del proyecto de investigación del folclorista ruso Artem Kozomin; y en el enlace <https://sites.ualberta.ca/~urban/Projects/English/Motif_Index.htm> (cons. 10/05/2022), a cargo de Shawn Urban, que también recoge la clasificación ampliada por Uther (Aarne-Thompson-Uther's *Tale Type Index*), las funciones de Propp y el análisis estructural del mito de Lévi-Strauss. Más completo –y más sofisticado– es el programa MOMFER (Meertens Online Motif FindER, <http://www.momfer.ml>, cons. 10/05/2022), elaborado por Folger Karsdorp, Marten van der Meulen, Theo Meder y Antal van den Bosch en 2015 (Karsdorp *et al.*, 2015), que posibilita la recuperación de motivos a través de un buscador que admite también búsquedas semánticas y devuelve no solo los motivos que incluyen los términos de la propria interrogación, sino también los relacionados conceptualmente con ellos.

Los proyectos de Humanidades Digitales no se dedicaron solo al *Index* de Thompson, sino también a otros índices de motivos, como por ejemplo el Proyecto sobre el Romancero pan-hispánico de Harriet Goldberg <http://depts.washington.edu/hisprom/> (cons. 10/05/2022), de 2000, que incluye una base de datos bibliográfica y una textual sobre los romances, y el *Índice de motivos folklóricos en el Romancero*,

«modificado y ampliado para optimizar su implementación y funcionalidad en la web»; los motivos aparecen ordenados en tres listas distintas según tres criterios: 1) por categoría, según las planteadas por Thompson, 2) por orden alfabético, y 3) por romance, con enlaces que relacionan los motivos con los textos de los romances donde aparecen. Otro proyecto digital que merece la pena destacar fue realizado por el grupo SAToR entre 1997 y 2020, con una base de datos de motivos de la literatura en lengua francesa anterior a la Revolución Francesa, ahora inaccesible, también organizada en tres índices: 1) por categorías de motivos: descriptivos, narrativos, y discursivos, 2) por motivos, y 3) por obras repertoriadas (Zoppi, 2019, 335 y Bueno Serrano 2007b, 31 y 38 ss.). Tomasi (2020, 139 ss.), en un amplio ensayo sobre las principales herramientas informáticas para el estudio de la literatura castellana del Siglo de Oro, señala unos proyectos que se dedican también a la clasificación de motivos, en particular la base de datos Calderón Digital <http://calderondigital.tespasiglodeoro.it/> (cons. 10/05/2022), dirigida por Fausta Antonucci, que incluye en los resúmenes de las obras las indicaciones de los motivos dramáticos detectados, elaborados a partir de la clasificación de Rotunda de los motivos de la *novella* italiana. A esta se añade el recurso MeMoRam, en fase de desarrollo, bajo la dirección de Claudia Demattè, que tiene el propósito de analizar las novelas caballerescas españolas de los siglos XVI-XVII precisamente a través del enfoque del reconocimiento de sus motivos.

El recurso MeMoRam se sitúa en el marco de un proyecto de investigación más amplio denominado *Mapping Chivalry. Spanish Romances of Chivalry from Renaissance to XXI century: a Digital approach (2018-2022)*, coordinado por Anna Bognolo, en el cual se inscribe también el trabajo del Progetto Mambrino para la realización de la base de datos de motivos caballerescos, objeto de la reflexión que nos ocupa. El proyecto también incluye la publicación de ediciones científicas digitales –Digital Scholarly Editions (DSE)– de las traducciones y continuaciones italianas, recogidas en una biblioteca digital –Digital Library (DL).

1. La metodología científica: unas reflexiones preliminares

1.1. La definición de motivo

En primer lugar, hay que determinar con claridad cuál definición de motivo queremos asumir para nuestro trabajo; lo cual está en el mismo fundamento de la clasificación que vamos a realizar, puesto que afecta tanto a la etiqueta que se atribuye al motivo, como al elemento narrativo que el propio motivo selecciona.

La flexibilidad y la aplicación del concepto de motivo a ámbitos culturales distintos, desde el folklórico al literario, ha representado un estímulo para la formulación de definiciones distintas. De hecho, el concepto de motivo se ha caracterizado y definido en múltiples maneras, según perspectivas críticas distintas; si esto implica un cuadro complejo desde el punto de vista teórico, este dinamismo también puede considerarse un recurso y una riqueza:

> Esta unidad narrativa […] requiere una formulación sistemática, adaptada a las peculiaridades de los objetos estudiados, a nuestro juicio a partir de esquemas narrativos o discursivos, aplicados con sistematicidad. A partir de su definición y características podremos construir nuevos índices, adaptando en mayor o menor grado los esquemas de Thompson o bien olvidándonos de estas herencias originarias (Cacho Blecua, 2020, 45).

En líneas generales, la perspectiva empírica de los etnógrafos y folkloristas, punto de partida del índice de Thompson, se opone a la aproximación teórica de los estructuralistas y formalistas, en particular de Propp, quien propone un sistema teórico abstracto que, a pesar de su minucioso diseño, parece acabar siendo poco eficaz bajo un criterio utilitario de aplicación práctica: la caracterización de Propp del motivo como una unidad invariable en cuanto a su función –así como la perspectiva estructuralista de Lévi-Strauss, que prefiere hablar de «mitema», núcleo narrativo invariable que compone la narración mítica– se halla en un ámbito de reflexión abstracta que no se traduce en una categoría productiva que tenga significación en la especificidad del análisis textual; en resumidas cuentas, la caracterización formalista del motivo no

considera su realización contextual como un factor que influya en la función que el propio motivo asume. Precisamente este punto de reflexión es el que se rechaza en las perspectivas más pragmáticas de definición de motivo: «los motivos deben ser analizados en relación con la función que desempeñan en la obra, teniendo en cuenta su situación en la intriga en la que se insertan y en conexión con una tradición que ratifican, renuevan o crean» (Cacho Blecua, 2002, 51). Para acercarnos a una definición de motivo que tenga un distinto valor narrativo y, por lo tanto, pueda confluir en una categoría funcional para el estudio crítico-literario, han sido imprescindibles las aportaciones de Joseph Courtés, Claude Brémond y Cesare Segre en el ámbito de la semiología, tematología y, consecuentemente, de la narratología (Luna Mariscal, 2013, 28).

En el contexto del estudio que nos ocupa, es decir, el reconocimiento de motivos en el corpus de los libros de caballerías de tradición castellana, es imprescindible la contribución de Juan Manuel Cacho Blecua (2002; 2012), quien formuló por primera vez la propuesta de la creación de «un índice completo de motivos de los libros de caballerías» (Cacho Blecua, 2002, 51). El estudioso, por lo tanto, nos ofrece el planteamiento más útil (en relación con nuestros objetivos) de la cuestión de los motivos, por lo menos en lo que atañe a su aplicación a un contexto narrativo. En la perspectiva del estudioso, el motivo se caracteriza por ser una unidad narrativa recurrente y estereotipada del contenido, que presenta cierta persistencia en la tradición, aunque pueda manifestarse con variaciones. La formulación de su propuesta teórico-metodológica parece ser deudora también de los estudios de Aurelio González (1990; 2003; 2012) aplicados al romancero tradicional (y más en concreto, a los romances caballerescos); los motivos se caracterizan como «unidades menores narrativas en las cuales se expresa el significado de las secuencias fabulísticas, o partes invariantes de la historia» (González, 2003, 381), relacionando el nivel narrativo (o del discurso) con aquel nivel de significación más profundo que identificamos como fábula. Luna Mariscal (2013, 40) resume este concepto en términos propios de la lingüística, afirmando que los dos niveles establecen la misma relación que vincula significante y significado. Cacho Blecua reelabora este planteamiento, pero recuperando la importancia de la presencia reiterada de un motivo en la

tradición, factor que González había dejado al margen de su reflexión y que, en cambio, es central en la perspectiva de Segre: a partir de una reflexión sobre el valor musical del motivo («minima unità musicalmente significativa», Segre, 1985, 340), el estudioso reflexiona sobre su carácter reiterativo y, a la vez, sobre su relación de interdependencia con el concepto de «tema», del que representa el núcleo germinal.

La definición de motivo proporcionada por Cacho Blecua nos parece la más adecuada en relación con las características propias del género caballeresco, porque destaca, además del carácter reiterado del motivo, su valor narrativo: la identificación de una serie de motivos –según la definición que abrazamos– permite trazar una guía narrativa de las acciones que componen la intriga de una obra a través de la identificación de unos esquemas repetitivos. Consecuentemente, a partir de esta definición, podemos también priorizar la acción como elemento central del motivo, tanto en la formulación de la etiqueta definitoria del mismo motivo, como en su catalogación (Laura Mariscal, 2013, 19-20), como se explicita más adelante.

1.2. El modelo estructural asumido para la clasificación

Hasta ahora hemos proporcionado un cuadro general de los principales catálogos de motivos que se han creado y que han originado la tradición metodológica y crítica que nos ocupa; a estos se añaden los proyectos que, sucesivamente, se han dedicado al traslado de estos instrumentos a un entorno digital, esencialmente en la forma de base de datos.

En este apartado nos centramos en considerar la aplicación de esta metodología a los libros de caballerías, objetivo del Progetto Mambrino y del proyecto PRIN 2017 *Mapping chivalry*, según se ha mencionado.

Como se ha indicado, varios géneros literarios se han estudiado también a la luz de la identificación de motivos; entre ellos se incluyen también los libros de caballerías españoles que, por su carácter repetitivo, representan un corpus especialmente adecuado para la aplicación de esta metodología de análisis.

En concreto, los trabajos de Bueno Serrano (2007b) y de Luna Mariscal (2013, 2017, 2020) se configuran como los puntos de partida para el desarrollo del tema y asimismo, modelos para la catalogación, al tratarse de sistemas de clasificación del corpus caballeresco castellano, a partir del cual se elabora la tradición italiana:

- *Índice y estudio de motivos en los libros de caballerías castellanos (1508-1516)* es la monumental tesis doctoral de Ana Carmen Bueno Serrano, de 2007, que se centra en los libros de caballerías castellanos publicados entre 1508 y 1516. En concreto analiza un corpus de siete obras: el *Amadís de Gaula*, las *Sergas de Esplandián*, el *Florisando*, el *Palmerín de Olivia*, el *Primaleón*, el *Lisuarte de Grecia* y el *Floriseo*.
- *Índice de motivos de las historias caballerescas breves*, otra tesis doctoral, defendida por Karla Xiomara Luna Mariscal en 2009 y publicada en 2013[1];
- *El motivo literario en* «El Baladro del sabio Merlín» *(1498 y 1535), con un índice de motivos de* «El Baladro del sabio Merlín» (*Burgos, 1498 y Sevilla, 1535)*, otro catálogo realizado por Karla Xiomara Luna Mariscal, publicado en 2017;
- *Índice de motivos de «La Demanda del Santo Grial» (Toledo, 1515)*, incluido en Luna Mariscal (2020)[2].

