Tecnologie OCR e *libros de caballerías:* un test sul *Florando de Inglaterra*[1]

Stefano Neri
Università degli Studi di Verona

Questo contributo si propone di esporre i risultati di una serie di test effettuati mediante la piattaforma di trascrizione automatica Transkribus su di un romanzo cavalleresco spagnolo del Cinquecento, il *Florando de Inglaterra*[2]. L'opera fu stampata a Lisbona da German Gallarde nel 1545 e non si ha notizia di edizioni successive. Il testo è molto esteso (256 carte in formato *in folio*) ed omogeneo, composto su doppia colonna con un carattere gotico appartenente a una famiglia piuttosto comune nella stampa spagnola e portoghese del Cinquecento, in particolare nei testi appartenenti al genere dei *libros de caballerías*[3]. Uno dei cinque esemplari esistenti del *Florando*, quello della British Library (C.62.h.14), oltre ad essere completo e in ottime condizioni di conservazione, è stato recentemente oggetto di una riproduzione fotografica digitale di buonissima qualità e messo a disposizione in rete sia nelle collezioni digitali della British Library, che su Google Books. L'appartenenza del *Florando* ad uno standard editoriale diffuso, la sua estensione, la disponibilità e la buona qualità della riproduzione digitale ne fanno un campo di prova ideale su cui testare un software di trascrizione automatica.

Sulla riproduzione fotografica del *Florando de Inglaterra* della British Library sono stati creati tre modelli di riconoscimento automatico dei caratteri. Tali modelli si distinguono tra loro per le regole di trascrizione imposte: il primo è decisamente modernizzante, mentre gli altri due sono più conservativi. Per la loro creazione si è seguito il procedimento standard previsto dalla piattaforma Transkribus, ossia la segmentazione delle aree di testo e la trascrizione manuale della quantità di caratteri utile alla macchina per allenare un modello di trascrizione automatica (*training set*). In quanto alla segmentazione si è utilizzato il modello di *layout analysis* P2PaLA (Page to Page Layout Analysis) per il riconoscimento della doppia colonna:

[1] Questo lavoro si colloca nell'ambito del PRIN 2017 Mapping Chivalry. Spanish Romances of Chivalry from Renaissance to XXI century: a Digital Approach (prot. 2017JA5XAR) e in particolare dell'Unità di Verona, il Progetto Mambrino; e del Progetto di eccellenza: Le Digital Humanities applicate alle lingue e letterature straniere (2018-2022) del Dipartimento di Lingue e letterature Straniere dell'Università di Verona

[2] Transkribus è una piattaforma per la digitalizzazione, il riconoscimento del testo basato su tecnologie di Intelligenza Artificiale, la trascrizione e la ricerca di documenti storici gestita dalla Società cooperativa Europea READ-COOP e sviluppata dall'Università di Innsbruck (<https://readcoop.eu/it/transkribus/> [28/12/2021]). Si veda il recente articolo di Mühlberger et al. (2019).

[3] Il *Florando de Inglaterra* è un romanzo cavalleresco ancora poco studiato del quale sto preparando l'edizione moderna per la collana "Los libros de Rocinante" dell'Instituto Universitario de Investigación Miguel de Cervantes. Gli unici approcci recenti all'opera sono quelli di Cristina Castillo Martínez (2001, 2002, 2005).

si tratta di uno strumento integrato in Transkribus che permette di riconoscere la struttura materiale del documento, individuando regioni di testo, colonne, righe. Tale strumento è stato configurato mediante un modello specifico per il riconoscimento della doppia colonna, creato da Stefano Bazzaco per testi dalle caratteristiche tipografiche simili al *Florando*. Per il *training set* dei primi due modelli sono state trascritte manualmente 25 pagine collocate all'inizio del testo (cc. 1r-13r corrispondenti a 2229 righe tipografiche e 17837 parole). Per il terzo modello il *traning set* è stato più esteso, in quanto sviluppato su di un maggior numero di opere stampate con tipografia gotica, come si vedrà più avanti. Una volta creati i modelli, per ognuno di questi è stato eseguito un test di trascrizione automatica, lanciato su 4 pagine del *Florando de Inglaterra* scelte casualmente: le cc. 42r-43v, corrispondenti a 14127 caratteri distribuiti su 376 linee tipografiche.

