La trascrizione automatica di documenti a stampa antichi. Appunti per un modello di riconoscimento della tipografia in corsivo

«Dig*Italia*» 1-2024
DOI: 10.36181/digitalia-00094

Stefano Bazzaco
Università di Verona

Il contributo intende descrivere il flusso di lavoro che ha condotto alla produzione di un modello di Handwritten Text Recognition (HTR) per la trascrizione automatica di testi veneziani del Cinquecento in corsivo. Nella prima parte, si definisce l'ambito di studio, con particolare attenzione allo stato dell'arte e ai recenti sviluppi nel campo dell'HTR per quanto riguarda i complex scripts, vale a dire testi a stampa antichi e manoscritti che per le loro caratteristiche ostacolano l'applicazione dei tradizionali sistemi di Optical Character Recognition (OCR). Nella seconda parte, si espongono le principali fasi del lavoro di addestramento della macchina per la creazione del modello Italics_VeniceXVIs, che costituisce un primo passo per l'interpretazione dei testi cavallereschi in carattere corsivo di interesse del Progetto Mambrino dell'Università di Verona. Infine, si individuano le principali caratteristiche del modello e, in un'ottica di accessibilità e riutilizzo, si segnalano i passi futuri del progetto, suggerendo possibili ricadute della ricerca svolta in relazione con altri ambiti di studio.

Introduzione

Negli ultimi anni il processo di digitalizzazione del patrimonio bibliografico sta sperimentando degli avanzamenti notevoli, che a loro volta si ripercuotono sui metodi e le pratiche degli addetti ai lavori, vale a dire umanisti, bibliotecari e in generale esperti della gestione dell'informazione. Tra i vari strumenti digitali che si impiegano per la conversione, rimediazione e analisi di materiali analogici testuali, stanno destando particolare attenzione le applicazioni che sfruttano l'intelligenza artificiale per svolgere compiti che richiederebbero altrimenti il dispendio di un'enorme quantità di tempo e di forza lavoro da parte degli specialisti. Si tratta di sistemi di *machine learning* che per risolvere particolari compiti integrano alla potenza di calcolo della macchina l'apprendimento su base di dati fornita dall'utente, aprendo la via a possibilità predittive che promettono di avere profonde ricadute sul trattamento stesso dei dati informativi e, conseguentemente, della conoscenza. Basti considerare, per esempio, la recente elaborazione di tecnologie come i

Large Language Models (LLM), modelli linguistici che, appoggiandosi a reti neurali per svolgere funzioni generative anche molto complesse, stanno scuotendo le fondamenta della ricerca e della didattica perché impongono di ripensare il patrimonio culturale come una sorta di *big data*, suscettibile di analisi quantitative per l'estrapolazione, la produzione e la valorizzazione automatica dell'informazione[1].
Precedentemente allo sviluppo dei LLM, l'ambito di ricerca maggiormente esposto alle nuove proposte dell'intelligenza artificiale è stato quello della trascrizione automatica, un campo di studi da troppo tempo imbrigliato dalle limitazioni e i *bias* dell'insufficienza tecnologica. Nello specifico, l'introduzione di reti neurali ricorrenti, vale a dire processi che simulano il funzionamento della mente umana ai fini di addestrare la macchina a svolgere funzioni complesse[2], ha avuto un notevole impatto sulle applicazioni di Automatic Text Recognition (ATR), modificandone le prestazioni e suggerendone un maggior impiego su larga scala. Grazie alla preparazione e distribuzione di modelli di riconoscimento di testo altamente affidabili, infatti, si assiste a una proliferazione di progetti che mirano a rendere accessibile e processabile il patrimonio testuale globale avvalendosi di strumenti di trascrizione automatica che consentono di trasformare il testo contenuto nelle immagini in formato elettronico (*machine readable*).
Tra le tendenze più significative che si registrano nel contesto della ATR, si segnala il vertiginoso sviluppo di sistemi di Handwritten Text Recognition (HTR), che hanno rapidamente soppiantato in termini di impiego i più famosi, ma datati, sistemi di Optical Character Recognition (OCR). Questa sostituzione ha in qualche modo alimentato l'erronea opinione che l'HTR sia l'evoluzione naturale dell'OCR, quando invece si tratta di tecnologie in parte distinte, visto che la più antica, l'OCR, si basa sul riconoscimento delle realizzazioni grafiche di singoli caratteri, mentre la più recente, l'HTR, cerca di individuare *pattern* ricorrenti di segni grafici in una linea di testo, affidandosi alla predizione dei risultati di trascrizione grazie alle informazioni contestuali e all'addestramento preliminare fornito dall'utente.
L'allenamento della macchina, che si avvale di processi di *machine learning* come le reti neuronali ricorrenti (RNN), costituisce la principale caratteristica dei sistemi di HTR ed ha favorito un significativo miglioramento dei risultati di trascrizione automatica, sia nel caso di testi manoscritti che di testi a stampa. Di conseguenza, la riflessione circa l'impiego di questi strumenti per la digitalizzazione del patrimonio culturale testuale si sta spostando verso problematiche specifiche, come la riduzione del lavoro umano nella creazione del materiale di addestramento (*training set*) e il riuso di dati e modelli già esistenti.

[1] Nell'attualità esistono diversi modelli linguistici generativi: il più comune e diffuso, per ragioni di distribuzione e risonanza mediatica, è ChatGPT sviluppato da Open AI, ma altri LLM sono stati resi pubblici negli ultimi mesi, come Bard (Google), Claude 2 (Anthropic) e LLAMA (Meta). Per uno studio approfondito delle caratteristiche e delle possibilità che potrebbero offrire queste applicazioni alla ricerca umanistica si vedano Ciotti 2023; Roncaglia 2023.
[2] Schwartz-Ricci 2022, p. 172.

Il presente studio si colloca in questo contesto e mira a descrivere le principali scelte che hanno governato il progetto di creazione di un modello di HTR per l'interpretazione dei caratteri in corsivo del Rinascimento. Inizialmente, si propone un'indagine storica dei sistemi di HTR, ponendo l'accento sullo stato dell'arte e il flusso di lavoro. In secondo luogo, si presentano le piattaforme disponibili e si definiscono le fasi progettuali che caratterizzano la creazione di un modello di riconoscimento, con riferimento in particolare al riutilizzo dei dati, allo sviluppo di modelli HTR misti e all'operazione di *fine tuning* di modelli preesistenti. Segue quindi una descrizione delle principali attività nel campo della trascrizione automatica intraprese dal gruppo di ricerca Progetto Mambrino dell'Università di Verona nell'ambito del Progetto PRIN "Mapping Chivalry" (2018-2022), del Progetto di Eccellenza "Inclusive Humanities" (2023-2027) e del Centro dipartimentale DAIH (Digital Arena for Inclusive Humanities). Infine, si chiariscono le caratteristiche dei modelli di HTR creati per l'interpretazione del corsivo, suggerendo possibili prospettive di ricerca future che possano favorirne il perfezionamento e la distribuzione su larga scala.

Trascrizione automatica e sistemi di HTR: breve analisi storica

La trascrizione automatica è un campo di studi che sta destando interesse grazie all'impressionante incremento delle prestazioni di calcolo che offrono le nuove tecnologie. In particolare, i sistemi di HTR, impiegati in larga scala per la conversione di intere collezioni digitalizzate in testo elettronico, hanno sperimentato un notevole sviluppo negli ultimi anni, qualificandosi come uno dei settori in maggior espansione nell'ambito delle Digital Humanities.

In origine, con il termine HTR si fa riferimento a strumenti che a partire dall'interpretazione di un'immagine digitalizzata di un manoscritto consentono di ottenere come *output* la trascrizione del contenuto del documento in formato testo elettronico. Per questa ragione, spesso ci si riferisce a queste tecnologie con la denominazione di "OCR for manuscripts", utilizzando un'espressione che risente di un certo grado di inesattezza perché accomuna due sistemi di riconoscimento distinti e al contempo limita l'applicabilità di entrambi a contesti predeterminati. E se, da una parte, è evidente che le rispettive storie dei sistemi di OCR e HTR si intreccino e sovrappongano, con il risultato che spesso non ci siano differenze sostanziali rispetto ai metodi e procedimenti alla base del loro utilizzo (Stokes e Kiessling 2024), dall'altro vanno chiarite le specificità dei secondi per spiegare l'interesse che stanno suscitando e il loro impiego sempre più diffuso.