[1] Merece la pena señalar también otra tesis doctoral, la de Kristin M. Neumayer, *Index and study of plot motifs in some Spanish libros de caballerías*, de 2008, defendida en la Universidad de Wisconsin-Madison, que ofrece un estudio, fundado en la metodología de Propp, sobre motivos caballerescos del *Amadís de Gaula*, *Florisando* y *Lisuarte de Grecia* de Feliciano de Silva; Neumayer plantea unas categorías principales sin llegar a crear un catálogo concreto. Añadimos también la aportación de Lucía Megías (1996), que compara *La leyenda del Cavallero del Cisne* y el *Libro del caballero Zifar*, para crear un esquema estructural cuya aplicación se puede extender también a otros textos; en concreto, proporciona un índice que organiza el discurso en tres niveles: el motivo, la fórmula y la expresión formularia, relacionando la realización de la misma fórmula con un contexto determinado (Bueno Serrano, 2007a, 142).

[2] Este trabajo de Luna Mariscal, así como el anterior sobre el *Baladro del sabio Merlín*, pertenece al ámbito de estudio de la literatura artúrica; por las relaciones de este ámbito con la literatura caballeresca los asumimos como modelos próximos a nuestros intereses; consideramos también las características pragmáticas que iremos detallando y que representan un importante punto de partida –teórico y práctico– para nuestra catalogación.

Las dos estudiosas, alumnas de Juan Manuel Cacho Blecua, fundan sus clasificaciones en la misma definición de motivo que hemos asumido. A pesar de esto, el mismo Cacho Blecua a la hora de plantearse el problema de cómo relacionar un catálogo de motivos de la tradición caballeresca con el índice de Thompson había trazado dos líneas posibles: 1. la de aceptarlo en su conjunto general, con la conciencia de deber «incluir nuevos apartados o acomodar profundamente los existentes» (Cacho Blecua, 2002 ,52); esta es la vía seguida por Luna Mariscal, que, en sus clasificaciones, decide aceptar implícitamente la pragmática propuesta por Thompson (2013, 17). Además, en su índice se integra la propuesta de Guerreau-Jalabert, por la vinculación que establece entre la literatura artúrica y la tradición folklórica y «por suponer un intento de adecuación, aún no superado, del trabajo de Stith Thompson a un corpus de literatura caballeresca» (Luna Mariscal, 2010, 128). 2. La segunda posibilidad planteada por Cacho Blecua es la de proponer un esquema distinto, a partir de una definición más precisa de motivo. Esta es la vía elegida por Bueno Serrano, que considera el modelo de Thompson inadecuado para catalogar los motivos del corpus caballeresco:

> la aplicación del *Motif-Index* a nuestro corpus se hace más por aproximación que por identificación literal de formas o paradigmas y, por ello, el género caballeresco queda inexplicado o mal explicado en muchos casos, y estas carencias no se solucionan en su totalidad con adiciones a la estructura predeterminada (2007b, 89).

La catalogación de motivos a la que nos dedicamos para nuestra base de datos se fundará esencialmente en el modelo asumido por Luna Mariscal, aunque se considerará la propuesta de Bueno Serrano en su planteamiento teórico, así como en las categorías propuestas, que representan una inestimable mina de informaciones que hay que tener en cuenta. Las razones que nos han llevado a tomar esta decisión son esencialmente dos: en primer lugar, la intención de participar en un diálogo científico internacional sobre los motivos, que ya está establecido y que se sigue fundando en el modelo de Thompson, a pesar de los innegables fallos detectados. La mayor parte de los índices existentes, como acabamos de ver, siguen sus pautas, aplicándolo a varios géneros literarios y a

distintas literaturas nacionales y, entre ellos, casi la totalidad de los índices dedicados a la literatura artúrica (Luna Mariscal, 2020, 56). En palabras de Uther (2009, 26), autor de la revisión, corrección y ampliación del sistema tipológico de Aarne-Thompson, «in spite of the criticism concerning current classification systems and in spite of the imprecise definitions of the type and the motif, no feasible countermodel has been suggested. It seems that the only choice is to stick to the old systems and try to improve them whenever possible».

En segundo lugar, nunca podemos olvidar que estos índices se fundan en un criterio utilitario: su propósito es de convertirse en una herramienta de consulta lo más sencilla e intuitiva posible. Si bien Bueno Serrano (2007b, 89) comparte este objetivo general, el índice que propone nos parece más complejo en su estructura, así que su manejo puede resultar más dificultoso.

Un nudo central de esta clasificación es la distinción entre motivos sintagmáticos y paradigmáticos; este planteamiento no pertenece solo a la configuración teórica de Bueno Serrano y, como veremos en el apartado siguiente, atañe a uno de los problemas principales con los que nos enfrentamos en la elaboración de un índice, es decir, la determinación del nivel de abstracción de los motivos. En la perspectiva de Bueno Serrano, el aspecto paradigmático de un motivo se halla en su contenido estable y, por lo tanto, más abstracto, mientras que el aspecto sintagmático representa su realización concreta, que suele manifestar formas variables y funciones distintas en la cadena narrativa. De hecho, la catalogación de Bueno Serrano proporciona una extraordinaria cantidad de información precisamente sobre las variantes de los motivos y su acomodación a varias circunstancias textuales; sin embargo, por lo menos en las fases iniciales del trabajo, nuestro punto de vista se mantendrá en un nivel más abstracto, intentando formular categorías de motivos más genéricas que pueden abarcar y describir de forma pertinente más variantes y que, por lo tanto, manifiesten una recurrencia relevante. Por lo tanto, nos centraremos en la identificación de los motivos sin explicitar nuevos enunciados para identificar cada una de las variantes detectadas, limitándonos a señalar solo las más significativas.

Como ya se ha mencionado, mantendremos como esquema central el proporcionado por Luna Mariscal, que sigue la ordenación alfanumérica de Thompson, integrando los motivos folklóricos ya presentes en el *Motif-Index* con las clasificaciones de Bordman, Guerreau-Jalabert, Goldberg y Birkhan; a estos se añadirán las nuevas indicaciones que se detectarán rastreando los textos. Considerando las especificidades del corpus que nos ocupa, parece oportuno tener en cuenta también el índice de Rotunda, que, al dedicarse a la *novella* italiana, puede revelarse útil en el análisis de un corpus italiano renacentista. En sentido general, aplicaremos un criterio de uniformidad con esta tradición teórico-metodológica: se intentará conformarse con los índices existentes, aplicando a nuestros textos las categorías ya formuladas para evitar la proliferación de etiquetas definitorias análogas. Al mismo tiempo, no queremos perder de vista nuestro objetivo principal, es decir, elaborar un catálogo funcional a la descripción del corpus, así que nos desviaremos de los modelos existentes cada vez que se consideren insuficientes a la luz de este objetivo, integrándolos con nuestras propuestas originales. De esta manera, esperamos compaginar el propósito de descripción de las especificidades del corpus con el diálogo con la tradición folklórica y medieval de la que los libros de caballerías son deudores y, contemporáneamente, con el diálogo con la tradición metodológica en la que nos insertamos, aprovechando el carácter de interoperabilidad de un sistema de catalogación infinitamente expandible.

1.3. La determinación del nivel de abstracción

Luna Mariscal (2013, 28) reconoce que la principal dificultad teórica en la elaboración de un catálogo de motivos es la determinación del nivel de abstracción en el que definir el motivo. Se trata de un problema que la mayoría de los estudiosos se ha planteado, aunque en ámbito de disciplinas distintas: representa de hecho el interrogativo central al que Segre (1985) intenta contestar al analizar la relación entre «tema» y «motivo», así como el núcleo de los numerosos estudios de Courtés y de Brémond, que precisamente a partir de esta reflexión criticaron el sistema de catalogación

de Thompson: por una parte, Courtés propone llevar el concepto de motivo a un nivel más abstracto respecto al planteamiento de Thompson, que lo reduciría a un mero elemento de clasificación empírica; Courtés (1980c) emplea, entonces, una nueva terminología definitoria, hablando de «motivema», de la que se excluyen las variables de sus realizaciones, para considerar solo su valor mitológico y antropológico y, por lo tanto, estable. Por otra parte, Brémond (1987, 124) habla de «proposición narrativa elemental», internamente organizada como un relato, que sería, de hecho, un elemento de un relato más extenso (Brémond, 1980, 18): da, entonces, cuenta de un proceso que incluye una acción (verbo), los actantes (sujeto) y las eventuales circunstancias (atributos); evidentemente, el enfoque de Brémond relaciona el motivo con lo particular, es decir, con sus manifestaciones narrativas y su consecuente función en la intriga, de la que hace depender su relación con la tradición.

Ya se ha presentado la distinción propuesta por Bueno Serrano entre motivo sintagmático y paradigmático que relabora la perspectiva de González (2003), según la cual el motivo se caracteriza por ser una unidad de significación que expresa un contenido fabulístico en el plano de la intriga/discurso, es decir expresa una secuencia propia de la fábula como significado específico de la historia contada (intriga, la historia que concreta como texto el plano de la fábula) a través de un significante (discurso, en el que los motivos se configuran como fórmulas): intentando simplificar, los motivos expresarían el significado no literal (es decir más abstracto, que hace referencia a la fábula) de las fórmulas (2003, 379).

De hecho, no es posible encontrar una solución definitiva a esta cuestión ni identificar una respuesta unívoca, puesto que, al fin y al cabo, «el nivel de abstracción depende del criterio del analista, del conocimiento del corpus y del contexto cultural» (Vázquez Recio, 2000, 19). A pesar de esto, se trata de un problema central en nuestro trabajo, tanto bajo el punto de vista teórico, como práctico, puesto que afecta la misma organización del catálogo y su realización visual en el ámbito digital, determinando la jerarquización en niveles y subniveles.

En la delimitación del nivel de abstracción de nuestra propuesta se mantiene presente la distinción entre un nivel sintagmático y uno paradigmático, aunque sin separarlos en dos índices distintos según el

modelo de Bueno Serrano. Un mismo motivo, entonces, tendrá, en un nivel de abstracción mayor, un valor paradigmático, que atañe a la mitología, al folklore y a la esfera cultural en general, y, en un nivel de abstracción menor, un valor sintagmático, es decir narrativo, que expresa un significado concreto en una obra concreta. A partir de la relación entre estos dos niveles, es decir, de cómo un motivo se concreta en un texto como expresión de una unidad de contenido estable que forma parte de la tradición, se puede desarrollar un análisis sobre la recepción histórica de un texto (estudio sincrónico), así como un estudio comparativo sobre la transformación y evolución de un motivo en la tradición y su adaptación a lo largo del tiempo a textos distintos (estudio diacrónico).