Modello 1 - Modernizzante

Questo modello è stato creato allo scopo di allestire un'edizione cartacea senza apparati critici come quelle della collana "Los libros de Rocinante" dell'Instituto Universitario de Investigación Miguel de Cervantes (ex Centro de Estudios Cervantinos). I criteri di trascrizione sono interpretativi e modernizzanti. In sintesi:

-Si sciolgono le abbreviazioni

-Si modernizzano i caratteri e i segni tipografici attualmente non in uso

-Si modernizza la separazione delle parole secondo l'uso attuale

-Si inserisce l'accentuazione secondo la norma attuale

-Si regolarizza l'uso delle maiuscole

-Si segnalano i dialoghi come discorso diretto (: -)

-Si distingue u,v,b secondo il valore vocalico o consonantico

-Si regolarizza y/i secondo l'uso moderno

-Si modernizza l'uso del gruppo qu- davani a a,o,u (*quando>cuando, qual>cual)*;

-Si modernizza l'uso della rr (*rrazón>razón*, *honrra>honra*);

-Si modernizzano in gruppi colti *ph >f* , *th >t, ch > c*, *ct>t (Christo > Cristo*)

La punteggiatura viene attualizzata secondo le norme vigenti e in funzione interpretativa, cercando, in ogni caso, di tener conto della punteggiatura del testo base.

Si mantengono le alternanze tra s/ss, j/x, m/n nasali davanti a p/b, y/e congiunzione.

Uno di questi criteri, l'attualizzazione della punteggiatura, è stato escluso dalla trascrizione manuale realizzata per il *training set* di Transkribus in modo da poter-

ne avere il controllo interpretativo: la macchina trascriverà i segni di interpunzione esattamente come appaiono nella riproduzione fotodigitale affinché l'editore possa interpretarli e inserirli manualmente secondo le norme attuali.

Il modello di riconoscimento dei caratteri che è stato creato ha un Character Error Rate (CER) stimato al 2,13% (*validation set*). Questo significa che la macchina calcola di non essere riuscita a riconoscere circa 2,13 caratteri su 100 (spazi inclusi), pertanto, dei circa 3500 caratteri che compongono una pagina dell'originale, nella trascrizione ce ne saranno 70 di sbagliati, cioè quasi uno (0.74) per riga (considerando la doppia colonna).

A questa percentuale andranno aggiunti gli interventi collegati all'inserimento dell'interpunzione (compreso l'inserimento delle maiuscole dopo il punto), nonché altri tre tipi di interventi che il calcolo del CER non considera:

-interventi di spaziatura: nella composizione a stampa gli spazi tipografici sono spesso irregolari, per cui l'unione o la separazione di parole saranno una delle tipologie di intervento più frequente.

-interventi di accentuazione: nel testo originale non è presente l'accentuazione e quindi, nonostante l'allenamento, è probabile che la macchina non riesca a introdurla automaticamente, ad esempio nel caso della distinzione tra *él* pronome e *el* articolo (fig.1).

-interventi sulle maiuscole: nella composizione tipografica le maiuscole sono più rare e meno omogenee. È probabile che il programma riesca a riconoscere in modo corretto solamente le maiuscole presenti nel *training set* (fig. 2)

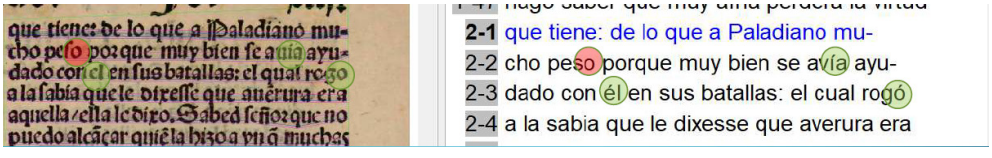


Fig.1 Riconoscimento automatico degli accenti

Fig. 2. Errore nel riconoscimento delle maiuscole (B>D)

Dopo aver esportato il file della trascrizione automatica in un programma di videoscrittura, sono stati eseguiti gli interventi di correzione manuale. Il confronto tra la versione di output e la versione con correzione manuale ha evidenziato la presenza delle tipologie di interventi previste, alle quali si è aggiunta un'ulteriore tipologia di interventi su errori di trascrizione originatisi durante la fase di *layout analysis*, cioè