La prima differenza è legata alle modalità di interpretazione dei segni grafici presenti nella fonte. Al riguardo, si noterà come già a livello terminologico gli strumenti di OCR si basano sulla decodificazione di caratteri isolati (Optical *Character* Recognition), mentre quelli di HTR su porzioni di testo (Handwritten *Text* Recognition), vale a dire linee o *patterns* di segni grafici contigui. Come naturale

conseguenza, si tende a considerare che i primi siano adatti all'interpretazione di testi a stampa, vista la loro predisposizione a trattare fenomeni unitari e discreti; al contrario, i secondi risponderebbero all'esigenza di trascrivere testimoni manoscritti, in cui lettere e parole tendono a essere collegate tra loro in modo asistematico, dipendendo da vari fattori, come le specificità della grafia dell'autore o le circostanze di scrittura. Questa distinzione però risulta immotivata, perché i sistemi di HTR vengono impiegati anche per la trascrizione automatica di testi a stampa antichi (ad esempio un libro in caratteri gotici) o di difficile interpretazione (ad esempio i giornali tedeschi dei primi del Novecento in Fraktur), raggiungendo un grado di affidabilità impensabile per i sistemi di OCR coetanei.

In secondo luogo, seppur gli sviluppi nel campo dell'OCR interessino diverse tecnologie gemelle, la storia dell'HTR si articola in modo parzialmente indipendente. Tra il 1965 e il 1973, quando si stavano già diffondendo i sistemi hardware di OCR di seconda generazione[3], la trascrizione automatica dei testi manoscritti iniziava a muovere i primi passi e occupare le riflessioni di Sayre. Lo studioso segnalava le difficoltà legate all'interpretazione di scritture continue e proponeva come soluzione tecnica di regolarizzare i caratteri che compongono una parola, con l'intento di semplificare la segmentazione e il posteriore riconoscimento con tecniche di *template matching*[4]. Queste teorie trovarono scarsa applicazione, e bisognerà attendere sino agli anni '80 perché i sistemi di riconoscimento di testo vengano per la prima volta impiegati per identificare brevi scritture manoscritte, come codici postali e sequenze di assegni bancari, vale a dire materiali specifici dove la segmentazione era normalmente agevolata dalla disposizione ben separata dei caratteri.

Il quadro rimaneva poi essenzialmente immutato per un decennio, con la ricerca che si orientava verso lo sviluppo di dispositivi hardware per il riconoscimento gestuale e della grafia per sostituire tastiera e mouse nell'interazione uomo-macchina: questo campo, comunemente riconosciuto come *pen computing*, in quanto implica l'uso di una penna digitale e di una tavoletta grafica per l'elaborazione del segnale analogico in ingresso, si sviluppò a partire dal 1983, quando l'azienda Pencept elaborò prototipi come il PenPad 200 e il PenPad 320, con il secondo già integrato alle applicazioni MS DOS dei primi personal computer di IBM.

In seguito, tra gli anni '80 e '90, si diffusero i primi sistemi software per il riconoscimento di testo manoscritto, ma il limite principale di questi strumenti continuava ad essere quello di seguire un approccio *segmentation based*, dove l'interpreta-

[3] Bazzaco et al. 2022, p. 73-74.

[4] La questione di cui si interessa lo studioso prende il nome di "paradosso di Sayre" e si appoggia sulla seguente concezione: una parola manoscritta per essere interpretata deve essere suddivisa nei caratteri che la compongono, ma, inevitabilmente, per essere segmentata, la stessa dovrà essere prima interpretata. Come risoluzione del problema, lo studioso propone di suddividere in frammenti una linea di testo continua (segmentazione implicita) tentando di individuare una corrispondenza su base probabilistica tra un set di parole regolarizzato e gli stessi frammenti. Sull'argomento si vedano: Sayre 1973 e Vinciarelli 2003.

zione passava per l'isolamento dei caratteri e la loro successiva interpretazione per sovrapposizione con un set predefinito di segni grafici. Questa disposizione presupponeva il ricorso a modelli euristici e algoritmi *ad hoc*, con evidenti ricadute sulla scalabilità, perché implicava la dipendenza del software dai dati di origine e una minore robustezza dell'applicazione con nuovi materiali da processare[5].

Solo nel 2009, con l'introduzione di processi di *machine learning* che simulano l'attività umana, si assiste al superamento del paradigma iniziale e all'evoluzione progressiva e generalizzata dei sistemi di ATR. Gli studiosi, appoggiandosi sulle nuove tecnologie, intendono sfruttare la funzione di memoria predittiva che sta alla base delle reti neurali ricorrenti (RNN), come i processi Long Short-Term Memory (LSTM), per accrescere l'affidabilità dei software di trascrizione automatica. Gli strumenti che traggono maggiore vantaggio dallo sviluppo di questi componenti sono proprio i sistemi di HTR, poiché le loro prestazioni si basano su un addestramento manuale che potenzialmente deve consentire di trascrivere testi sconosciuti, che non sono stati precedentemente processati dalla macchina.

Le caratteristiche proprie dei sistemi di *machine learning* incidono anche sulle modalità di allenamento del software di HTR, in quanto il set di dati fornito dall'utente rappresenta il nucleo dell'intero processo e va confezionato seguendo modalità che ne assicurino successivamente l'efficienza e l'esportabilità a contesti diversi. Secondo questa prospettiva, il lavoro di trascrizione manuale deve rispettare dei criteri che assicurino un certo grado di robustezza del modello (*consistency*), il quale occasionalmente può essere sottoposto a operazioni di messa a punto (*fine tuning*). Così come avviene per i LLM, che su questo aspetto somigliano molto ai sistemi di HTR, l'addestramento della macchina è tanto più efficiente quanto più si incrementa la quantità e qualità dei materiali processati; ciò nonostante, la varietà interna a volte si deve preferire alla costituzione di un dataset molto esteso di materiali dello stesso tipo, ai fini di limitare l'iperspecializzazione del modello.

Ciò che accomuna le recenti tecnologie di Intelligenza Artificiale è proprio la necessità di un addestramento su materiali di base forniti dall'utente (*ground truth*) che rappresentano il modello ideale. Nel caso di strumenti come ChatGPT, la creazione di *ground truth* è un'operazione costosa e faticosa che richiede la revisione di miliardi di parole. Nel caso della trascrizione automatica, invece, i dati immessi dall'utente possono essere di quantità nettamente inferiore, ma non per questo si riducono i costi e le implicazioni etiche dell'operazione. In entrambi i casi descritti, infatti, il materiale di allenamento, che può essere considerato una sorta di big data testuale, è frutto di una selezione operata dall'utente, che così alimenta lo sviluppo della nuova tecnologia. Ma la scelta dei materiali processabili chiaramente è motivata da trend che governano la cultura dominante, per cui si preferisce lavorare su oggetti culturali di enorme interesse, più accattivanti per la disseminazio-

[5] Vinciarelli 2003, p. 8-9.

ne, la collaborazione e la ricerca[6]. Questo comporta il rischio che interi patrimoni culturali già scarsamente visibili diventino totalmente opachi, mentre altri, che già stavano al centro dell'attenzione ed erano disponibili in grande quantità, accrescano a dismisura la loro visibilità e disponibilità, rinsaldando così la posizione culturale dominante di chi gestisce quegli stessi materiali. Secondo questa prospettiva, la creazione di *ground truth* per alimentare sistemi di *machine learning* è un'operazione nient'affatto banale, che deve necessariamente confrontarsi con i concetti di accessibilità ed equità nella produzione e riuso digitale dei beni culturali, in linea con le disposizioni della nuova agenda europea in materia di digitalizzazione.

Sistemi di HTR: stato dell'arte e flusso di lavoro

I sistemi di HTR, per le ragioni fin qui esposte, stanno suscitando l'interesse degli studiosi di materie umanistiche perché assicurano un grado di affidabilità elevato nella trascrizione automatica di documenti antichi. Seguendo questa prospettiva, il loro impiego è vantaggioso soprattutto con documenti a stampa antichi, perché spesso queste tecnologie consentono di ottenere un testo elettronico che si avvicina al 99% di precisione (vale a dire 99 caratteri corretti su 100, spazi inclusi). Si tratta di risultati impensabili fino a qualche decennio fa, quando gli strumenti di OCR con i medesimi documenti promettevano risultati meno attendibili (con un margine di precisione che si aggirava intorno all'80-85% nel migliore dei casi), alimentando una certa reticenza o avversione da parte degli specialisti del testo (*bias*), i quali ne segnalavano la completa inutilità per studi di carattere accademico.

Nell'attualità, invece, il panorama è sostanzialmente mutato, perché i sistemi di HTR permettono di risolvere, generalmente senza ostacoli, problemi prima insormontabili, come la trascrizione automatica in presenza di legature, abbreviature e caratteri speciali. Questi segni grafici, infatti, seppur facilmente interpretabili da parte di un lettore umano, condizionavano i risultati del riconoscimento con software di OCR, abbassandone drammaticamente le prestazioni. Lo stesso discorso si potrebbe estendere anche ai documenti manoscritti, che tuttavia costituiscono una sfida ancora maggiore per le nuove tecnologie perché presentano un'elevata variabilità interna anche a livello di realizzazione grafica di singole lettere.