En sentido general (sin considerar las excepciones que sin duda se irán encontrando en la práctica), un motivo se identificará a partir de un enunciado que enfoque una acción (verbo) y un/os agente/s o paciente/s (sujeto/s), al que se añadirán eventuales atributos calificativos, tanto de la acción como del sujeto (de tiempo, lugar, manera, etc.). La formulación del enunciado intentará respetar la relevancia de los rasgos distintivos del propio motivo, así que se dará prioridad a la acción o al sujeto dependiendo de qué elemento se configure como el principal en determinar el significado del motivo.

Los catálogos ya existentes son sin duda un patrimonio del que nos podremos aprovechar también para solucionar eventuales casos dudosos, respetando nuestro propósito de dialogar con la tradición científica anterior y de perpetuarla cuando deje constancia de su eficacia según el ya mencionado criterio utilitario.

Para la incorporación de nuevos motivos se considerará esencialmente su recurrencia, tanto intertextual como intratextual; criterio alternativo será la evaluación de la incidencia simbólica o estructural de un motivo en el corpus: aunque no se detecte una recurrencia significativa, un motivo puede representar un elemento caracterizador, que vincula la obra al género al que pertenece (en el plano sintagmático) o que establece una dependencia de un contexto cultural. Evidentemente, en este segundo caso, es fundamental la personal evaluación del investigador y, sobre todo, su conocimiento científico del contexto histórico-cultural y del género literario en su conjunto; podremos entonces aprovechar nuestro

conocimiento de los libros de caballerías castellanos y evaluar cómo el corpus italiano se relaciona con ellos en la elaboración de motivos para comprobar si se puede apreciar una diferencia concreta procedente de los distintos contextos culturales.

Precisamente con este mismo criterio se considerará la posibilidad de incluir variantes de motivos: en primera instancia, la intención es la de identificar los motivos más relevantes, tanto desde un punto de vista sintagmático como paradigmático; la integración de variantes se limitará a las ocurrencias más significativas, no solo por ser frecuentes, sino también porque representan un desvío que aporte nuevas informaciones, por ejemplo, anclando el motivo al universo específico de referencia de la corte renacentista italiana, otra vez en relación con su importancia simbólica y estructural.

2. Para una base de datos de motivos: ventajas e hipótesis de estructura

Como se ha mencionado, el proyecto de una base de datos digital de motivos se inscribe en un contenedor más amplio que explora varias posibilidades en el ámbito de las Humanidades Digitales: en concreto, el Progetto Mambrino se va configurando como una biblioteca digital de acceso abierto que pretende recoger las reproducciones digitales en facsímil de los libros de caballerías italianos, visualizadas de forma sinóptica y acompañadas por una transcripción de los propios textos. El proceso de transcripción se está realizando de manera automática –o sería más adecuado decir semi-automática– gracias al empleo de *software* OCR (Bazzaco, 2018; Bognolo y Bazzaco 2019; Bazzaco 2020): se trata de herramientas de reconocimiento óptico que, a través de la lectura de la imagen digital, convierten los signos gráficos en un texto procesable, a partir del cual se podrá realizar la edición científica digital de los volúmenes. Para conseguir este resultado, los textos serán indexados con un etiquetado TEI, manteniendo la relación con las imágenes digitalizadas de las fuentes originales con una visualización *split screen* (Bognolo y Bazzaco, 2019, 29-31).

La modelización posibilita la elaboración de los datos textuales sobre los que se quiere llamar la atención (en concreto, nombres de personajes, topónimos y motivos). La elaboración de algunos de estos datos (en particular de los nombres de personajes) ya se ha llevado a cabo en dos repertorios, uno dedicado a las obras en italiano del ciclo amadisiano y otro a las del ciclo palmeriniano (Bognolo, Cara y Neri, 2013; Bognolo, Neri, Bellomi y Zoppi, en prensa); los dos volúmenes, que ofrecen asimismo resúmenes de las distintas novelas, se podrán consultar en el marco de la biblioteca digital.

Al configurar los motivos, según lo indicado, como una guía narrativa del enredo de una obra, nos parece adecuado aislarlos por medio del etiquetado TEI a partir de los resúmenes conseguidos, que ya nos han proporcionados unas pautas iniciales en el proceso de extracción de la información y simplificación del contenido, para intentar restituir –y reconstruir–, también de manera visual, la lógica de nuestro procedimiento de trabajo. Merece la pena recordar que el índice de Birkhan (2005-2010) se estructura según un planteamiento análogo, ofreciendo resúmenes de las secuencias narrativas de los textos que forman parte del corpus repertoriado, «con el fin de hacer lo más precisa posible la relación de un motivo con su contexto narrativo» (Luna Mariscal, 2020, 63); en el ámbito de los proyectos digitales, también Calderón Digital, ya citada, incluye los resúmenes de las piezas del autor, incorporando en ellos la indicación de los motivos rastreados.

Como se ha visto, a partir del análisis de algunos de los catálogos consultados, la mayoría de ellos acaba proporcionando más índices, ordenando el material según criterios distintos. En cambio, la estructura que estamos planteando elimina la necesidad de componer índices distintos dependiendo del criterio de búsqueda, con el objetivo de agilizar la consulta para los usuarios, así como el proceso de catalogación.

De hecho, el proyecto está orientado a la creación de un catálogo, no de un índice ordenado según un criterio alfabético o alfanumérico, como obligaría la publicación en papel. El soporte digital, y en concreto la base de datos, nos permite explorar una nueva organización de la información, que se mantendrá más dependiente del texto también en el aspecto visual, en conformidad con el propósito fundamental de estudiar los motivos

como elementos descriptivos del corpus. Por lo tanto, el catálogo se configura como una estructura de árbol, en el que cada motivo representa una rama distinta; de esta manera se posibilita la incorporación de ulteriores ramificaciones, según un criterio de expandibilidad.

En concreto, a cada motivo le corresponderá una ficha descriptiva, que representaría un eslabón de conexión entre la base de datos y los textos; se podrá acceder a las fichas a través de un buscador que el usuario podrá interrogar y que permitirá localizar la ficha correspondiente a un motivo sin recurrir al sistema clasificatorio alfabético o alfa-numérico, que en un contexto informático va perdiendo su utilidad. Por otra parte, las fichas estarán directamente conectadas a los textos y se podrá acceder a ellas a partir de la lectura de los mismos: para acceder al contenido de la ficha bastaría con hacer clic sobre el fragmento de texto marcado con etiquetado TEI que selecciona en el resumen el motivo identificado.

La ficha va precisamente a describir el motivo encontrado, proporcionando unas informaciones básicas: en primer lugar, el enunciado que describe el contenido del motivo: puede tratarse de una etiqueta original, de nuestra formulación, o de un motivo ya existente, sacado de otro catálogo, que describe adecuadamente nuestro texto (en este segundo caso se indicará también de qué catálogo procede, respetando el sistema empleado por Luna Mariscal en sus trabajos). A esto se añadirá, si es oportuno, la referencia a otros índices cuando el mismo motivo aparezca en varios catálogos formulado con expresiones distintas o cuando merezca la pena recoger eventuales repeticiones, es decir etiquetas diferentes que se atribuyen al mismo motivo (como es frecuente en el *Motif-Index*).

La ficha incluirá también las indicaciones de otras ocurrencias del motivo en el corpus; con la intención de ofrecer una descripción más completa, se quiere incluir asimismo una bibliografía de estudios pertinentes sobre el motivo o sobre los episodios en los que se localiza. Finalmente, se dará cuenta de las secuencias en las que el motivo suele aparecer con cierta frecuencia, según su carácter combinatorio (Bognolo, en prensa). La composición de una guía narrativa a través de los motivos, de hecho, nos permitirá averiguar el carácter combinatorio de un motivo y con qué otros motivos suele aparecer reiteradamente, planteando unas concatenaciones de unidades narrativas mínimas que representen el

desarrollo de acciones más articuladas (compuestas o complejas), tanto de forma vertical (antecedentes y causas, desarrollo principal, consecuencias) como horizontal (en el caso de pruebas menores que compongan, como piezas de un mosaico, una aventura compleja).

Según una propuesta de articulación inicial (Bognolo, en prensa), que nos permita empezar el trabajo de manera eficaz y esbozar un esquema que pueda ser sujeto a ampliaciones y actualizaciones, la identificación de los motivos se centrará esencialmente en tres ámbitos, tres macro-áreas dentro las cuales se rastrea un amplio abanico de motivos: 1. Aventura, 2. Corte y amor, 3. Maravilla, dentro de los cuales podemos clasificar la mayoría de las acciones que constituyen la intriga; a ellos habrá que añadirse también otra categoría que atañe a los motivos metanarrativos (manuscrito encontrado, falsa traducción, intervenciones del narrador, etc.). Siguiendo las pautas de Thompson y, más cerca a nuestro campo de investigación, de Luna Mariscal, se propondrá un esquema jerárquico que tenga en cuenta un nivel macroestructural (fábula) del que depende el nivel microestructural (motivos). En este marco general, mantenemos la intención declarada de limitar el registro de variantes de motivos a las más significativas, tanto por su recurrencia como por su valor sintagmático o paradigmático.

No pretendemos con este estudio introductorio proporcionar un cuadro detallado definitivo del catálogo que queremos crear sino, más bien, argumentar las premisas científicas y metodológicas bajo las que queremos construir nuestra base de datos, puesto que sin duda el avance del trabajo nos planteará nuevos interrogantes. Ni siquiera pretendemos afirmar que nuestra aproximación al estudio de los motivos y a la estructura de la base de datos que hemos esbozado no conlleven una serie de problemas a la vez que, supuestamente, solucionan otros. Por ejemplo, marcar el texto de los resúmenes con el etiquetado TEI para identificar los motivos nos plantea el problema de realizar una segmentación que no siempre es lineal: bajo el punto de vista del contenido, un motivo puede describir una porción de texto que se extiende a lo largo de varios capítulos, dificultando la inmediatez de la visualización en la pantalla. Análogamente, habrá que considerar los casos (frecuentes) en los que se presenten concatenadas aventuras distintas segmentadas, es decir, cuando

una aventura compuesta por más episodios menores se entrelace con otra aventura con la misma estructura narrativa: será necesario segmentar claramente el texto y señalar los motivos correspondientes teniendo en cuenta su aspecto combinatorio.

Sin embargo, el objetivo general de nuestro catálogo y de la organización que aquí se propone es el de mantener la conexión entre los motivos y los textos, también de manera visual, estableciendo un vínculo circular que une la base de datos, las fichas sobre los motivos y el corpus. Se mantiene, entonces, la necesidad de partir de los textos como los objetos principales que queremos describir, considerando los motivos como los instrumentos a través de los que conseguir este objetivo; por lo tanto, iremos trazando nuestra propia experiencia de análisis, intentando establecer un diálogo en una perspectiva continuadora con la tradición científica que hemos registrado, pero sin olvidar nunca el enfoque esencialmente pragmático con el que esta misma metodología ha nacido, que la hace especialmente flexible para acomodarse a ámbitos distintos, entre los cuales estamos convencidos de que podremos enumerar también los libros de caballerías italianos de tradición hispánica.