Fig. 3. *Layout analysis:* problemi di riconoscimento delle aree di testo in corrispondenza di xilografie e (sotto) interventi manuali

l'individuazione automatizzata dello specchio di scrittura e delle righe. Questo procedimento fallisce e causa errori di trascrizione in corrispondenza di xilografie o di iniziali xilografiche (tipicamente all'inizio dei capitoli): nel caso specifico, in prossimità della xilografia e dell'iniziale illustrata che appaiono nella colonna A della c. 42v all'inizio del capitolo 28, la macchina non riesce ad individuare con precisione i confini del testo e quindi la trascrizione risulta lacunosa nelle zone adiacenti alle xilografie, compresa la colonna B (fig. 3).

Sull'output di trascrizione automatica delle cc. 42r-43v generato con questo modello sono stati necessari 520 interventi, per un CER effettivo del 3,53%, corrispondente a 1,38 errori per riga. Ciò significa che lavorando con una impostazione standard di videoscrittura che preveda righe di circa 100 caratteri (spazi inclusi), l'editore moderno dovrà mettere in conto circa tre interventi (3,45) per ogni riga di testo trascritto.

Va puntualizzato, tuttavia, che la trascrizione automatica sarà tanto più soggetta a errori quanto più modernizzanti e interpretativi saranno i criteri imposti. Questo modello, pertanto, è prevedibilmente più impreciso rispetto a modelli più conservativi. Il

servizio reso all'editore moderno rimane comunque molto apprezzabile, a tratti sorprendente; per lo scioglimento delle abbreviazioni il software è molto efficace, così come per la distinzione tra u/v/b; gli interventi sugli accenti in queste quattro pagine sono stati 85, ma va sottolineato che in 181 occasioni il software è riuscito ad accentuare correttamente modo automatico. Gli interventi di punteggiatura, come detto, andrebbero esclusi dal calcolo di affidabilità di Transkribus: nel testo originale ci sono 188 segni di interpunzione e su quasi tutti è stato fatto un intervento. Se li escludessimo dal calcolo, il CER si ridurrebbe al 2,25%, cioè di poco superiore a quello stimato dalla piattaforma per questo modello (2,13%).

Modello 2 – Semidiplomatico

Sullo stesso campione testuale corrispondente alle cc. 42r-43v del Florando, applicando lo stesso modello di layout analysis realizzato con P2PaLa, è stato creato un modello di riconoscimento di tipo conservativo. Il training set è costituito da una trascrizione manuale realizzata con criteri conservativi, che vanno verso un'edizione semidiplomatica. Gli interventi interpretativi previsti da questo modello sono molto limitati:

- si regolarizza la separazione di parole e gli "a capo"

- si regolarizza l'uso delle maiuscole (es. nomi propri)

- si correggono errori materiali evidenti (per inversione dei tipi, capovolgimenti, ecc)

Si mantine:

- la punteggiatura e tutti i segni diacritici

- le abbreviazioni

- le oscillazioni irregolari tra u/v – y/e – x/j

- l'uso di diversi caratteri tipografici per identici grafemi (es. d, r, s) e per le legature

I criteri di trascrizione più conservativi imposti al modello hanno prodotto una trascrizione che richiede un numero di interventi decisamente minore rispetto al modello modernizzante. Transkribus ha conservato tutte le abbreviazioni, tutti i caratteri tipografici dell'originale, tutti i segni diacritici e di interpunzione (fig. 4). Gli interventi di revisione riguardano prevalentemente l'unione o separazione di parole e qualche mancato riconoscimento dei segni di abbreviazione. Anche qui, inoltre, si è dovuto intervenire per sanare errori causati dal *layout analysis* in corrispondenza della xilografia e dell'iniziale ornata che appaiono all'inizio del capitolo 28. Si sono inoltre ripresentati gli stessi errori del modello modernizzante nel riconoscimento dei caratteri tipografici maiuscoli. Il CER del modello (*validation set*) previsto dalla piattaforma è del 1.35%. Gli interventi manuali di revisione sono stati 164 e hanno portato il tasso di errori effettivo al 1,16%, un valore addirittura inferiore a quello

previsto dal software. Gli errori per riga di testo tipografico sono stati 0,43, che in un foglio di videoscrittura calibrato su 100 caratteri per riga (spazi inclusi) si tradurrebbe in circa un errore per ogni riga (1,09).