Sulla scorta delle inedite potenzialità offerte dai sistemi di HTR attuali, sono stati inizialmente i grandi gruppi editoriali e le aziende tecnologiche private a integrare questi strumenti tra i loro servizi. Adam Matthew Digital, per esempio, impiega gli HTR per offrire all'utente una base di ricerca all'interno delle proprie collezioni digitalizzate; Google, invece, ha sviluppato di recente la piattaforma *Fabricius*, che si serve dell'IA per la trascrizione automatica di geroglifici. In questi casi, però, non si forniscono dichiarazioni riguardo gli algoritmi impiegati e non si diffonde il codice sorgente, pertanto l'implementazione resta appannaggio di una comunità ridotta di utenti[7].

[6] Stokes e Kiessling 2024.
[7] Terras 2022, p. 184.

Sul versante opposto si situano invece le piattaforme di HTR sviluppate nel contesto di una comunità scientifica più estesa e che si caratterizzano per essere dei *Virtual Research Environment*, vale a dire delle applicazioni collaborative che possono essere utilizzate da qualsiasi tipologia di utente per finalità specifiche.
Questi strumenti propongono un flusso di lavoro che si compone di varie fasi:
1. *Data collection e preprocessamento delle immagini*. In primo luogo, si procede con la definizione dei materiali che costituiranno la base dell'allenamento del modello HTR. Questi possono essere estratti da un unico documento (*single document*) o da documenti diversi (*multiple documents*), accomunati tra loro dal fatto di presentare caratteristiche linguistiche, grafiche o tipologiche condivise. In occasione, questi stessi materiali, prima di essere importati nella piattaforma, possono essere processati in locale con software che consentono di migliorarne la leggibilità, riducendo il rumore causato da elementi, come macchie o trasparenze di inchiostro, che complicano il riconoscimento. Nel caso dei sistemi di OCR, la critica sottolinea come questa operazione sia indispensabile ai fini di ottenere risultati di segmentazione e trascrizione potenzialmente affidabili: nella maggior parte dei casi per limitare il problema del rumore presente nelle digitalizzazioni si sceglie di convertire l'immagine in bianco e nero (*bynarization*). I sistemi di HTR recenti, al contrario, integrano applicazioni che preferiscono immagini a colori ad altissima risoluzione, in cui la macchina è capace di determinare le zone di testo presenti nella pagina a partire dalla distribuzione dei pixel, individuando e separando tra di loro aree di maggiore e minore densità di colore.
2. *Segmentazione*. Una volta importate le immagini digitalizzate all'interno della piattaforma, si compie una prima operazione fondamentale nel processo di trascrizione automatica che prende il nome di *Layout Analysis* e corrisponde alla segmentazione della pagina in regioni, righe di testo e, se necessario, parole. Normalmente, un tool di HTR integra tra le sue funzionalità un sistema di segmentazione automatica abbastanza affidabile, che distingue principalmente la/e cassa/e di testo e le corrispondenti linee che la/e compongono, mentre si trascura solitamente la suddivisione di linee in entità minori come le parole, poiché, come già chiarito, il riconoscimento si basa su *pattern* di segni grafici ricorrenti distribuiti in una linea di testo. Successivamente, l'utente ha la possibilità di correggere o ritoccare la segmentazione manualmente per mezzo di appositi strumenti integrati alla piattaforma.
3. *Creazione della ground truth*. Si tratta di una delle fasi fondamentali del flusso di lavoro e corrisponde alla corretta segmentazione e trascrizione manuale (secondo criteri prestabiliti) di alcune pagine di testo, che costituiscono il materiale di base (*golden standard*) su cui si fonda l'allenamento del software. L'importanza di questa attività è determinata da due ragioni principali: da un lato, deve rispettare criteri di qualità e coerenza metodologica poiché è la base per la creazione di un modello di HTR solido (*consistent*); dall'altro, rappresenta il nucleo del lavoro di

ricerca, e pertanto deve essere sottoposta ad una valutazione per la preservazione a lungo termine dei dati, secondo formati standard che ne consentano la futura esportazione ad altri contesti. Inoltre, questa fase può avvalersi di strumenti che facilitano la trascrizione manuale, per esempio si potrebbe prevedere l'impiego di modelli HTR già disponibili ai fini di generare una trascrizione preliminare dei contenuti, che, sebbene imprecisa, serva da base al ricercatore per la creazione del dataset finale[8].

4. *Allenamento e creazione di un modello HTR*. Una volta realizzata la trascrizione di un numero adeguato di pagine (la cui quantità dipende dalle caratteristiche del testo sottoposto a riconoscimento), si procede con l'avvio della fase di *training*, vale a dire l'addestramento vero e proprio della macchina. Di solito l'allenamento prevede la suddivisione della *ground truth* in due parti: la prima, corrispondente all'insieme di allenamento (*training set*), è composta dal 90% del dataset e costituisce la base concreta dell'addestramento; la seconda, conforme all'insieme di validazione (*validation set*), è composta dal 5-10% del materiale trascritto dall'utente e permette di valutare l'efficienza del modello ottenuto. L' addestramento si articola in fasi che prendono il nome di *epochs*, vale a dire periodi successivi in cui l'applicazione tenta ricorsivamente di "avvicinarsi" alla trascrizione modello del *validation set* basandosi sulle informazioni acquisite durante il *training*. A tal riguardo, è bene considerare che non sempre la scelta di un maggior numero di *epochs* corrisponde a una miglior produttività: un modello di HTR, infatti, può tendere all'iperspecializzazione sui materiali di training, perdendo in efficienza se applicato a materiali affini. Questo fenomeno di atrofizzazione del modello si conosce come *overfitting* e viene normalmente controbilanciato attraverso la funzione di *early stopping*, che implica l'interruzione dell'addestramento se non si verifica un progressivo avvicinamento tra gli indici di errore di *training* e *validation set*[9].

5. *Valutazione del modello*. L'indice di errore del modello, in inglese Character Error Rate (CER), viene calcolato determinando la distanza tra la trascrizione corretta fornita dall'utente e la trascrizione generata dal software. In concreto, questo numero percentuale viene stabilito in base alla sommatoria delle operazioni (aggiunte, eliminazioni, sostituzioni) che consentono di ottenere il testo corretto a

[8] Al riguardo, ricerche attuali segnalano possibili soluzioni per produrre una *ground truth* sintetica, per esempio Perdiki 2023.

[9] Su questi temi si veda Rabus, che afferma: «Using many epochs typically leads to the asymptotic approximation of the training data CER curve to zero, which means that the model has more or less completely adapted to the training data. However, while the fact that using a great number of epochs during training takes up many computational resources is getting less and less significant nowadays due to improved hard and software capabilities, another important unintended effect brought about by using many training epochs must not be neglected. Since more training epochs adapt the model to the training data better and better, it may happen that model performance on test data stagnates or even becomes worse. This phenomenon is called overfitting and is quite common when training neural networks» (2019, p. 12).

partire dal testo grezzo trascritto dalla macchina[10]. Da ciò deriva che l'indice non corrisponde a un dato assoluto, ma relativo, perché fondato su un dataset di prova che non include testi totalmente estranei al modello, ma documenti spesso impiegati per lo stesso addestramento. Per questa ragione, il CER è un indicatore a volte fuorviante, che va rivalutato ogni volta che si sottopongono al modello nuovi materiali da interpretare.

Sviluppi recenti: piattaforme disponibili, modelli di HTR misti e fine tuning

Per svolgere le operazioni descritte all'interno di un unico spazio di lavoro, il ricercatore ha a disposizione due piattaforme ad accesso aperto: Transkribus (READ Coop)[11] ed eScriptorium[12]. Si tratta delle due applicazioni più famose e utilizzate nell'ambito della ricerca universitaria e archivistica, perché entrambe dispongono di un'interfaccia grafica intuitiva e non richiedono competenze nell'ambito della programmazione. Inoltre, entrambi i software seguono grossomodo il flusso di lavoro descritto nei precedenti paragrafi, consentendo la creazione di modelli di HTR a partire dai dati di specifico interesse dell'utente

La differenza principale tra Transkribus ed eScriptorium si basa sulle prospettive di uso e di distribuzione alla base di ciascun progetto. Gli sviluppatori di eScriptorium, infatti, hanno deciso di seguire una filosofia totalmente Open Access, elaborando un'applicazione web che si appoggia all'infrastruttura di Kraken e permette l'esportazione in locale dei modelli HTR creati dall'utente[13]. Transkribus, al contrario, si presenta come un'infrastruttura che segue un'ottica più commerciale, perché richiede l'ottenimento di crediti per il riconoscimento, una volta esauriti i crediti mensili gratuiti a disposizione dell'utente, e permette la condivisione dei modelli solamente all'interno dello spazio cloud della piattaforma, senza la possibilità di un'esportazione in locale.