§

Bibliografía citada

Ardanuy Baró, Jordi, «MOMFER: una eina de cerca de motius folklòrics en el context de les humanitats digitals», *BiD: textos universitaris de biblioteconomia i documentació*, 36 (2016). DOI: <https://dx.doi.org/10.1344/BiD2016.36.21> (cons. 11/06/2021).

Barthes, Roland, «Introduzione all'analisi strutturale dei racconti», en *L'analisi del racconto*, Milano, Bompiani, 1969, pp. 5-47.

Bazzaco Stefano, «El Progetto Mambrino y las tecnologías OCR: estado de la cuestión», *Historias Fingidas*, 6 (2018), pp. 257-272.

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*, 9 (2020), pp. 534-561.

Birkhan, Helmut, Karin Lichtblau y Christa Tuczay (eds.), *Motif-Index of German Secular Narratives from the Beginning to 1400*, 7 vols., Berlin-New York, Walter de Gruyter-Austrian Academy of Sciences, 2005-2010.

Bognolo, Anna, «El Proyecto Mambrino: para una Base de Datos de motivos caballerescos», en *Siglo de Oro: nuevas perspectivas,* Actas XII Congreso AISO, Neuchâtel, 2-6 noviembre de 2020, Kassel, Reichenberger, en prensa.

Bognolo, Anna, Giovanni Cara y Stefano Neri, *Repertorio delle continuazioni italiane ai romanzi cavallereschi spagnoli. Vol. 1. Ciclo di Amadis di Gaula*, Roma, Bulzoni, 2013.

Bognolo, Anna y Stefano Bazzaco, «Tra Spagna e Italia: per l'edizione digitale del Progetto Mambrino», *eHumanista/IVITRA*, 16, 2019, pp. 20-36.

Bognolo Anna, Stefano Neri, Paola Bellomi y Federica Zoppi, *Repertorio delle continuazioni italiane ai romanzi cavallereschi spagnoli. Vol. 2. Ciclo di Palmerino di Oliva*, Roma, Bulzoni, en prensa.

Brémond, Claude, «A Critique of the Motif», en *French Literary Theory Today: A Reader*, ed. Tsvetan Todorov, Cambridge, Cambridge University Press, 1982, pp. 125-146.

—, «Comment concevoir un index des motifs», *Le Bulletin du Groupe de Recherches sémio-linguistiques*, 16 (1980), pp. 15-29.

—, «Sobre la noción de motivo en el relato», *La crisis de la literariedad*, ed. Miguel Angel Garrido Gallardo, Madrid, Taurus, 1987, pp. 115-124.

—, «Vers un Index des actions narratives», *Crisol, Typologie des formes narratives brèves au Moyen Âge (domaine roman)*, II/4 (2000), pp. 243-250.

Bueno Serrano, Ana Carmen, «Aproximación al estudio de los motivos literarios en los libros de caballerías castellanos (1508-1516)», en *De la literatura caballeresca al Quijote*, eds. Ana Carmen Bueno Serrano, Patricia Esteban Erlés, Karla Xiomara Luna Mariscal, Zaragoza, Prensas Universitarias de Zaragoza, 2007a, pp. 95-113.

—, *Índice y estudio de motivos en los libros de caballerías castellanos (1508-1516)*, tesis de doctorado, dir. Juan Manuel Cacho Blecua, Universidad de Zaragoza, Filología Hispánica (Literaturas Española e Hispánicas), 2007b, 4 vols.

Cacho Blecua, Juan Manuel, «El 'Motif-Index' de S. Thompson y sus aplicaciones en la literatura caballeresca», *Historias Fingidas*, 8 (2020), pp. 5-54. DOI: <https://doi.org/10.13136/2284-2667/729> (cons. 10/05/2022).

—, «El motivo en la literatura caballeresca. Presentación», *Revista de poética medieval*, 26 (2012), pp. 11-30.

—, «Introducción al estudio de los motivos en los libros de caballerías: la memoria de Román Ramírez», en *Libros de caballerías (de Amadís al Quijote). Poética, lectura, representación e identidad*, Salamanca, SEMyR, 2002, pp. 27-53.

Courtés, Joseph, «La "lettre" dans le conte populaire merveilleux français. Contribution à l'étude des motifs», *Documents de Recherche du Groupe de Recherches Sémio-Linguistiques de l'Institut de la Langue Française*, 9 (1979a), pp. 1-44.

—, «La "lettre" dans le conte populaire merveilleux français (seconde partie). Contribution à l'étude des motifs», *Documents de Recherche du Groupe de Recherches Sémio-Linguistiques de l'Institut de la Langue Française*, 10 (1979b), pp. 3-27.

—, «La"lettre" dans le conte populaire merveilleux français. Contribution à l'étude des motifs (Troisième partie)», *Documents de Recherche du Groupe de Recherches Sémio-Linguistiques de l'Institut de la Langue Française*, 14 (1980a), pp. 4-32

—, «Le motif selon S. Thompson», *Le Bulletin du Groupe de Recherches sémio-linguistiques (EHESS) -Institut de la Langue Française (CNRS), «Le Motif en Ethno-Littérature»*, 16 (1980b), pp. 3-14.

—, «Le motif, unité narrative et/ou culturelle?», *Le Bulletin du Groupe de Recherches sémio-linguistiques-Institut de la Langue Française (Le motif en ethno-litterature),* 16 (1980c), pp. 44-54.

—, «Motif et Type dans la tradition folklorique. Problèmes de typologie», *Littérature*, 45 (1982), pp. 114-127.

Díez-Borque, José María, *El libro: de la tradición oral a la cultura impresa*, Barcelona, Montesinos, 1995.

DQ = Cervantes, Miguel de, *Don Quijote de la Mancha*, ed. Francisco Rico, Instituo Cervantes, Barcelona, Crítica, 1998. URL: <https://cvc.cervantes.es/literatura/clasicos/quijote/> (cons. 11/06/2021).

González, Aurelio, «El concepto de motivo: unidad narrativa en el Romancero y otros textos tradicionales», en *Propuestas teórico-metodológicas para el estudio de la literatura hispánica medieval*, ed. Lilian von der Walde Moheno, México, Universidad Nacional Autónoma de México-Universidad Autónoma Metropolitana, 2003, pp. 353-384.

—, *El motivo como unidad narrativa a la luz del romancero tradicional*, México, El Colegio de México, 1990.

—, «El motivo: unidad narrativa en los romances caballerescos», *Revista de poética medieval*, 26 (2012), pp. 129-147.

Greimas, A. J., «Elementi per una teoria dell'interpretazione del racconto mitico», en *L'analisi del racconto*, Milano, Bompiani, 1969, pp. 47-95.

Karsdorp, Folgert, Marten Van der Meulen, Theo Meder y Antal Van den Bosch, «MOMFER: A Search Engine of Thompson's *Motif-Index of Folk Literature*», *Folklore*, 126/1 (2015), pp. 37-52.

Lucía Megías, José Manuel, «Dos caballeros en combate: batalla y lides singulares en *La leyenda del caballero del Cisne* y el *Libro del caballero Zifar*», en *La literatura en la época de Sancho IV. Actas del Congreso Internacional* (21-24 febrero 1994), Alcalá de Henares, Universidad de Alcalá/Servicio de Publicaciones, 1996, pp. 427-452.

—, «Los libros de caballerías en la floresta digital: aventuras jamás contadas ni imaginadas», *Historias Fingidas*, 7 (2019), pp. 5-34. DOI: <https://doi.org/10.13136/2284-2667/151> (cons. 10/05/2022).

Luna Mariscal, Karla Xiomara, «De la metodología o la pragmática del motivo en el índice de motivos de las historias caballerescas breves», *eHumanista*, 16 (2010), pp. 127-135.

—, «De Stith Thompson a las plataformas digitales: algunas reflexiones (con un Índice de motivos de la *Demanda del Santo Grial*, Toledo, 1515)», *Historias Fingidas*, 8 (2020), pp. 55-128. DOI: <https://doi.org/10.13136/2284-2667/164> (cons. 10/05/2022).

—, *El motivo literario en «El Baladro del sabio Merlín» (1498 y 1535). Con un Índice de motivos de «El Baladro del sabio Merlín» (Burgos, 1498 y Sevilla, 1535)*, México, El Colegio de México, 2017.
—, *Índice de motivos de las historias caballerescas breves*, Vigo, Academia del Hispanismo, 2013.
Pini, Donatella, Carmen Castillo Peña, «Cervantes. Elogio de la imperfección», *Artifara*, 13 bis: *Extraordinario monográfico: Las Novelas ejemplares en su IV centenario* (2013), pp. 265-284.
Propp, Vladimir, *Morfologia della fiaba*, Torino, Einaudi, 1966.
Sarmati, Elisabetta, *Le critiche ai libri di cavalleria nel Cinquecento spagnolo (con uno sguardo sul seicento). Un'analisi testuale*, Pisa, Giardini Editori, 1996.
Segre, Cesare, «Analisi del racconto, logica narrativa e tempo», en *Le strutture e il tempo*, Torino, Einaudi, 1974, pp. 3-77.
—, «Tema/motivo», en *Avviamento all'analisi del testo letterario*, Torino, Einaudi, 1985, pp. 331-359.
Todorov, Tzvetan, «Le categorie del racconto letterario», en *L'analisi del racconto*, Milano, Bompiani, 1969, pp. 227-270.
Tomasi, Giulia, «Las Humanidades Digitales y la base de datos MeMoRam: para un enfoque sistemático hacia los motivos en los libros de caballerías», *Historias Fingidas*, 8 (2020), pp. 129-156. DOI: <https://doi.org/10.13136/2284-2667/155> (cons. 10/05/2022).
Uther, Hans Jörg, «Classifying tales: remarks to Indexes and Systems of Ordering», *Narodna Umjetnost: Croatian Journal of Ethnology and Folklore Research*, 46/1 (2009), pp. 15-32.
—, *The Types of International Folktales. A Classification and Bibliography based on the System of Antti Aarne and Stith Thompson*, 3 vols., Helsinki, Academia Scientiarum Fennica (FF Communications, 285), 2004.
—, «Type-and Motif-Indices 1980-1995: An Inventory», *Asian Folklore Studies*, 55, 2 (1996), pp. 299-317.
Vázquez Recio, Nieves, *Una «yerva enconada»: sobre el concepto de «motivo» en el romancero tradicional*, Cádiz, Servicio de Publicaciones de la Universidad-Fundación Machado, 2000.
Zoppi, Federica, «Aproximación al estudio de los motivos cómicos en los libros de caballerías: unos ejemplos de los *Palmerines* italianos», *Historias Fingidas*, 7 (2019), pp. 313-340.

PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Realización de una base de datos de los motivos caballerescos: presentación y avances de MeMoRam

Giulia Tomasi

(Università di Trento)*

Abstract

En el artículo se presenta un proyecto nacional italiano dividido en cuatro unidades que abarcan la literatura caballeresca hispánica desde distintas perspectivas. Se trata de *Mapping Chivalry: Spanish Romances of Chivalry from Renaissance to 21st Century: a Digital Approach*. Tras su presentación general, en el artículo se hace referencia especialmente a la unidad que se está desarrollando en la Università di Trento, que se dedica a la realización de una base de datos centrada en los motivos de los libros de caballerías castellanos de los siglos XVI y XVII, que constituyen las fuentes de las demás unidades que conforman el proyecto. Se mencionan las tareas llevadas a cabo sobre el corpus y se apuntan los actuales objetivos de nuestras investigaciones: definir un índice de motivos con la ayuda de las herramientas digitales, sin olvidar el enfoque crítico tradicional. Los resultados obtenidos se ofrecerán a los usuarios en la base de datos MeMoRam.
Palabras clave: libros de caballerías; base de datos; índice de motivos; Humanidades Digitales

The article presents an Italian national project divided into four units covering Hispanic Chivalric Literature from different perspectives. It is *Mapping Chivalry: Spanish Romances of Chivalry from Renaissance to 21st Century: a Digital Approach*. After its general presentation, the article refers in particular to the unit being developed at the Università di Trento, which is dedicated to the creation of a database focusing on the motifs of 16th and 17th century Castilian romances of chivalry, which constitute the sources for the other units that make up the project. We mention the tasks carried out on the corpus and point to the current objectives of our research: to define an index of motifs with the help of digital tools, without forgetting the traditional critical approach. The results obtained will be made available to users in the MeMoRam database.
Keywords: Romances of Chivalry; database; Motif-Index; Digital Humanities

§

* Este trabajo se ha realizado en el marco del Proyecto de Investigación PRIN 2017 «Mapping Chivalry: Spanish Romances of Chivalry from Renaissance to 21th Century: a Digital approach», concedido por el MIUR.

El proyecto *Mapping Chivalry: Spanish Romances of Chivalry from Renaissance to 21st Century: a Digital Approach* obtuvo una financiación trienal por el Ministero dell'Università e la Ricerca en el ámbito de los proyectos de interés nacional (PRIN). En el equipo se integran cuatro unidades de investigación de distintas universidades italianas, que abarcan el tema de la literatura caballeresca desde perspectivas que se diferencian bajo el aspecto geográfico, genérico y diacrónico, compartiendo el fin de ofrecer a la comunidad académica, y no solo, unas herramientas de libre acceso para el estudio de dicha literatura[1]. En la unidad de la *Università di Trento*, bajo la dirección de Claudia Demattè, tomamos en consideración los libros de caballerías castellanos de los siglos XVI y XVII, con el objetivo de identificar, estudiar y coleccionar sus motivos y personajes en la base de datos MeMoRam Tomasi (2020b); el grupo de la *Università di Verona*, encabezado por Anna Bognolo, estudia las traducciones de los libros de caballerías y sus *aggiunte* en Italia, junto con su fortuna europea, a través de la realización de una biblioteca digital del ciclo italiano de *Amadís de Gaula* (Bognolo y Bazzaco, 2019); la *Università di Salerno*, con Daniele Crivellari como líder de su unidad, está preparando una base de datos donde se recoge información sobre el teatro caballeresco (TeatroCaballeresco) partiendo del repertorio de Demattè (2005); y el objeto de la *Università di Roma La Sapienza*, cuya unidad es dirigida por Elisabetta Sarmati, es elaborar la base de datos *AmadísSigloXX* para el estudio del fenómeno de las reescrituras de obras caballerescas en la época contemporánea.

El lugar que ocupa la base de datos MeMoRam en el marco de la unidad de Trento es fundamental, ya que nos dedicamos a las obras que constituyen las fuentes de las demás elaboraciones caballerescas objeto de estudio de las otras unidades. Así pues, la primera fase del flujo de trabajo que nos planteamos fue centrada en la definición del corpus, que se basó

[1] En este ámbito se señalan también la página web Corpus of Hispanic Chivalric Romances, dirigida por Ivy Corfis de la University of Winsconsin-Madison https://textred.spanport.lss.wisc.edu/chivalric/texts.html y el reciente proyecto *Universo de Almourol. Base de dados da Matéria Cavaleiresca Portuguesa*, desarrollado en la Universidade do Porto y dirigido por Aurelio Vargas Díaz-Toledo https://parnaseo.uv.es/UniversoDeAlmourol/ (Vargas Díaz-Toledo, 2019).

en varios catálogos de libros de caballerías[2]. La lista completa cuenta, finalmente, con 82 títulos entre obras originales, traducciones y reescrituras. Se organizaron luego unas fichas, donde se recogen sus metadatos en la plataforma *Muruca Bibliography*, una herramienta creada por el equipo de Net7, la empresa informática con la que colaboramos para la realización de las distintas bases de datos que conforman el proyecto completo[3]. Dedicamos varios meses a la elaboración de la ficha-tipo, decidiendo y descartando los campos relativos a los textos. El resultado final consta de varios campos donde se recoge, en primer lugar, un código diferenciador para cada texto mediante el que es posible crear vínculos con las demás bases de datos y unidades de investigación. A partir del corpus caballeresco, que todas las unidades comparten como fuente, resulta establecida, por ejemplo, la asociación entre el texto en lengua original (y su código) y la traducción italiana estudiada en Verona; se encuentran luego noticias eminentemente bibliográficas, como el año y el lugar de la primera edición a nuestro alcance y el autor, cuando lo conocemos; se apunta la naturaleza del texto, es decir, si se trata de un impreso, o bien de un manuscrito, ya que esta es una de las características que diferencia los paradigmas de evolución de los libros de caballerías (Lucía Megías, 2002, 25-32); asimismo, se detalla la pertenencia de las obras a un ciclo, siendo otro rasgo fundamental de la poética del género[4]; en fin, para cada obra se encuentra un enlace a la plataforma *Amadís* de la base de datos bibliográfica *Clarisel* de la Universidad de Zaragoza, donde se coleccionan los acercamientos críticos que se centran en los libros de caballerías.

[2] A partir de Gayangos (1963), hasta llegar a los más recientes y ya canónicos catálogos de Eisenberg y Marín Pina (2000) y Lucía Megías (2002 y 2019b), los estudiosos proporcionaron distintas clasificaciones para las obras.

[3] Net7 es una empresa italiana líder en las Humanidades Digitales. Su equipo realizó distintos proyectos de ámbito hispánico, como Calderón Digital, la Biblioteca Teatral Gondomar y Casa di Lope en el marco de TeSpa (Investigadora Principal: Fausta Antonucci): <http://tespasiglodeoro.it> (cons. 10/05/2022), y la base de datos AUTESO, que forma parte del proyecto de Theatheor (Investigadora Principal: Sònia Boadas): <http://theatheor-fe.netseven.it> (cons. 10/05/2022).

[4] Sobre este aspecto véanse Gutiérrez Trápaga (2017), Hinrichs (2017), Ramos Nogales (2017) y Sales Dasí (2002 y 2017). Demattè (en prensa), propone una nueva clasificación de las obras y sus ciclos. En mayo de 2021 se celebraron en Ciudad de México las II Jornadas de Literatura Caballeresca, *Estructura, poética y género de los ciclos caballerescos* en el ámbito del Seminario de Estudios sobre Narrativa Caballeresca.

De forma paralela, realizamos MeMoRam, la base de datos donde se encuentran referencias a las primeras ediciones a nuestro alcance de las obras del corpus y a sus ediciones modernas. Cabe mencionar el convenio que tenemos con la Universidad de Alcalá, gracias al que podemos contar con el permiso de utilizar las ediciones de las obras publicadas en la colección Libros de Rocinante[5], que cuenta actualmente con cuarenta libros de caballerías publicados desde los años noventa. Además de los datos editoriales coleccionados, en la plataforma tenemos un campo donde podemos cargar los textos en formato XML, para investigar en su contenido a través de herramientas de minería de datos y de textos. Una de las primeras tareas con la que nos enfrentamos fue la revisión de las ediciones modernas, que tuvimos que modificar de acuerdo a los criterios de redacción de los documentos en un lenguaje de marcado, para la que utilizamos Oxygen XML Editor. Ahora bien, tuvimos que intervenir en los textos en función del programa, eliminando cualquier elemento gráfico que pudiera obstaculizar la lectura por su parte. Por ejemplo, las normas editoriales de la colección de la Universidad de Alcalá imponen que las erratas de las obras originales se registren entre ángulos y esto invalida la redacción del texto en XML. Decidimos, pues, fijar el texto enmendado, sin necesidad de que fuese evidente la intervención del editor moderno, ya que nuestro objetivo es contar con un texto limpio de erratas para que pueda procesarse con las herramientas de *text mining*. Por la misma razón, otros símbolos que tuvimos que eliminar fueron los guiones y los números de folios que, en ocasiones, separan las sílabas. Tampoco la <&> es compatible con la redacción en XML de los textos, pues la sustituimos con <y>, o <e>, según los casos. Tras esta minuciosa labor de *restyling*, pudimos empezar a recopilar las obras utilizando una etiquetadura básica. Como nuestro objetivo es el de proporcionar la localización puntual de los motivos en el interior de las obras, solo marcamos las partes principales en que ellas se dividen, es decir, el título, los libros (o partes) y los capítulos, como se desprende en la figura 1:

[5] Agradecemos a Carlos Alvar y José Manuel Lucía Megías su disponibilidad en compartir estos textos y su apoyo a nuestro proyecto.

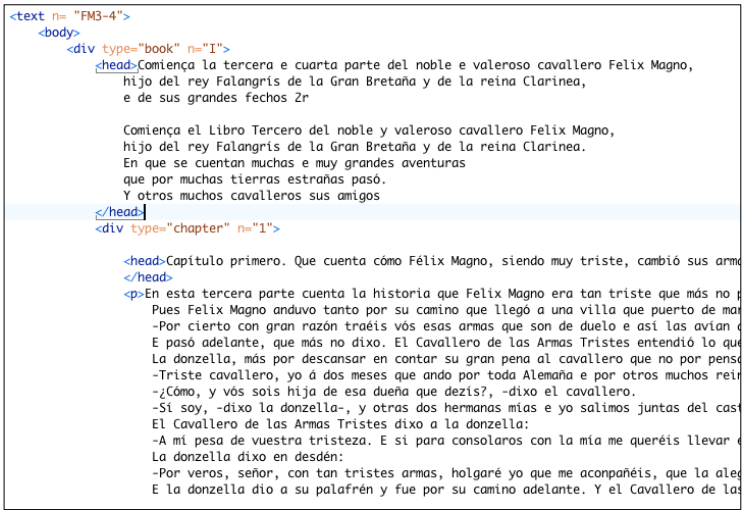


Fig. 1. Etiquetado realizado con el programa Oxygen XML Editor de *Félix Magno (III-IV)*, ed. de Claudia Demattè, Centro de Estudios Cervantinos, 2001.