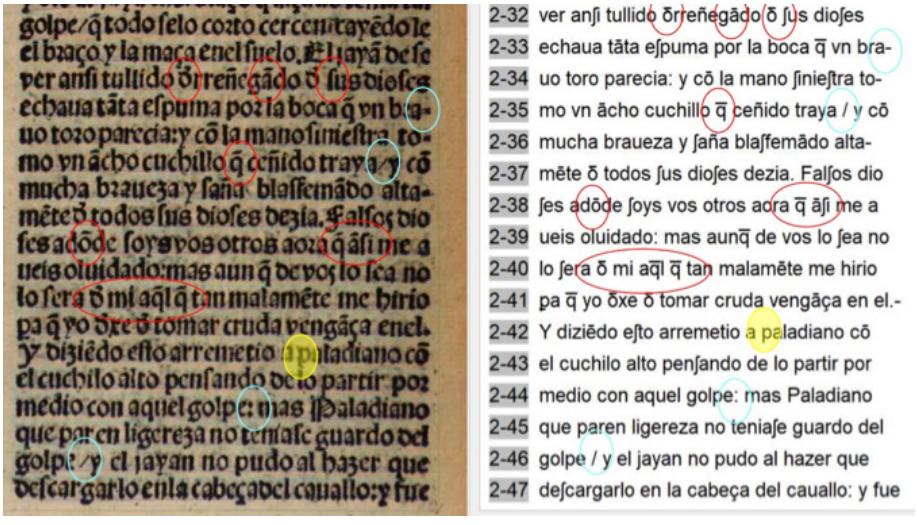


Fig. 4. *Output* di trascrizione automatica con il modello 2 semidipomatico

Modello 3 – Spanishgothic_XV-XVI_Extended

L'ultimo modello è stato creato da Stefano Bazzaco su di un corpus ragguardevole di edizioni spagnole del XVI e XVII secolo con caratteristiche materiali simili al *Florando de Inglaterra* (Bazzaco: 2020[4]) . Il *training set* su cui è stato allenato il software differisce, quindi, dai due modelli precedenti, basati unicamente sul *Florando,* essendo composto da maggior numero di testi (15), righe di testo (16816) e caratteri (150137): ognuno di questi valori è praticamente 10 volte più alto rispetto agli altri

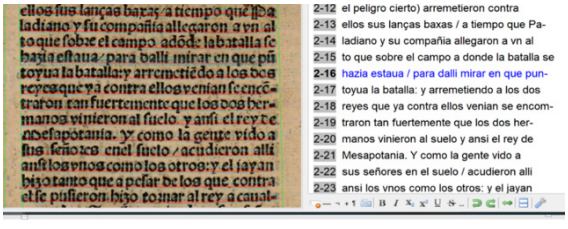
[4] Il modello è stato creato all'interno della attività di ricerca inerenti alla creazione della biblioteca digitale del Progetto Mambrino, finanziate dal progetto PRIN 2017 – Progetti di Ricerca di rilevante interesse nazionale del Ministero dell'Università e della Ricerca (MUR), prot. 2017JA5XAR, Mapping Chivalry. Spanish Romances of Chivalry from Renaissance to 21th century: a Digital Approach (P.I. Anna Bognolo). Sulla stessa tematica, sempre all'interno delle attività del Progetto Mambrino si veda anche Bazzaco-Bognolo (2019), Bazzaco (2018) e Mancinelli (2016). Il modello è disponibile ad accesso aperto nella sezione Public Models di Transkribus e una sua descrizione è disponibile nel sito della piattaforma < https://readcoop.eu/it/model/spanish-gothic-15th-16th-century/ > (20/12/2021). Una descrizione più estesa può essere consultata al seguente link GitHub: < https://github.com/stefanobazzaco/HTR-model-SpanishGothic_XV-XVI_extended > (28/12/2021) e l'uso del modello è sotteso al rispetto del copyright così come dichiarato nella corrispondente pagina del repository Zenodo < https://zenodo.org/record/4888927#.Yc122GjMI2x > (20/12/2021).

due modelli.