Per questa ragione, si è sviluppato negli ultimi anni un intenso dibattito intorno a Transkribus e il suo impiego all'interno di progetti universitari, con un ampio numero di critiche provenienti dai sostenitori dell'Open Access e della libera circolazione delle risorse digitali. A ben vedere, però, bisogna considerare una differenza sostanziale alla base dell'utilizzo delle due piattaforme, soprattutto per quanto ri-

[10] Comunemente la distanza tra due testi si denomina *edit distance* o distanza di Lavenshtein: il dato si ottiene calcolando il numero di operazioni (aggiunte, soppressioni, sostituzioni) che consentono di passare da un testo A ad un testo B.

[11] L'applicazione è disponibile in due versioni: l'*expert client*, scaricabile alla seguente pagina web <https://readcoop.eu/transkribus/>, e l'applicazione web, accessibile dalla medesima pagina. Per maggiori informazioni relative a Transkribus si veda Mühlberger et al. 2019.

[12] Per maggiori informazioni su eScriptorium si consulti Stokes et al. 2021. La documentazione relativa all'installazione e all'utilizzo della piattaforma si trova disponibile al seguente link: <https://escriptorium.readthedocs.io/en/latest/>.

[13] Pinche 2023; Chagué-Clérice 2020-23.

guarda l'installazione, lo sviluppo e la risoluzione dei problemi lato utente. Di fatto, eScriptorium è un software, non un servizio, pertanto non è accessibile da un unico sito web, ma da distinti server alloggiati presso le istituzioni che decidono volontariamente di installare l'applicazione e metterla a disposizione di utenti esterni[14]. Questo comporta che l'uso dell'applicazione sia totalmente libero e sostenibile, ma d'altro canto impone delle limitazioni: nel caso di installazione locale, l'utente necessita di competenze informatiche avanzate e di una macchina che possa sostenere un'elevata potenza di calcolo; se invece si decide di appoggiarsi a server già attivi, inevitabilmente si incorrerà in problematiche specifiche del server, per esempio continui aggiornamenti o paralisi di alcuni comandi, a cui spesso si accompagna un supporto tecnico meno efficiente, perché legato alla disponibilità dei gestori del server.

Gli sviluppatori di Transkribus, invece, hanno preferito sviluppare un sistema cloud di appoggio alla piattaforma, attualmente ospitato nei server dell'Università di Innsbruck. Questo comporta che l'accesso al software non richieda all'utente particolari competenze informatiche o tecnologiche, né vi sia la necessità di un supporto istituzionale per l'installazione. La conseguenza diretta di queste scelte è che Transkribus ha visto una più estesa diffusione a partire dal 2018, con un sostanzioso allargamento del bacino di utenti registrati e un incremento del numero di istituzioni partner che partecipano al progetto[15]. Poiché i sistemi di *machine learning* alla base di queste piattaforme progrediscono con la quantità di materiali processati, questa espansione della comunità di utenti di Transkribus corrisponde a una maggior efficienza dell'applicazione, che attualmente assicura risultati di segmentazione e trascrizione nettamente migliori rispetto a quelli offerti da eScriptorium. Inoltre, la semplicità di installazione e utilizzo di Transkribus corrisponde a una maggiore distribuzione della piattaforma in contesti accademici, dove gli utenti spesso non hanno esperienza con linea di comando e altre specificità tecniche.

Le piattaforme di HTR come Transkribus ed eScriptorium consentono, come si diceva, di creare un modello di riconoscimento adatto all'interpretazione del testo contenuto in documenti a stampa e manoscritti. In prima battuta, questo modello può basarsi su di un unico documento, del quale l'utente trascrive manualmente un numero determinato di pagine all'interno della piattaforma, con la volontà poi di lanciare il modello HTR creato sulle restanti pagine. In tal caso, si parla di modello HTR individuale, vale a dire di uno strumento specificatamente concepito per l'interpretazione di un singolo testo. Da una parte, questa operazione assicura

[14] Stokes 2021.

[15] Nel momento in cui scrivo la comunità di READ Coop vanta all'incirca 160 membri provenienti da più di 30 Paesi diversi. Tra i membri si contano più di 70 istituzioni partner, tra cui la British Library, archivi nazionali come quelli di Zurigo, Lussemburgo, Norvegia, e varie università europee ed extra-europee. La lista è disponibile al seguente indirizzo: <https://readcoop.org/members>.

alcuni vantaggi, per esempio un maggior controllo del flusso di trascrizione automatica nel rispetto dei criteri di trascrizione stabiliti, dal quale deriva la possibilità di prevedere in che punti del testo la macchina potrebbe incontrare maggiori difficoltà. D'altro canto, il limite di un modello individuale è quello di essere efficace solo nel contesto di opere di una certa lunghezza: in presenza di testi brevi, infatti, l'utente non dispone di materiale sufficiente per creare una *ground truth* adeguata. Pertanto, si può con ragione affermare che al momento i modelli individuali sono particolarmente utili nel contesto di testi manoscritti molto estesi, in cui i problemi derivati dalle variazioni interne della grafia, che generalmente complicano l'intero processo di riconoscimento, possono essere superati con la creazione di un imponente dataset di addestramento.

In risposta alla necessità di diffondere l'uso dei sistemi di HTR ad opere di breve estensione, come relazioni, frammenti, articoli di giornale ecc., è stata recentemente introdotta la possibilità di creare modelli misti (*general models*), ossia modelli basati su diversi documenti che vengono raccolti in un unico set di dati e costituiscono la base dell'allenamento[16]. I modelli misti possono essere di tipo diverso, a seconda dalle opere che compongono il dataset. Nel caso della trascrizione automatica di documenti a stampa, spesso la scelta è quella di includere in un unico modello misto diverse realizzazioni grafiche di uno stesso set di caratteri, per esempio comprendendo opere affini per tipologia o pubblicate da una stessa stamperia in un lasso di tempo determinato. Viceversa, per il riconoscimento di testi manoscritti si preferisce ricorrere ad opere di uno stesso autore, per fare in modo che l'addestramento tenga conto delle diversità della grafia che, benché provenga da una stessa mano, dipende dalle condizioni di scrittura, dal tempo di composizione e dai supporti impiegati. In entrambe le situazioni, inoltre, si consiglia di fondare il modello su opere di diversa provenienza, presupponendo che la varietà interna derivata da tecniche di acquisizione e preservazione distinte estenda le possibilità del modello.

Recentemente, la creazione di modelli misti molto estesi sta attirando l'attenzione della comunità scientifica, anche perché gli stessi modelli misti possono costituire la base di un allenamento personalizzato. In tal caso si parla di *fine tuning*, vale a dire di perfezionamento di un modello misto esteso, che viene usato come *base model* nella creazione di un modello individualizzato che include specifici materiali di *training* trascritti manualmente dall'utente. Questa operazione è largamente praticata, poiché limita la necessità di creare dataset *ex novo* e incentiva il riuso di materiali già processati, ma va sempre sottoposta a un controllo da parte dell'utente perché può comportare una scarsa sistematicità dei risultati di trascrizione, che derivano dall'impiego di un modello HTR di base in parte sconosciuto.

Per tale ragione, la corretta descrizione dei dataset che si sono impiegati per la

[16] Hodel et al. 2021.

creazione di modelli misti comincia a rivestire un'importanza sempre maggiore, poiché solamente chiarendo le caratteristiche del modello e i criteri di trascrizione utilizzati si assicura che quegli stessi materiali possano essere utili ad altri utenti. I progetti OCR-D[17] e HTRUnited[18], che sono i principali canali di distribuzione di dataset per la trascrizione automatica, insistono proprio sulla definizione dettagliata dei requisiti di ciascun modello, fomentando la libera circolazione di materiali di training presso altri progetti.

Il Progetto Mambrino: esperienze e prospettive nel campo della trascrizione automatica

Il Progetto Mambrino è un gruppo di ricerca dell'Università di Verona diretto da Anna Bognolo e Stefano Neri che si occupa dello studio dei *libros de caballerías*, romanzi che si diffusero per più di un secolo in tutta la Penisola Iberica a partire dalla pubblicazione nel 1508 dell'*Amadís de Gaula* di Garci Rodríguez de Montalvo. Il genere ebbe un enorme successo in Spagna, con la pubblicazione di più di 60 titoli e diverse riedizioni dei libri appartenenti ai cicli più famosi come quelli di Amadís e Palmerín, sino a varcare i confini nazionali e venire distribuiti in tutta Europa, in Italia, Francia, Inghilterra, Germania, convertendosi in uno dei primi fenomeni di letteratura di massa. In Italia questi romanzi circolarono in un primo momento in lingua originale, con epicentro di produzione prima a Roma e poi a Venezia, grazie alle imprese editoriali dei Nicolini da Sabbio e Giovanni Battista Pederzano[19]; ma ben presto, a causa del successo che riscontrava il genere, si iniziarono a stampare traduzioni e continuazioni originali, le principali ad opera di Mambrino Roseo da Fabriano, che tra il 1550 e il 1580 operava in relazione con l'editore Michele Tramezzino, un importante stampatore veneziano[20].