El asunto principal de nuestras investigaciones es el motivo literario, que posibilita un enfoque privilegiado para orientar al lector en el rico contenido de estas obras de inmenso tamaño. Como señala de manera exhaustiva Luna Mariscal (2020), existe una multitud de índices dedicados a *corpora* distintos y todos se basan en el modelo inaugurado por Thompson en la elaboración de su monumental *Motif-Index of Folk Literature* (1975)[6]. Al detallar los problemas que plantea la creación de un índice de motivos literarios[7], la estudiosa subraya que, a pesar de los conocidos límites que conlleva, al fin y al cabo, toda investigación que apunte a la realización de un instrumento para localizar los motivos, tiene que seguir al sistema Thompson «no sólo por la inmensa cantidad de materiales que recoge, sino por haberse convertido en una obra de referencia internacional» (Luna Mariscal, 2020, 57). Pese al acierto de estas

[6] Especialmente interesantes resultan los índices dedicados a los diversos *corpora* caballerescos y artúricos, como, entre otros, Ruck (1991), Guerreau-Jalabert (1992), Birkhan (2005-2010), Luna Mariscal (2013, 2017 y 2020).

[7] Luna Mariscal evidencia que «el significado o los sentidos de un relato son resultado de un sistema de relaciones y en ningún caso poseen una naturaleza sustancial, la indexación del sentido es, entonces, por definición, imposible» (2020, 57) y añade que los índices son instrumentos de localización de motivos, que no pueden proporcionar resultados críticos sobre el sentido de los mismos. Véase también Luna Mariscal (2010) para un enfoque pragmático de la cuestión.

conclusiones, en nuestro trabajo sobre los motivos otra obra ocupa un lugar de referencia: es el *Índice y estudio de motivos en los libros de caballerías castellanos (1508-1516)* de Bueno Serrano[8], donde, además de los motivos identificados en el *Motif-Index*, se registran otros, más eminentemente caballerescos, que la autora elaboró *ad hoc* para el corpus seleccionado, es decir, los siete primeros libros de caballerías: *Amadís de Gaula*, *Las Sergas de Esplandián*, *Lisuarte de Grecia* (de Feliciano de Silva), *Palmerín de Olivia*, *Primaleón*, *Florisando* y *Floriseo*[9]. Esta tarea supuso la definición del motivo a nivel teórico y, a la vez, su formulación práctica. Como afirma la autora, se trata de expresar «por medios finitos […] un contenido que se realiza por medios infinitos» (Bueno Serrano, 2007, 148). Podemos, pues, identificar al motivo como un enunciado que define metaliterariamente las unidades literarias y que se formula a través de sustantivos deverbales[10]. Se distinguen, además, varios niveles de abstracción del motivo, que Bueno Serrano llama Nivel 1 (del sintagma, más concreto) y Nivel 2 (del paradigma, más abstracto)[11]. Ambos estratos resultan fundamentales, ya que el del paradigma descubre las constantes básicas de la acción narrativa, que se coloran mediante las múltiples combinaciones de los elementos que se aprecian en el nivel sintagmático, que garantiza la variedad en dichas constantes. El índice cuenta con cuatro volúmenes, en los que, además de una parte introductoria muy extensa y actualizada, se encuentran las páginas dedicadas a los motivos folclóricos en orden alfabético (Bueno Serrano, 2007, 641-938)[12]; el tercer volumen ofrece el orden alfabético-morfológico de los motivos sintagmáticos simples (Bueno Serrano, 2007, 1267-1929); las páginas siguientes van dedicadas a los motivos

[8] La tesis doctoral de la estudiosa, dirigida por Cacho Blecua, saca partido de los planteamientos del mismo acerca del motivo en los libros de caballerías (Cacho Blecua, 2002), sobre los que volvió a reflexionar en distintas ocasiones (Cacho Blecua, 2012; 2020).

[9] Contar con tal índice nos parece un avance no desdeñable, ya que, como afirma Antonucci acerca de los motivos en el teatro de Calderón, en los índices de motivos folclóricos a menudo se omiten aspectos fundamentales para ámbitos distintos del folclore, como el universo cortesano en nuestro caso (Antonucci, 2018, 87).

[10] En lugar de *Un niño es abandonado por su criado*, el motivo será: *Abandono de un niño por criado* (Bueno Serrano 2007, 153).

[11] Se distinguen también los motivos paradigmáticos compuestos: dos motivo de Nivel 2 que se combinan entre sí y llevan al Nivel 4.

[12] Esta parte del índice sigue al sistema de catalogación del *Motif-Index* de Thompson.

paradigmáticos simples (Bueno Serrano, 2007, 1935-2024) y, en fin, se recogen los motivos paradigmáticos compuestos (Bueno Serrano, 2007, 2133-2206). Para facilitar la consulta, además del orden alfabético, se proporciona siempre otra posibilidad de búsqueda por orden alfabético-morfológico (páginas 941-1262; 2024-2130 y 2209-2361 respectivamente). La articulación de este trabajo es muy compleja y los resultados sobre el contenido del corpus proceden de una lectura y un conocimiento profundos del mismo, ambos aspectos que nos facilitan datos rigurosos y detallados acerca de los motivos de las obras. Sin embargo, nos parece que no siempre el instrumento resulta asequible, debido a la necesidad de combinar manualmente los datos sacados de índices distintos, y a la excesiva especificidad de ciertos motivos, que tienden a excluir muchas realizaciones contextuales afines[13].

Por estas razones, pese a tener en el *Índice* de Bueno Serrano un precioso aliado del que no podemos prescindir en las investigaciones sobre un corpus tan extenso como el que nos ocupa, somos concientes de la necesidad de adecuarnos a un sistema de más amplio alcance, es decir, al modelo Thompson, para que el instrumento que queremos proporcionar sea lo más productivo posible desde la perspectiva del acceso al material textual y su uso en los estudios sobre la literatura caballeresca[14]. Es nuestro objetivo, entonces, combinar estas dos herramientas, para sacar de ambas sus mejores aportaciones: de los volúmenes de Bueno Serrano la naturaleza misma de las etiquetas, que logran perfectamente descubrir y describir la esencia de las obras de nuestro corpus; del índice de Thompson, queremos aprovechar el sistema de clasificación alfanumérico, su uso en el ámbito de otros géneros y literaturas —lo que posibilita la comparación— y las añadiduras de otros estudiosos, fruto de la aplicación del índice a los distintos *corpora*. Se vincula a este último aspecto también

[13] En el largo listado de motivos paradigmáticos simples encontramos *Amenaza de suicidio para obligar a aceptar un servicio caballeresco* (identificado en las *Sergas de Esplandián*, Rodríguez de Montalvo, 2002, 5, 15, 212). El grado de precisión del enunciado permite localizar puntualmente el acontecimeinto; sin embargo, nos parece que, al detectarse solo una vez, se contradice el principio de la recurrencia. Además, los matices que se engloban en el enunciado no tienen un nivel de abstracción que pueda adecuarse a un corpus diferente y más amplio. Por ello se requiere una revisión de las entradas identificadas por Bueno Serrano a la hora de ampliar el material narrativo a las que ellas irían refiriéndose.

[14] En efecto, no tendría sentido anhelar a realizar un instrumento de acceso amplio y abierto sin tener en cuenta la posibilidad de comparación que el sistema de clasificación de Thompson garantiza.

la necesidad de contar con un instrumento más ágil, que cuente con un número de entradas suficiente pero no exagerado y que sea inteligible también por parte de usuarios no expertos en la literatura de caballerías, con el fin de favorecer el enfoque multidisciplinar.

En lo que atañe a la revisión de las entradas propuestas por Bueno Serrano en su índice, la cuestión que nos parece relevante es definir qué grado de precisión y detalles debe considerarse adecuado en el ámbito de los libros de caballerías para reconocer a un motivo nuevo, frente a una posible variación. La misma autora reflexionó sobre el problema: «El enunciado de un motivo está en constante reformulación porque cada nuevo dato hace necesario valorar el conjunto y sus relaciones, y replantearse el grado de abstracción» (Bueno Serrano, 2007, 165), a lo que se añade el grado de pertenencia, es decir, la posibilidad de considerar los matices como parte de un nuevo motivo, susceptibles de añadirse al enunciado ya existente del mismo, o bien simplemente como variaciones de un motivo que puede expresarse mediante un enunciado más abstracto[15]. Pensamos que el enfoque computacional puede sustentar nuestro trabajo de interpretación de los elementos textuales mediante la aportación de datos estadísticos. En efecto, solo la recurrencia nos parece un valor adecuado para establecer qué matices deberían ser valorados en el enunciado del motivo y cuáles, en cambio, quedarían fuera del mismo por no repetirse un número suficiente de veces en el conjunto de las obras tomadas en cuenta. El acercamiento al corpus, cuya amplitud aumenta según su segmentación en motivos, se convierte en una tarea difícil de llevar a cabo sin la ayuda de las herramientas digitales.

Nuestras actuales investigaciones se están encaminando precisamente hacia la definición en plan abstracto de los motivos, su identificación, y la organización del material obtenido en la plataforma. Junto al enfoque basado en estudios críticos ya existentes[16], es nuestra

[15] Véase el ejemplo de *Amenaza de suicidio para obligar a aceptar un servicio caballersco* en nota 17. Pensamos que los estudiosos necesitan contar con un instrumento que les permita localizar en el corpus la frecuencia del motivo más general *Amenaza de suicidio*, para dar cuenta, mediante el tradicional acercamiento crítico a los textos, de los matices que se aprecian en sus distintas realizaciones contextuales.

[16] Antes de todo, el índice de Bueno Serrano en sus distintas declinaciones, pero también los demás índices caballerescos, los estudios de Cacho Blecua (2002; 2012), de Luna Mariscal (2010; 2013; 2017;

intención poner a prueba las potencialidades de las herramientas de *text mining* para la investigación literaria. Así pues, en la estela de trabajos recientes sobre distintos *corpora* literarios[17], tenemos por objetivo identificar en un grupo de obras determinadas las nubes de palabras que se crean en torno a motivos específicos y utilizarlas para definir y localizar los mismos en el corpus completo, corroborando o bien contradiciendo, al mismo tiempo, su validez y eficacia en gran escala[18]. Un rastreo que nos devuelva resultados estadísticos sobre los textos, nos otorgaría respuestas acerca del nivel de abstracción adecuado para describir de manera conveniente al universo de los libros de caballerías, puesto que nos ayudaría a descifrar fácilmente y con más inmediatez cuándo nos encontramos ante un nuevo motivo, o bien simplemente a un desvío de uno ya registrado.