I criteri di trascrizione sono di tipo conservativo, anche se si introducono alcuni elementi di interpretazione che si discostano dal modello precedente. In particolare:

- si sciolgono le abbreviazioni
- si normalizzano i caratteri tipografici secondo l'uso moderno
- si normalizza la separazione di parole
- si mantengono gli "a capo"
- si mantiene l'uso delle maiuscole
- si mantengono gli errori materiali evidenti

Il modello di riconoscimento automatico creato e allenato da Stefano Bazzaco è stato usato sullo stesso campione testuale (42r-43v), dopo aver effettuato il medesimo procedimento di *layout analysis*. Transkribus ha riconosciuto e sciolto efficacemente quasi tutte le abbreviazioni, ha riconosciuto la maggior parte delle separazioni di parole, ha trasformato correttamente i caratteri tipografici e ha conservato i segni di interpunzione. Gli interventi manuali realizzati sulla trascrizione automatica generata da Transkribus sono stati pochi e hanno riguardato soprattutto la separazione di parole, lo scioglimento di alcune abbreviazioni e il controllo dei passaggi in cui si sono verificati problemi di *layout analysis* in corrispondenza della xilografia, come nei due modelli precedenti (fig. 5). Il CER calcolato da Transkribus è 0,92% (*validation set*). Gli interventi manuali sono stati solo 82, pari ad un CER dello 0,57%, un valore sensibilmente inferiore al previsto che, ipotizzando l'uso di un programma di videoscrittura impostato sui 100 caratteri per riga (spazi inclusi), per l'editore moderno si tradurrebbe in un intervento ogni due righe (0,54 per riga).



~~taua~~acertaua golpe a cauallero que no matasse / lo que viendo los aquileos empeçaron a desmayar algun tanto: y conociendo esto sus enemigos los seguian de tal manera que ya muy destroç~~a~~ados andauan: que ni valia bozes de los hermanos ni cosa alguna que les dixessen para hazerles cobrar animo: y viendo ellos que de todo esto eran causadores: el rey de Mesapotania / y el de Mentapolin: como aquellos que mucho estimauan su ~~honrra~~caun~~honrra~~ (aun que en ello era el peligro cierto) arremetieron contra ellos sus lanças baxas / a tiempo que Paladiano y su compañia allegaron a vn ~~al to~~alto que sobre el campo ~~a donde~~adonde la batalla se hazia estaua / para dalli mirar en que ~~puntoyua~~punto yua la batalla: y arremetiendo a los dos reyes que ya contra ellos venian se enco~~m~~ntraron tan fuertemente que los dos hermanos vinieron al suelo y ansi el rey de Mesapotania. Y como la gente vido a sus señores en el suelo / acudieron alli ansi los vnos como los otros: y el jayan

Fig. 5. *Output* di trascrizione automatica con il modello 3 Spanishgothic_XV-XVI_Extended e (sotto) interventi di correzione manuale.

Conclusioni

Il test ha confermato che la trascrizione automatica diventa tanto più affidabile quanto più conservativi sono i criteri imposti e quanto più estesa è la base del *training set.* Dei nostri tre modelli, infatti, quello che dà risultati migliori in termini di "pulizia" della trascrizione è senza dubbio il terzo, Spanishgothic_XV-XVI_Extended, per il suo carattere conservativo e per la maggior consistenza del corpus sul quale il software è stato allenato (Tab. 1).

La scelta tra un modello e un altro, tuttavia, dipende soprattutto dall'uso a cui è destinata la trascrizione.

Se lo scopo fosse quello di allestire un'edizione scientifica digitale, la scelta ricadrebbe sicuramente su uno dei due modelli conservativi. Il modello Spanishgothic_XV-XVI_Extended, in particolare, è stato creato espressamente a questo scopo, la realizzazione di edizioni scientifiche digitali che prevedono, da un lato la trascrizione conservativa del testo-fonte, dall'altro la possibilità di marcarlo con diverse tipologie di *tags* semantici e strutturali in linguaggio XML TEI. Sul testo trascritto automaticamente si avrà, ad esempio, la possibilità di marcare un errore e proporre un emendamento, di esporre le varianti, ma anche di registrare la struttura del documento (in capitoli, pagine, righe), segnalare la presenza di toponimi e antroponimi, inserire note esplicative, eccetera.