Al centro degli interessi del Progetto Mambrino c'è la volontà di riportare alla luce questo genere romanzesco, che tanta fortuna ebbe in Italia durante il Rinascimento, per mezzo dello studio delle opere e il censimento di edizioni ed esemplari conservati. Per rispondere a questa necessità, uno degli obiettivi principali del progetto è quello di pubblicare un'edizione moderna dei testi, realizzata seguendo criteri rigorosi e invalsi nella disciplina. Non si tratta però di un'operazione semplice, vista l'estensione del corpus e delle opere che lo compongono: al riguardo, si consideri che il solo Ciclo di Amadis di Gaula italiano si compone di 21 libri e ciascuno di essi supera le 400 carte[21].

[17] https://ocr-d.de/en/.
[18] https://htr-united.github.io/.
[19] Bognolo 2017.
[20] Bognolo 2012; Bognolo-Cara-Neri 2013.
[21] Per uno studio bibliografico delle traduzioni e continuazioni italiane dei *libros de caballerías* si consulti la sezione "Spagnole Romanzerie" del sito web del progetto: <https://www.mambrino.it/it/spagnole-romanzerie>.

A partire dal 2016, grazie all'esperienza maturata dai membri del progetto nel campo delle Digital Humanities, ha preso corpo l'idea di realizzare una Biblioteca Digitale dei romanzi cavallereschi italiani di derivazione spagnola, per mezzo della creazione di edizioni scientifiche digitali dei volumi che compongono i cicli di Amadis e Palmerino, le quali, integrate a database bibliografici e semantici, consentano lo studio del contesto editoriale e degli aspetti narrativi del genere[22]. Ai fini di realizzare un'edizione, in quella circostanza sono state esplorate per la prima volta le possibilità di applicare strumenti di trascrizione automatica per ottenere il testo grezzo delle opere in formato elettronico, poi editabile secondo i criteri stabiliti dal progetto[23].

Dopo le iniziali sperimentazioni con sistemi di OCR come ABBYYFineReader[24], che non assicuravano risultati affidabili se applicati a testi in caratteri corsivi, per generare le trascrizioni delle opere si è scelto di impiegare la piattaforma Transkribus (READ Coop), facendo affidamento sulle migliori prestazioni dei sistemi di HTR nell'ambito di testi a stampa antichi. I risultati furono sin da subito impressionanti, con la creazione di modelli individuali per ciascuna opera che si avvicinavano al 99% di precisione, indipendentemente dallo stato di conservazione e dalla qualità delle digitalizzazioni[25]. I criteri di trascrizione scelti per la creazione del dataset di ogni modello individuale implicavano una modernizzazione del testo, con lo scioglimento delle abbreviazioni e l'interpretazione di caratteri speciali del corsivo, come le legature e il segno tironiano, che complicavano l'applicazione di sistemi di OCR tradizionali. Al contrario, quegli stessi elementi grafici non costituivano un particolare problema per un sistema di HTR come Transkribus, che offriva risultati di trascrizione affidabili, generando testi adatti sia per la creazione di edizioni che per lo studio con strumenti computazionali, per esempio analisi quantitative di stilometria[26].

Mentre si proseguiva con la trascrizione automatica e la posteriore correzione dei testi ottenuti con l'intenzione di popolare la Biblioteca Digitale, l'esperienza maturata ispirò nuovi sviluppi nel campo del riconoscimento di testi a stampa antichi. Nel 2021, grazie al sostegno dei gruppi di ricerca spagnoli BIDISO (Biblioteca Digital Siglo de Oro[27]) e COMEDIC (Catálogo de obras medievales impresas en castellano[28]), fu possibile avviare un progetto collaborativo che condusse alla creazione di due modelli di HTR per l'interpretazione di testi a stampa iberici in caratteri gotici e romani[29]. In quell'occasione, si decise per la prima volta di lavorare alla

[22] Bognolo-Bazzaco 2019; 2024.
[23] Bazzaco 2018.
[24] Mancinelli 2016.
[25] Bazzaco 2018, p. 268.
[26] Bazzaco 2021.
[27] https://www.bidiso.es/.
[28] https://comedic.unizar.es/?es.
[29] Bazzaco et al. 2022.

creazione di modelli misti, che integrassero la trascrizione manuale di una porzione di opere che differivano per tipologia editoriale, genere e qualità delle fotoriproduzioni. Guidati dalla volontà che i modelli di HTR sviluppati fossero poi esportabili ad altri contesti di ricerca, per la costituzione della *ground truth* si preferì realizzare una trascrizione manuale conservativa, che riproducesse fedelmente i segni grafici presenti nelle fonti, ad eccezione della "s lunga", convertita in "s semplice" e del segno tironiano, trascritto come "&". Le abbreviazioni anche in questo caso si sciolsero, ai fini di ottenere testi finali più facilmente interpretabili e processabili da parte dei collaboratori del progetto[30]. Inoltre, i modelli si diffusero pubblicamente all'interno della piattaforma Transkribus e furono costantemente implementati, con l'aggiunta di nuove trascrizioni realizzate da altri studiosi interessati, che permisero un ampliamento del dataset di addestramento e, corrispondentemente, un sensibile incremento delle prestazioni di ciascun modello[31].
La qualità dei risultati ottenuti dimostrò l'alto grado di affidabilità dei sistemi di HTR misti in relazione con i documenti a stampa antichi, gettando le basi per lo sviluppo di un modello che consentisse l'interpretazione generalizzata di testi in caratteri corsivi, utile non solo per le opere al centro degli interessi del Progetto Mambrino, ma esportabile anche ad altri contesti di ricerca.

Verso un modello di HTR per il corsivo del XVI secolo

Nel 2023, alla luce dell'esperienza maturata dal Progetto Mambrino nel campo della trascrizione automatica, ha preso avvio lo sviluppo del modello di HTR misto *Italics_VeniceXVIc*[32].
1. *Selezione dei materiali di addestramento*. La prima fase per la creazione del dataset all'interno della piattaforma Transkribus corrisponde alla scelta delle opere che andranno a costituire la base per l'addestramento della macchina. Per eseguire questa operazione, sono stati selezionati alcuni romanzi cavallereschi italiani di interesse del progetto, secondo criteri di variabilità dei caratteri tipografici e diversità dei processi di acquisizione delle immagini (Fig. 1).

[30] Bazzaco et al. 2022, p. 95.
[31] Si segnalano di seguito le principali caratteristiche dei due modelli:
- Modello *SpanishGothic_XV-XVI_extended* (vers. 1.2.0). Stefano Bazzaco (coord.), Federica Zoppi, Giada Blasut, Nuria Aranda García, Ángela Torralba Ruberte, Ana-Milagros Jiménez Ruiz, Pedro Monteiro, José Manuel Fradejas, Eduardo Camero Santos, Laura Lecina Nogués, Almudena Izquierdo Andreu. Dataset disponibile all'indirizzo: <https://doi.org/10.5281/zenodo.4888926>.
- Modello *SpanishRedonda_XVI-XVII_extended* (vers. 1.2.0). Stefano Bazzaco (coord.), Gaetano Lalomia, Daniela Santonocito, Manuel Garrobo Peral, Mónica Martín Molares, Carlota Cristina Fernández Travieso, Giulia Tomasi, Alessia Fichera, Soledad Castaño Santos, Almudena Izquierdo Andreu. *Dataset* disponibile all'indirizzo: <https://doi.org/10.5281/zenodo.4889217>.

[32] Lo sviluppo del modello *Italics_VeniceXVIc* è stato realizzato da chi scrive in collaborazione con Giulia Lucchesi (Università di Verona), che ringrazio.

PRIMA PARTE DEL TERZODECIMO LIBRO DI AMADIS DI GAVLA,

nelqual si contiene le stupende, & marauigliose prodezze del Prencipe Sferamundi figliuolo del valoroso don Rogello, di Amadis d'Astra, et altri sforzati cauallieri.

Che il prencipe Sferamundi insieme con Amadis d'Astra, fu portato dalla naue del gran Serpente dal la sfera col mezzo della donzella che la guidaua innanzi l'imperator di Varti, perche gli ammosse cauallieri. Cap. I.