En el caso del motivo *Mantenimiento de una mala costumbre*, identificado en veintisiete distintas ocasiones en el corpus que Bueno Serrano toma en cuenta (2007, 1990), conducimos una búsqueda a través de *Voyant Tools*[19] para averiguar si la nube de palabras extrapolada por el programa se correspondía con nuestras hipótesis sobre los términos (y, de forma más amplia, los campos semánticos) más comunes en el desarrollo de dicho motivo. Nuestras experimentaciones se centraron en cuatro obras: *Floriseo*, *Lisuarte de Grecia*, *Palmerín de Olivia* y *Primaleón* que manejamos en word para nuestros objetivos. De ellos analizamos ocho capítulos, es decir, los que presentan el motivo *Mantenimiento de una mala costumbre* según el rastreo de Bueno Serrano. El resultado es interesante: entre las palabras sacadas por el programa, encontramos hasta trece veces «presos», a la que añadimos

2018; 2020) y las Guías de Lectura Caballerescas de la Universidad de Alcalá, que nos ofrecen una perspectiva amplia sobre los principales argumentos tratados en las obras y sus personajes.

[17] Véanse al respecto los trabajos de Blevins (2010), Jockers-Mimno (2012) y Karsdorp-Van der Bosch (2013).

[18] Vamos a comprobar la utilidad de las herramientas mediante el material contenido en el índice de Bueno Serrano, en el que se encuentran las referencias puntuales a los motivos en los siete primeros libros de caballerías. A continuación se proporciona un ejemplo de cómo hemos llevado a cabo esta tarea.

[19] Se trata de una herramienta de libre acceso en la web, creada por Stéfan Sinclair y Geoffrey Rockwell, que permite analizar textos mediante técnicas de *distant reading*. Para nuestros objetivos actuales utilizamos la función «Cirrus» para visualizar las nubes de palabras a partir de los *corpora* que el usuario decida cargar en la plataforma.

«prisioneros», que aparece dos veces. Asociando las prácticas de *distant* y *close reading* se desprende que en diez ocasiones se trata precisamente de las víctimas de la mala costumbre mantenida por gigantes, caballeros malvados, o bien dueñas engañosas. Es más, la palabra «dueña» sale veintisiete veces según la nube de palabras (Fig. 2) y solo en una ocasión no se relaciona con el motivo analizado. Asimismo, «castillo» se conecta en todas las ocasiones en que aparece (treinta y cuatro) con la costumbre de aprisionar caballeros, mientras que no se encuentra esta palabra en un episodio en el que «presos» no se refiere a las víctimas de la mala costumbre[20]. Así pues, podemos concluir que el motivo paradigmático simple *Mantenimiento de una mala costumbre* en los cuatro libros de caballerías tomados en cuenta, suele referirse en concreto a la práctica de aprisionar caballeros en castillos, a menudo por parte de dueñas malvadas:

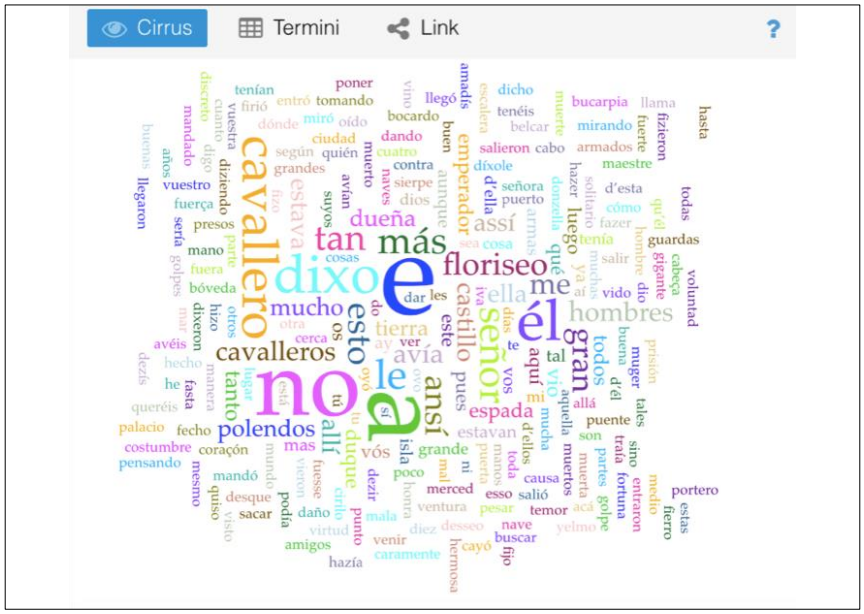


Fig. 2. Nube de palabras obtenida mediante la función Cirrus de Voyant Tools a partir de unos capítulos de los libros de caballerías *Floriseo*, *Lisuarte de Grecia*, *Palmerín de Olivia* y *Primaleón*

[20] La palabra «presos» en *Floriseo* se refiere a los enemigos del héroe epónimo, que este vence y aprisiona. En este episodio la mala costumbre no se mantiene en un castillo, sino que al pasar un puente para acceder a una ciudad, la compañía de Floriseo se topa con una guarda intimándole que los caballeros juren «que toda muger de diez años arriba que viéredes matéis en toda esta tierra» (Bernal, 2003, 1, 40, 72).

Completando la búsqueda a través de la ampliación del corpus a nuestro alcance, sería posible corroborar estos resultados que clasifican a los episodios analizados en la estela del primer libro de *Amadís de Gaula*, que se remonta al universo novelesco artúrico (Bognolo, 1996). Al mismo tiempo podría trazarse la evolución del motivo a medida que los libros de caballerías siguen difundiéndose y que, como afirma Bognolo, en el *Amadís de Gaula* actúa «como elemento connotativo de un género, rasgo que permite el reconocimiento del universo artúrico por parte del lector; pero se encuentra resemantizado en la nueva obra de Montalvo» (1996, 71). En el grado de resemantización adoptado por los autores de libros de caballerías reside la capacidad del género mismo de desarrollarse a medida que avanza el siglo.

Para dar cuenta de la complejidad del concepto de motivo, el diseño de la sección de la plataforma donde se coleccionarán las unidades contiene distintos campos: a través de un desplegable que ofrecerá un abanico de motivos a elegir, podrá accederse a las referencias puntuales a las obras (con sus divisiones en partes y capítulos) en las que el motivo buscado se desarrolla; a sus realizaciones en los distintos contextos, es decir, los fragmentos donde el motivo aparece; se encontrarán las posibles referencias a los homólogos del índice de Thompson[21]; se pondrá en evidencia tanto la relación con otros motivos de la lista, en el caso de secuencias muy estereotipadas, como las referencias a los personajes que actúan en el fragmento; el usuario tendrá a su alcance un enlace a la bibliografía crítica que tenga por objeto el motivo; y, en fin, se señalarán las posibles divergencias respecto al enunciado original del motivo, siempre y cuando estas afecten directamete a la trama, sin que puedan constituir un motivo nuevo[22]. Este último aspecto es crucial, ya que

[21] Al motivo *Destrucción de la propiedad*, por ejemplo, le corresponde: *Q595. Loss or destruction of property as punishment* en el *Motif-Index* de Thompson.

[22] Es el caso, por ejemplo, del motivo *Destrucción de la propiedad* que en *Valerián de Hungría* sufre un desvío que resulta singificativo en el relato, sin que dicho desvío se repita un número suficiente de veces como para conformar un motivo nuevo. En este texto, tras erradicar la mala costumbre de la maga Boralda, Valerián prende fuego a su castillo, que, sin embargo, no se quema del todo y, como descubrimos más adelante en la narración, la encantadora conserva en sus antiguas habitaciones un libro de conjuros que será fundamental para el hechizo de la princesa, que constituye el principal motor narrativo de la segunda parte de la obra de Dionís Clemente (2010, 2, 12, 522). La divergencia respecto al motivo original registrado se señala con un asterisco: *Destrucción de la propiedad *fallida*. Véase Tomasi (2020a, 103).

tenemos entre nuestros objetivos simplificar el índice proporcionado por Bueno Serrano, sin por ello renunciar a los desvíos que permiten corroborar la idea de que los libros de caballerías no son todos «una mesma cosa» (Cervantes, 2014, 1, 47, 618; Marín Pina, 1998)[23].

Al plantearnos los objetivos del proyecto y su actuación práctica, resonaron en nuestros oídos las palabras de Lucía Megías, según las cuales «las herramientas digitales solo pueden constituir la base de nuestro conocimiento futuro si se basan en los trabajos científicos anteriores» (2019a, 33): el desafío es favorecer el conocimiento a partir de la mucha información que tenemos a nuestro alcance. Así pues, para llegar a construir de manera eficaz una herramienta que pueda guiar a los usuarios en el inmenso mundo de los libros de caballerías, los datos sacados a la luz a través de instrumentos informáticos deben ser interpretados por estudiosos expertos en este género, que sepan elaborar etiquetas (el medio finito, metaliterario) vinculándolas a los fragmentos textuales en los que se desarrollan los motivos (los medios infinitos, literarios), ya que no siempre estos se expresan en el texto mediante las mismas palabras utilizadas en los marbetes que los describen y el conocimiento de los contenidos narrativos llega a ser, pues, imprescindible para realizar una indexación de este tipo[24].

Acercarse al estudio de los libros de caballerías mediante el soporte tecnológico no significa dejar de leer los textos y confiar todo el esfuerzo a las máquinas, sino respaldar los datos cuantitativos que ellas nos otorgan con la cualidad de unas conclusiones que solo pueden sacarse a través de un enfoque más tradicional. Para cumplir con este importante desafío de las Humanidades Digitales, tenemos que emprender un trabajo circular:

[23] En lo que atañe al grado de abstracción elegido para elaborar, mantener y descartar los enunciados, confluyen distintos aspectos, entre otros la «singularidad textual» (Luna Mariscal, 2010, 130) que un motivo, gracias a sus características específicas, le otorga a una obra. Sin embargo, tampoco debe olvidarse el principio de recurrencia y nos parece que el registro de las divergencias es un instrumento útil para marcar la singularidad frente a un motivo recurrente y mantener un equilibrio entre estos dos aspectos fundamentales.

[24] En el caso del motivo *Mantenimiento de una mala costumbre* sí es posible leer en todos los fragmentos los términos *mala* y *costumbre*, lo que permite identificar fácilmente al motivo. Pero piénsese en el motivo de la *Hospitalidad*, que suele expresarse mediante referencias a la mesa, la comida, el servicio, los atavíos, etc. como elementos siempre reconocibles, que, sin embargo, se remontan a la *Hospitalidad* solo a través de una interpretación connotativa.

precisamos vincular las sólidas bases críticas que manejamos sobre el género y las teorías sobre el motivo literario con los resultados estadísticos obtenidos a través de métodos computacionales para llegar a nuestro objetivo, es decir, ampliar el conocimiento crítico del corpus. En esta óptica, el mayor conocimiento futuro se alcanzaría mediante la interpretación de unos datos (sacados gracias a las tecnologías), que no puede prescindir de cierta familiaridad previa con los libros de caballerías, la cual solo es posible si volvemos a fijarnos en las mencionadas bases críticas que desde siempre nos guían en su universo.