Nel caso si volesse allestire un'edizione critica in cartaceo, la scelta ricadrebbe comunque sul modello Spanishgothic_XV-XVI_Extended, che garantisce la possibilità di intervenire manualmente su di un testo molto affidabile, mantenendo traccia della lezione originale. Il secondo modello può essere scelto se si ha lo scopo di preparare un'edizione semidiplomatica. Inoltre, essendo particolarmente conservativo anche rispetto al disegno dei caratteri tipografici, esso può essere usato per ricercare varianti di emissione o di stato in copie diverse della stessa edizione. Nel caso specifico, tutti e cinque gli esemplari dell'unica edizione del *Florando* si potrebbero trascrivere automaticamente e poi confrontare in maniera automatica in modo da individuare i luoghi in cui è probabile siano avvenuti cambiamenti nella composizione tipografica durante la tiratura. Questo suppone, tuttavia, di disporre di riproduzioni di analoga qualità per tutti gli esemplari: un'evenienza ancora lontana dalla realtà, purtroppo, non solo per il *Florando de Inglaterra*.

Con l'obiettivo, infine, di utilizzare la trascrizione generata da Transkribus per allestire un'edizione cartacea non critica e con criteri di trascrizione modernizzanti sul modello della collana "Los libros de Rocinante", il primo modello creato è quello che presenta maggiori vantaggi. La trascrizione generata da questo modello modernizzante fornisce una buona affidabilità, ma l'intervento dell'editore dovrà essere attento e analitico sia per far fronte agli errori generati dal processo di trascrizione automatica (2,25%) che per introdurre la punteggiatura interpretativa sulla base dei segni diacritici che sono stati conservati. Usando gli altri due modelli per allesti-

re un'edizione con queste caratteristiche, gli interventi dell'editore sarebbero molto maggiori, pari al 4,34% per il modello Spanishgothic_XV-XVI_Extended, e del 5,26% per il modello semidiplomatico.

MODELLO	(1) MODERNIZZANTE	(2) SEMIDIPLOMATICO	(3) SPANISHGOTHIC_ XV-XVI_EXTENDED
CER su Validation set %	2,13	1,35	0,92
CER post correzione %	3,53	1,16	0,57
Inteventi totali cc. 42r-43v	520	164	82
Interventi per riga (originale)	1,38	0,43	0,21
Interventi per riga (100 car.)	3,45	1,09	0,54

Tab. 1 – Confronto tra i tre modelli

Bibliografia citata

- Bazzaco, Stefano (2020), "El reconocimiento automático de textos en letra gótica del Siglo de Oro: creación de un modelo HTR basado en libros de caballerías del siglo XVI en la plataforma Transkribus", *Janus,* 9: 534-561 [28/12/2021] < https://www.janusdigital.es/articulo.htm?id=160> .

— (2018), "El Progetto Mambrino y las tecnologías OCR: estado de la cuestión", *Historias Fingidas,* 6: 257-272 [28/12/2021] <https://historiasfingidas.dlls.univr.it/article/view/89>

- Bognolo, Anna; Bazzaco, Stefano (2019) "Tra Spagna e Italia: per l'edizione digitale del Progetto Mambrino", eHumanista/IVITRA, 16: 20-36 [28/12/2021] < https://www.ehumanista.ucsb.edu/sites/default/files/sitefiles/ivitra/volume16/3.%20Bognolo.pdf>.

- Castillo Martínez, Cristina (2001), *Florando de Inglaterra (Partes I-II). Guía de lectura caballeresca,* Alcalá de Henares, Centro de Estudios Cervantinos.

— (2002), "Algunas consideraciones acerca del Florando de Inglaterra (1545)", *Edad de oro,* 21, 367-374.

— (2005), *Florando de Inglaterra (Parte III). Guía de lectura caballeresca,* Alcalá de Henares, Centro de Estudios Cervantinos, 2005.

- Mancinelli, Tiziana (2016), "Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work", *Historias Fingidas,* 4: 255-260 [28/12/2021] <https://historiasfingidas.dlls.univr.it/article/view/65>.

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