OSTO che la naue del marauiglioso Serpente de i Magli si mise in altro mare, dilegnandosi dalla uista dei Prencipi, & Prencipesse di Grecia che la stauan mirando, furono i duo valorosi Prencipi sferamundi, & Amadis d'Astra con somma uelocità portati, senza che ne la donzella Alchisa, ne i duo nani si adoperassero, nell'isola del fuoco, così chiamata, perche di lontano pareua à nauiganti, che tutta ardesse, & giunti al lito di essa ben si uedea il fuoco, & la gran fiamma, ma era l'isola di un aere temperato,

A

AGGIVNTA NOVAMENTE RITROVATA NE gli Annali de gli Imperadori Greci al libro dello inuittissimo cauallieve PRIMALEONE figliuolo dello Imperadore di Greci Palmerino di Oliua,

Intitolata quarta parte del libro.

Come, & donde nacque l'errore, perche il fine del libro terzo di Primaleone non ua ad affrontarsi con il libro di Platire per mancamento di alcuni fogli lasciati à dietro dal traduttore, & l'errore di alcuni nomi equiuocati. Cap. I.

MOLTI restano marauigliati, come nel fin del terzo libro di Primaleone facendosi mentione di alcune cose notate dall'historico, nel principio del libro chiamato Platir, che seguita, piu non se ne parli, anzi si ragiona di historia diuersa in parte, & in quelle cose che si accordano, i nomi sieno varij con le prouincie, & i luoghi. Et perche non resti il lettore marauigliato, è bene di venderne la ragione, prima che si cominci à ripigliare l'historia che per errore fu in parte lasciata à dietro. Nel tempo che da Turchi fu con tanta strage, presa la gran città di Costantinopoli,

A

DI AMADIS DI GAVLA. 4

uendo con le mie arti saputo i maneggi di questa guerra, ne i quali uoi, come cauallieri risentito, ui trauagliate tanto, son uenuto a trouarui per darui aiuto di opera, in quanto per me si può, & di consiglio, tenendo per fermo, che questa uolta sia per rinfarci ogni nostro disegno, ma ha bisogno, che andiamo a trouare il Re Arauigo uostro zio, nell'Isola Liconia, done se ne sta senza saputa alcuna di queste pratiche, & senza speranza di ricuperar mai il perduto. Sifardo si rallegrò molto di queste parole, & disse che uolena far quanto egli diceua, & che senza porre indugio douessero pigliar il camino, & Arcalaus riuolse il cauallo a mano stanca, che era la strada d'andare nell'Isola di Liconia.

Che Arcalaus & Sifardo giunsero nell'Isola di Liconia, & il ragionamento notabile fatto fra Arcalaus & il Re Arauigo, & la conclusione di esso. Cap. II.

CON grande allegrezza caualcauano Arcalaus & Sifardo uerso l'Isola di Liconia, uenendo Sifardo ragguagliandolo de i Re & principi pagani, che haueano a concorrere in questa impresa, & tutte le pratiche che egli hauea fatte in diuersi luoghi, & come Arnoldo suo fratello era andato a cercare Sifardo

A iiij

Figura 1. *Esempi di materiali inclusi nel corpus, tratti da Sferamundi di Grecia. Prima parte; La quarta parte di Primaleone; Aggiunta al quarto libro di Amadis di Gaula (in ordine da sinistra a destra).*

Si offre di seguito la lista delle opere incluse nel modello, seguendo l'ordine cronologico di pubblicazione di ciascun esemplare:

Autore/traduttore e Opera	Data	Editore	Provenienza digitalizzazioni
Mambrino Roseo da Fabriano, *Flortir*	1558	Michele Tramezzino	Praha, Národní knihovna České republiky, 9 J 000225
Mambrino Roseo da Fabriano, *Sferamundi di Grecia. Prima parte*	1558	Michele Tramezzino	Madrid, Biblioteca Nacional de España, 5-4978
Francesco Portonaris, *Palmerino d'Inghilterra*	1558	Francesco Portonaris	Wien, Österreichische Nationalbibliothek, 8.J.33.(Vol.1)
Mambrino Roseo da Fabriano, *Sferamundi di Grecia. Seconda parte*	1560	Michele Tramezzino	Madrid, Biblioteca Nacional de España, 5-4978
Mambrino Roseo da Fabriano, *La quarta parte di Primaleone*	1560	Michele Tramezzino	Wien, Österreichische Nationalbibliothek, 0.M.40.(Vol.2)
Mambrino Roseo da Fabriano, *Aggiunta al quarto libro di Amadis di Gaula*	1563	Michele Tramezzino	Santiago de Compostela, Biblioteca Universitaria, 13996
Mambrino Roseo da Fabriano, *Aggiunta al Florisello (Le prodezze di don Florarlano)*	1564	Michele Tramezzino	Verona, Biblioteca Civica, Cinq. E 350-12
Mambrino Roseo da Fabriano, *Aggiunta a Rogello di Grecia*	1564	Michele Tramezzino	Verona, Biblioteca Civica, Cinq. 350-15

Le digitalizzazioni di queste opere, che erano già state importate all'interno della piattaforma in occasione della creazione di modelli individuali specifici, sono state raccolte in un'unica collezione all'interno della piattaforma: di ciascuna di esse si sono selezionate 20 pagine interne corrispondenti al testo in prosa, in modo da confezionare un dataset sufficientemente equilibrato, in cui nessun testo prevaricasse sugli altri in termini di quantità.

2. *Definizione dei criteri di trascrizione*. Una volta selezionati i materiali di addestramento, si è passati alla definizione dei criteri di trascrizione. Come già anticipato, questa operazione richiede un'attenta valutazione da parte dello studioso, poiché le scelte di trascrizione incidono in maniera determinante sui risultati finali dell'allenamento del modello. Come stabilito in Heigl, «the machine must learn from a transcription that is as accurate as possible so that it can later reproduce exactly what is written on the sheet» (2019), pertanto è preferibile una trascrizione diplomatica. Questo suggerimento, che non genera problemi, per esempio, nel caso della trascrizione del segno tironiano come "&" o nel mantenimento della distinzione "u/v", si complica nel caso delle abbreviature, che possono corrispondere a più di un carattere nel testo trascritto. L'utente al riguardo ha a disposizione tre diverse possibilità: sviluppare l'abbreviazione, per rendere il testo finale più comodamente ricercabile; mantenere il carattere speciale in corrispondenza dell'abbreviazione, possibilmente facendo ricorso a un set di caratteri standard, come "Latin Extended-D" di Unicode; trascrivere il carattere speciale come carattere semplice, aggiungendo però una metadatazione di tipo sintattico che indichi che siamo in presenza di un'abbreviazione, per esempio aggiungendo il tag <abbr>[33]. Quest'ultima soluzione, come si può immaginare, richiede un maggior tempo di produzione rispetto alle precedenti; per questo è una pratica poco seguita.

La decisione sui criteri è sempre una fase delicata, che risponde a delle necessità concrete e che comporta dei vantaggi, ma anche, nella maggior parte dei casi, una perdita di informazione. Per esempio, nel caso si confezioni una *ground truth* che presenta una modernizzazione del testo fonte, la trascrizione che si potrebbe ottenere applicando il modello non terrebbe conto della presenza di eventuali "s lunghe" (che verrebbero a uniformarsi con le "s semplici"). In caso contrario, una trascrizione diplomatica potrebbe complicare una normalizzazione automatica dei caratteri, per esempio limitando la regolarizzazione di "u/v" secondo l'uso attuale.

Nel nostro caso, per lo sviluppo di un primo modello di riconoscimento per i caratteri corsivi si è scelto di adottare dei criteri di trascrizione conservativi, secondo una prospettiva di preservazione di segni grafici speciali e soluzioni tipografiche che potrebbero risultare di interesse per settori di studio specifici, per esempio la

[33] Le medesime indicazioni si ritrovano in Heigl 2019 e nelle Guidelines di Transkribus: <https://help.transkribus.org/data-preparation>.

paleografia o la bibliografia materiale dei testi a stampa[34]. In futuro si prevede però la pubblicazione di un secondo modello di riconoscimento che consenta la produzione di un testo modernizzato, in modo da favorire il processamento con strumenti di analisi quantitativa.

3. *Segmentazione e trascrizione manuale*. L'operazione di *Layout Analysis* è stata avviata all'interno della piattaforma per mezzo di due comandi principali: in primo luogo, attraverso la funzione "Printed Block Detection" sono state individuate le regioni di testo; successivamente, sono state individuate in modo automatico le linee di testo, avendo cura di interrompere ciascuna riga in corrispondenza del perimetro delle regioni. Infine, si è provveduto alla trascrizione manuale delle porzioni di ogni opera, secondo i criteri prestabiliti.

4. *Creazione del modello*. L'addestramento del modello *Italics_VeniceXVIc* è stato realizzato selezionando dalla *ground truth* il materiale di addestramento e segnalando una corrispondente porzione di testo, pari al 10% dell'intero dataset, come materiale di validazione. Nello specifico, sono state impiegate 4.848 linee di testo trascritto manualmente, corrispondenti a 40.982 parole processate.