§

Bibliografía citada

Antonucci, Fausta «Una nueva herramienta para el estudio del teatro clásico español: Calderón Digital. Base de datos, argumentos y motivos del teatro de Calderón», *Bulletin of the Comediantes*, 70/1 (2018), pp. 79-95.

Bernal, Fernando, *Floriseo*, ed. Javier Guijarro Ceballos, Alcalá de Henares, Centro de Estudios Cervantinos, 2003.

Birkhan, Helmut (dir.), Karin Lichtblau and Christa Tuczay (eds.), *Motif-Index of German Secular Narratives from the Beginning to 1400*, 7 vols., Berlin-New York: Walter de Gruyter-Austrian Academy of Sciences, 2005-2010.

Blevins, Cameron «Topic Modeling Martha Ballard's Diary», 2010. URL: <https://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/> (cons. 05/04/2022).

Bognolo, Anna, «La desmitificación del espacio en el Amadís de Gaula los "castillos de la mala costumbre"», en *Studia aurea: Actas del III Congreso de la AISO (Toulouse, 1993)*, coords. Ignacio Arellano Ayuso, Carmen Pinillos Salvador, Marc Vitse, Frédéric Serralta, vol. 3, 1996.

Bognolo, Anna y Stefano Bazzaco, «Tra Spagna e Italia: per l'edizione digitale del Progetto Mambrino», *eHumanista/IVITRA*, 16 (2019), pp. 20-36.

Bueno Serrano, Ana Carmen, *Índice y Estudio de Motivos en los Libros de Caballerías Castellanos (1508-1516)*, Tesis de Doctorado, dir. Juan Manuel Cacho Blecua, Universidad de Zaragoza, Filología Española (Literaturas Española e Hispánica), 2007.

Cacho Blecua, Juan Manuel, «Introducción al estudio de los motivos en los libros de caballerías: la memoria de Román Ramírez», en *Libros de caballerías, (de «Amadís» al «Quijote»). Poética, lectura e identidad*, eds. Eva Belén Carro Carbajal, Laura Puerto Moro, María Sánchez Pérez, Salamanca, Seminario de Estudios Medievales y Renacentistas, Sociedad de Estudios Medievales y Renacentistas, 2002, pp. 27-57.

— (ed.), *Revista de Poética Medieval. El motivo en la literatura sapiencial*, 26, 2012.

—, «El 'Motif-Index' de S. Thompson y sus aplicaciones en la literatura caballeresca», *Historias Fingidas*, 8 (2020), pp. 5-54. DOI: <https://doi.org/10.13136/2284-2667/729> (cons. 10/05/2022).

Cervantes, Miguel de, *Don Quijote de la Mancha*, ed. John Jay Allen, Madrid, Cátedra, 2014, 2 vols.

Clarisel: *Amadís Base de datos de literatura caballeresca*. URL: <https://clarisel.unizar.es/paginas/index.php?base=amadis&opcion=presentacion> (cons. 05/04/2022).

Clemente, Dionís, *Valerián de Hungría*, ed. Jesús Duce García, Alcalá de Henares, Centro de Estudios Cervantinos, 2010.

Corpus of Hispanic Chivalric Romances. URL: https://textred.spanport.lss.wisc.edu/chivalric/index%20english.html (cons. 05/04/2022).

Demattè, Claudia, *Repertorio bibliografico e studio interpretativo del teatro cavalleresco spagnolo del sec. XVII*, Trento, Editrice Università degli Studi di Trento, 2005.

—, «Ciclos de caballerías hispánicos *versus* libros de caballerías 'únicos' y 'sueltos': una propuesta de nueva clasificación», en *Libros de caballerías: aproximaciones a la poética de un género literario*, eds. Daniel Gutiérrez Trápaga y María Gutiérrez Padilla, Ciudad de México, Facultad de Filosofía y Letras, UNAM, en prensa.

Eisenberg, Daniel y María Carmen Marín Pina, *Bibliografía de los libros de caballerías castellanos*, Zaragoza, Prensas Universitarias, 2000.

Félix Magno (III-IV), ed. Claudia Demattè, Alcalá de Henares, Centro de Estudios Cervantinos, 2001.

Gayangos, Pascual de, *Libros de caballerías. Con un discurso preliminar y un catálogo razonado por don Pascual de Gayangos*, BAE, 40 Madrid, Atlas, 1963.

Guerreau-Jalabert, Anita, *Index des motif narratifs dans le romans arthuriens français en vers (XII-XIII siècles)*, Geneve, Droz, 1992.

Gutiérrez Trápaga, Daniel, *Rewritings, Sequels and Cycles in Sixteenth-Century Castilian Romances of Chivalry*, Woodbridge, Tamesis, 2017.

Hinrichs, William, «La novela y la secuela. De cómo la prosa narrativa del Siglo de Oro inventó la continuación literaria» en *La escritura inacabada. Continuaciones literarias y creación en España. Siglos XIII a XVII*, Madrid Casa de Velázquez, 2017, pp. 19-29. URL: <http://books.openedition.org/cvz/3324> (cons. 05/04/2022).

Jockers, Matthew L. y David Mimno, «Significant Themes in 19th-Century Literature», *Faculty Publications*, *Department of English*, 105 (2012). URL: <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1105&context=englishfacpubs> (cons. 05/04/2022).

Karsdorp, Folgert y Antal Van der Bosch, «Identifying Motifs in Folktales using Topic Models», en *Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning*, Nijmegen, 2013, pp. 41-49.

Lucía Megías, José Manuel, «Libros de caballerías castellanos: textos y contextos», *Edad de Oro XXI*, 2002, pp. 9-61.

—, «Los libros de caballerías en la floresta digital: aventuras jamás contadas ni imaginadas», *Historias Fingidas*, 7 (2019a), pp. 5-34.

—, «Género literario, corpus y difusión de los libros de caballerías castellanos», en *Libros de caballerías castellanos. 2. Género literario, corpus y difusión*, dir. Marta Haro Cortés, Monografías *Aula Medieval*, 9 (2019b), pp. 5-47.

Luna Mariscal, Karla Xiomara, «De la metodología o la pragmática del motivo en el índice de motivos de las historias caballerescas breves», *eHumanista*, 16 (2010), pp. 127-135.

—, *Índice de motivos de las historias caballerescas breves*, Vigo, Editorial Academia del Hispanismo, 2013.

—, *El motivo literario en «El Baladro del Sabio Merlín» (1498 y 1535)*, México, El Colegio de México, 2017.

—, «El motivo y los libros de caballerías», *Lingüistíca y Literatura*, 2018, pp. 78-90. URL: <https://revistas.udea.edu.co/index.php/lyl/article/view/334474> (cons. 05/04/2022).

—, «De Stith Thompson a las plataformas digitales: algunas reflexiones (con un Índice de motivos de la *Demanda del Santo Grial*, Toledo, 1515)», *Historias Fingidas*, 8 (2020), pp. 55-128. DOI: <https://doi.org/10.13136/2284-2667/164> (cons. 10/05/2022).

Marín Pina, María Carmen, «Motivos y tópicos caballerescos», en *Don Quijote de La Mancha*, ed. Francisco Rico, Barcelona, Crítica, 1998, pp. 857-902.

Palmerín de Olivia, ed. Giuseppe di Stefano, intr. María Carmen Marín Pina, Alcalá de Henares, Centro de Estudios Cervantinos, 2004.

Primaleón, ed. María Carmen Marín Pina, Alcalá de Henares, Centro de Estudios Cervantinos, 1998.

Ramos Nogales, Rafael, «Las continuaciones y la configuración genérica de los libros de caballerías», en *La escritura inacabada. Continuaciones literarias y creación en España. Siglos XIII a XVII*, Madrid Casa de Velázquez, 2017, pp. 121-143.

Ruck, Elaine Heather, *An index of themes and motifs in 12th century French Arthurian poetry*, Arthurian studies 25, Cambridge, D. S. Brewer, 1991.

Sales Dasí, Emilio José, (2002), «Las continuaciones heterodoxas (el *Florisando* [1510] de Páez de Ribera y el *Lisuarte de Grecia* [1526] de Juan Díaz) y ortodoxas (el *Lisuarte de Grecia* [1514] y el *Amadís de Grecia* [1530] de Feliciano de Silva) del *Amadís de Gaula*», *Edad de Oro XXI*, pp. 117-152.

—, «¿Continuador o creador? "Las entricadas razones del famoso Feliciano de Silva"», en *La escritura inacabada. Continuaciones literarias y creación en España. Siglos XIII a XVII*, Madrid Casa de Velázquez, 2017, pp. 145-161. URL: <https://books.openedition.org/cvz/3333#bodyftn43> (cons. 05/04/2022).

Rodríguez de Montalvo, Garci, *Sergas de Esplandián*, ed. Carlos Sainz de la Maza, Madrid, Castalia, 2002.

Silva, Feliciano de, *Lisuarte de Grecia*, ed. Emilio J. Sales Dasí, Alcalá de Henares, Centro de Estudios Cervantinos, 2002.

TeSpa Siglo de Oro. URL: <http://tespasiglodeoro.it> (cons. 05/04/2022).

Theateor. URL: <http://theatheor-fe.netseven.it> (cons. 05/04/2022).

Thompson, Stith (1975), *Motif-Index of Folk Literature*, Bloomington and London, Indiana University Press.

Tomasi, Giulia, «Hacia un repertorio de personajes divergentes y motivos caballerescos: unas notas sobre *Valerián de Hungría* y *Cirongilio de Tracia*», en *En línea caballeresca. Lecciones del Seminario de Estudios sobre Narrativa Caballeresca*, eds. Axayácatl Campos García Rojas, Yordi Enrique Gutiérrez Barreto, 2020a, pp. 89-111. URL: <http://ru.atheneadigital.filos.unam.mx/jspui/handle/FFYL_UNAM/3493> (cons. 5/04/2022).

—, «Las Humanidades Digitales y la base de datos MeMo-Ram: para un enfoque sistemático hacia los motivos en los libros de caballerías», *Historias Fingidas*, 8 (2020b), pp. 129-156. DOI: <https://doi.org/10.13136/2284-2667/155> (cons. 10/05/2022).

Universo de Almourol. Base de dados da Matéria Cavaleiresca Portuguesa. URL: <https://parnaseo.uv.es/UniversoDeAlmourol/> (cons. 05/04/2022).

Vargas Díaz-Toledo, Aurelio, «Universo de Almourol. Base de dados da Matéria Cavaleiresca Portuguesa», *Historias Fingidas*, 7 (2019), pp. 459-461. DOI: <https://doi.org/10.13136/2284-2667/148> (cons. 01/11/2021).

Voyant Tools. URL: <https://voyant-tools.org/?lang=es> (cons. 01/11/2021).