Per l'allenamento si è scelto un numero massimo di *epochs* pari a 350, con una funzione di *early stopping* fissata a 20 *epochs*. Ciò significa che, valutando la coincidenza tra le curve di apprendimento di *training* e *validation set*, e quindi l'inefficacia di ulteriori fasi di addestramento, l'applicazione poteva interrompere l'allenamento dopo il ventesimo periodo: nel caso concreto del modello creato, l'allenamento si è interrotto all'ottantacinquesima *epoch* (Fig. 2).

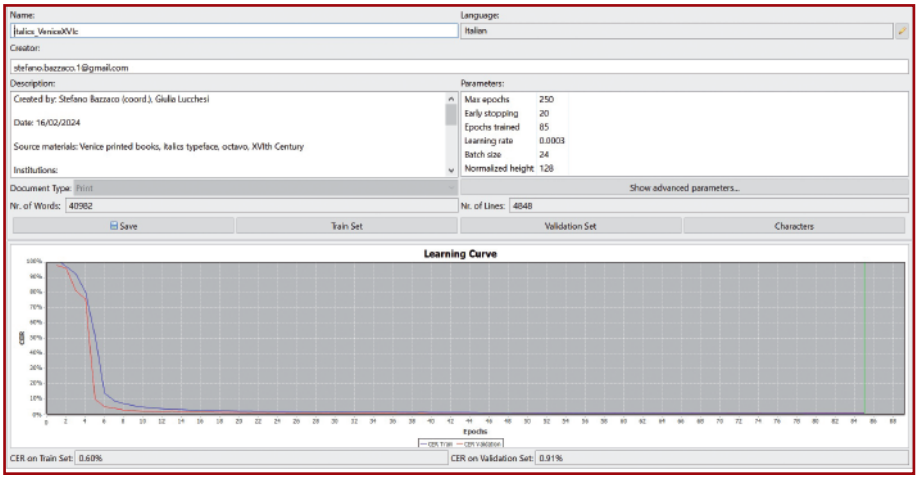

Figura 2. *Dettagli e curve di apprendimento del modello Italics_VeniceXVIc*

[34] Per esempio, si sono conservati caratteri speciali presenti nella fonte come ſ (per la 's lunga') o p̄ (per le abbreviazioni di "per", "pro", "pre"). Una visualizzazione dettagliata dei criteri è consultabile alla pagina: <https://github.com/stefanobazzaco/HTR-model-italics_VeniceXVIc>.

5. *Valutazione e Fine Tuning*. I risultati dell'addestramento, come si è detto, vengono trasmessi secondo un indice chiamato CER, che nel contesto che ci riguarda arriva al 0.91%. L'indice di errore, però, riguarda materiali simili a quelli processati dall'applicazione, vale a dire quelli inclusi nel set di validazione; per questa ragione, è necessario testare la concreta efficienza del modello su nuovi materiali sconosciuti alla macchina. Questa fase di valutazione concreta dei risultati ottenibili, nel contesto del modello creato, è stata particolarmente efficace, perché il riconoscimento su altri testi cavallereschi affini ha rivelato una difficoltà di interpretazione di alcuni elementi presenti sulla pagina, per esempio i titoli e le intestazioni in lettera maiuscola.
Per risolvere queste limitazioni, è possibile realizzare un'operazione di *fine tuning*, vale a dire di perfezionamento del modello. Questa fase potrebbe prevedere l'integrazione di più modelli o l'aggiunta di nuovi materiali al dataset iniziale, opportunamente scelti in relazione con le sezioni del testo dove il riconoscimento automatico rivela maggiori debolezze. Nell'ambito del modello qui descritto è stato impiegato un altro modello di HTR pubblico, chiamato *Transkribus_PrintM1*, come *base model*[35]. L'operazione si è rivelata estremamente utile, perché ha permesso di ridurre ulteriormente il CER, sceso al 0.70%, e ha assicurato risultati più affidabili anche in corrispondenza dei caratteri che non venivano interpretati correttamente in precedenza (Fig. 3).

Figura 3. *Perfezionamento dei risultati grazie all'operazione di Fine Tuning (al centro, modello Italics_VeniceXVIc; a destra, modello Italics_VeniceXVIc+basemodel)*

[35] Una descrizione delle caratteristiche di *Transkribus_PrintM1* può essere consultata nella sezione "Public Models" di READCoop: <https://readcoop.eu/it/model/transkribus-print-multi-language-dutch-german-english-finnish-french-swedish-etc/>.

Come si può osservare, il modello perfezionato non risente dei criteri di trascrizione del modello di supporto *Transkribus_PrintM1*, che prevale solamente in presenza di punti particolarmente critici per il modello *Italics_VeniceXVIc*. Per esempio, si noti la preservazione della "ꞵ", presente nel modello perfezionato ma trascritta come "s semplice" nel *base model*.

6. *Distribuzione e sviluppi futuri.* Forti dell'esperienza maturata con i testi a stampa antichi di area iberica, nell'orbita del Progetto Mambrino si sta valutando la possibilità di implementare il modello attraverso la collaborazione di altri studiosi, che potranno procedere con l'elaborazione di nuove trascrizioni manuali di porzioni di testi simili a quelli già inclusi ed integrarle ai materiali di addestramento. In tal senso, si prevede la distribuzione di una prima versione del modello *Italics_VeniceXVIc* nel mese di giugno 2024, suscettibile di future integrazioni. La *ground truth* di base del modello è attualmente disponibile nella piattaforma Zenodo[36] e contiene i materiali estratti dalla piattaforma Transkribus: le immagini originali, i corrispondenti file nei formati standard ALTO XML e PAGE XML, nonché i file in testo piano (TXT). La costituzione del dataset pertanto rispetta le indicazioni di distribuzione e riutilizzo dei dati suggerite dal progetto HTR-United[37] e, indipendentemente dai futuri aggiornamenti, potrà essere facilmente individuabile a partire da un unico DOI (Digital Object Identifier), che restituirà all'utente l'ultima versione pubblicata.

In seguito, per favorire la creazione di uno strumento che consenta lo scioglimento automatico di abbreviazioni e simboli speciali, adattando quindi le trascrizioni risultanti a nuovi scenari di impiego, si contempla la possibilità di modificare la *ground truth* di partenza secondo nuovi criteri di trascrizione e di procedere all'allenamento di un secondo modello chiamato *Italics_VeniceXVIc_modernized*. Entrambi i modelli saranno resi disponibili nella piattaforma Transkribus (sezione "Public Models") e le rispettive *ground truth* verranno successivamente impiegate nella creazione di modelli affini con la piattaforma eScriptorium, in modo da estendere il loro utilizzo presso una comunità scientifica più ampia.

Conclusioni

Come già chiarito, lo studio che si propone in questo articolo costituisce un primo passo verso la creazione di un modello di HTR che consenta il riconoscimento automatico e generalizzato di testi a stampa in carattere corsivo. L'argomento risulta di estrema attualità, visti i recenti avanzamenti nel campo della trascrizione automatica che, di pari passo con lo sviluppo tecnologico, sta sperimentando un rinnovato interesse nel contesto accademico e archivistico; basti pensare ai recenti pro-

[36] <https://zenodo.org/records/10674282>, DOI: 10.5281/zenodo.10674282.
[37] Chagué-Clérice 2020-23.

getti di area italiana che hanno preso avvio negli ultimi anni e che propongono un'integrazione dei sistemi di HTR alla ricerca storica e letteraria[38].

Nel gruppo di ricerca Progetto Mambrino, in quanto ispanisti interessati alle relazioni tra la letteratura spagnola e quella italiana, ci si è avvicinati a questo campo della ricerca guidati dalla necessità di trascrivere dei romanzi cavallereschi molto estesi, ma la volontà è quella di condividere la nostra esperienza, che risulterebbe molto utile anche a diversi campi dell'italianistica, con cui ci auspichiamo di collaborare. La proposta di lavoro che si descrive in queste pagine, infatti, mira a incentivare lo sviluppo e impiego di sistemi di HTR nel contesto di opere a stampa cavalleresche della metà del Cinquecento, ma la sperimentazione con questi strumenti si può estendere anche ad altri set di caratteri[39]. Nel nostro caso, i sistemi di HTR, già allo stato attuale, assicurano con il corsivo risultati di trascrizione altamente affidabili, con l'ottenimento di testi elettronici che presentano un margine di errore inferiore all'1%. Raggiungere un tale margine di precisione rappresenta un avanzamento notevole, perché gli stessi testi possono poi essere processati con strumenti computazionali, essere impiegati nello sviluppo di edizioni scientifiche digitali e amplificare le possibilità di collazione nel caso di tradizioni testuali complesse, suggerendo nuovi percorsi di integrazione tra le diverse aree di studio che compongono l'estesa galassia delle Digital Humanities.

[38] Zappulli-Iorio 2018; Schwarz-Ricci 2022; Spina 2022; Malatesta 2023.

[39] Nel momento in cui scrivo, siamo coinvolti in due progetti italiani che prevedono lo sviluppo di modelli HTR per incunaboli e testi a stampa:

- Progetto PRIN 2022 *spaNice: Spanish cultural models in Early Modern Venice (the development and circulation of Spanish literature and language in 16th-17th century Italy)*, dir. Federica Zoppi (prot. 202297ATKC);
- Progetto PRIN 2022 *Chivalric Spaces. A digital landscape of texts, stories and narrative motifs of printed popular chivalric literature in Italy and Spain*, direttrice Annalisa Perrotta (prot. 2022HT3XYP).

The contribution aims to describe the workflow that led to the production of a Handwritten Text Recognition (HTR) model for the automatic transcription of sixteenth century Venetian texts printed in italics. In the first part, the scope of the study is defined, with particular attention to the state of the art and recent developments in the field of HTR regarding complex scripts, namely historical printed texts and manuscripts that, due to their characteristics, hinder the application of traditional Optical Character Recognition (OCR) systems. The second part outlines the main phases of the training process for the creation of the Italics_VeniceXVIs model, which represents a first step towards the interpretation of chivalric texts in italics of interest to the Mambrino Project at the University of Verona. Finally, the main characteristics of the model are identified, and, with a view to accessibility and reusability, future steps of the project are highlighted, suggesting possible implications of the research carried out in relation to other fields of study.

L'ultima consultazione dei siti web è avvenuta nel mese di giugno 2024

Riferimenti bibliografici

Bazzaco 2018 — Stefano Bazzaco. *El Progetto Mambrino y las tecnologías OCR: estado de la cuestión*. «*Historias Fingidas*», 6 (2018), p. 257-272. <http://dx.doi.org/10.13136/2284-2667/89>.

Bazzaco 2021 — Stefano Bazzaco. *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)*. In: *"Prenga xascú ço qui millor li és de mon dit". Creació, recepció i representació de la literatura medieval*, coord. Meritxell Simó, ed. G. Avenoza, A. Contreras, G. Sabaté, L. Soriano. Cilengua: San Millán de la Cogolla, 2021, p. 149-166.

Bazzaco et al. 2022 — Stefano Bazzaco — Mónica Martín Molares — Ana Milagros Jiménez Ruiz — 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*, p. 67-125. <https://historiasfingidas.dlls.univr.it/article/view/1190>.

Bognolo 2012 — Anna Bognolo. *El libro español en Venecia en el siglo XVI*. In: *Rumbos del hispanismo en el umbral del Cincuentenario de la AIH*, coord. Patrizia Botta, Aviva Garribba, María Luisa Cerrón Puga, Debora Vaccari, vol. 3 (III. Siglo de Oro (prosa y poesía)). 2012, p. 243-258.

Bognolo 2017 — Anna Bognolo. *Entre Celestinas, novela sentimental y libros de caballerías. La empresa editorial de los Nicolini da Sabbio y Juan Bautista Pederzano en Venecia alrededor de 1530*. In: *Serenísima palabra. Actas del X Congreso de la Asociación Internacional Siglo de Oro (Venecia, 14-18 de julio de 2014)*. Venezia: Edizioni Ca'Foscari, 2014, p. 727-738.

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

Bognolo-Bazzaco 2024 — Anna Bognolo — Stefano Bazzaco. *Editar libros de caballerías en la era digital: la Biblioteca Digital del Proyecto Mambrino*. In: *Editar el Siglo de Oro en la era digital*, ed. E. Fosalba, S. Allés Torrent. Studia Aurea Monográfica, n. 9, p. 19-48.

Bognolo-Cara-Neri 2013 — Anna Bognolo — Giovanni Cara — Stefano Neri. *Repertorio delle continuazioni italiane ai romanzi cavallereschi spagnoli. Ciclo di Amadis di Gaula*. Roma: Bulzoni Editore, 2013.

Chagué-Clérice 2020-2023 — *HTR-United* (2020-2023), ed. by Chagué, Alix — Thibault Clérice. HTR-United Catalog. <https://github.com/HTR-United/htr-united/>.

Ciotti 2023 — Fabio Ciotti. *Minerva e Il Pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale*. «*Testo E Senso*», (dicembre 2023), n. 26, p. 289-15. <https://doi.org/10.58015/2036-2293/671>.

Cordell-Smith 2018 — Ryan Cordell — David Smith. *Report: A Research Agenda for Historical and Multilingual Optical Character Recognition*. Northeastern University Library. <http://hdl.handle.net/2047/D20297452>.

Heigl 2019 — Elisabeth Heigl. *Transcription Guidelines*. «*Rechtsprechung Im Ostseeraum: Digitization & Handwritten Text Recognition*» (blog), 2019. <https://rechtsprechung-im-ostseeraum.archiv.uni-reifswald.de/transcriptionguidelines/>.

Hodel et al. 2021 — Tobias Hodel — David Schoch — Christa Schneider — Jake Purcell. *General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example*. «Journal of Open Humanities Data», 7 (July 2021). <https://doi.org/10.5334/johd.46>.

Malatesta 2023 — Serena Malatesta. *(Proto)modello di trascrizione automatica per i manoscritti danteschi in Littera Textualis (XIV-XV sec.) con Transkribus*. «Dig*Italia*. Rivista del digitale nei beni culturali», 18 (2023), n. 2, p. 282–285. <https://digitalia.cultura.gov.it/article/view/3016>.

Mancinelli 2016 — Tiziana Mancinelli. *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», (2016), n. 4, p. 255-260.

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

Pavlopoulos et al. 2022 — John Pavlopoulos et al. *Error Correcting HTR'ed Byzantine Text*. «*HTREC* 2022» p. 1-16. <http://dx.doi.org/10.21203/rs.3.rs-2921088/v1>.

Perdiki 2023 — Elpida Perdiki. *Preparing Big Manuscript Data for Hierarchical Clustering with Minimal HTR Training*. «Journal of Data Mining and Digital Humanities. Special Issue: Historical documents and automatic text recognition»*, 2023*. <https://hal.science/hal-03880102v4>.

Pinche 2023 — Ariane Pinche. *Generic HTR Models for Medieval Manuscripts. The CREMMALab Project*. «Journal of Data Mining and Digital Humanities. Special Issue: Historical documents and automatic text recognition», 2023. <https://hal.science/hal-03837519v3>.

Rabus 2019 — Achim Rabus. *Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus*. «Scripta & e-Scripta», (2019), n. 19, p. 9–32.

Roncaglia 2023 — Gino Roncaglia. *L´architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT*. Bari: Laterza, 2023.

Sayre 1973 — Kenneth M. Sayre. *Machine recognition of handwritten words: A project report*. «Pattern Recognition», 5 (1973), n. 3 p. 213-228. <https://doi.org/10.1016/0031-3203(73)90044-7>.

Schwarz-Ricci 2022 — Vera Isabel Schwarz-Ricci. *Handwritten Text Recognition per registri notarili (secc. XV-XVI): una sperimentazione*. «Umanistica Digitale», (2022), n. 13, p. 171-181. <https://doi.org/10.6092/issn.2532-8816/14926>.

Spina 2022 — Salvatore Spina. *Historical Network Analysis and HTR tools for a digital methodological historical approach to the Biscari Archive of Catania*. «Umanistica Digitale», (2022), n. 14, p. 163–181. <https://doi.org/10.6092/issn.2532-8816/15159>.

Stokes et al. 2021 — Peter A. Stokes — Benjamin Kiessling — Daniel Stökl Ben Ezra — Robin Tissot — El Hassane Gargem. *The EScriptorium VRE for Manuscript Cultures.* «Classics@ Journal, Ancient Manuscripts and Virtual Research Environments» ed. C. Clivaz e G. V. Allen, 18 (2021). <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/>.

Stokes-Kiessling 2024 — Peter A. Stokes — Benjamin Kiessling. *Sharing Data for Handwritten Text Recognition (HTR)*. «Digital Humanities in Practice», In corso di stampa. ffhal-04444641 <https://hal.science/hal-04444641/document>.

Terras 2022 — Melissa Terras. *Chapter 7: Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription*. In: *Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections*, ed. by L. Jaillant. Bielefeld University Press, 2022, p. 179-204. <https://doi.org/10.14361/9783839455845-008>.

Vinciarelli 2003 — Alessandro Vinciarelli. *Offline Cursive Handwriting: From Word To Text Recognition*. IDIAP Research Record 03-24, 2003. <https://infoscience.epfl.ch/record/82879>.

Zappulli-Iorio 2018 — Andrea Zappulli — Sabrina Iorio. *La digitalizzazione dell'Archivio Storico del Banco di Napoli*. «Dig*Italia*. Rivista del digitale nei beni culturali», 13 (2018), n. 2 p. 46-51. <https://digitalia.cultura.gov.it/article/view/2169>.