Transkribus e la trascrizione automatica di manoscritti. Un caso studio sulle carte di Vittorio Alfieri

«DigItalia» 2-2025
DOI: 10.36181/digitalia-00147

Sara Gallegati
Università degli studi di Macerata

Lo studio esplora l'uso della piattaforma Transkribus per la trascrizione automatica di manoscritti, utilizzando i testi inclusi nel manoscritto Alfieri 6 (conservato presso la Biblioteca Medicea Laurenziana di Firenze). Dopo una prima fase di test con il modello generico Italian Handwriting M1, i cui risultati si sono dimostrati insoddisfacenti, è stato creato un modello personalizzato a partire dal modello M1, che ha ottenuto un CER del 13.40%. Un secondo addestramento, ampliando il set di ground truth, ha ridotto il CER al 6.9%, con trascrizioni più accurate. Nonostante i progressi, permangono alcune difficoltà nella distinzione tra lettere simili; ma lo studio conferma l'efficacia di Transkribus per la trascrizione di testi manoscritti, suggerendo l'importanza di un numero adeguato di dati per migliorare la qualità della trascrizione automatica.

Introduzione

Transkribus è una piattaforma basata sull'intelligenza artificiale e sulle tecnologie di Handwritten Text Recognition (HTR); consente alle macchine di riconoscere testi manoscritti e a stampa, convertendo le immagini in trascrizioni codificate, ricercabili ed esportabili in vari formati.

L'idea alla base della piattaforma è che «the collaboration between (1) archives/libraries, (2) humanities scholars, (3) computer scientists/technology providers and (4) the public (crowd users, volunteers) is key for the success of an innovative platform which aims at substantially improving access to historical documents on the long term»[1].

Uno dei maggiori vantaggi di Transkribus è il suo essere *user-friendly*: gli utenti possono gestire l'intero processo di digitalizzazione senza bisogno di una formazione informatica specifica, possono inoltre creare i propri modelli di riconoscimento di scrittura a mano o addestrarne di già esistenti[2].

[1] Sebastian Colutto — Philip Kahle — Günter Hackl — Günter Mühlberger, *Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents*, «IEEE Computer Society» (2019), p. 463-466: 463. DOI: 10.1109/eScience.2019.00060.

[2] I modelli pubblici disponibili (allo stato attuale 140, <https://readcoop.eu/transkribus/public-models/>) consentono agli utenti di far analizzare i documenti manoscritti e a stampa senza bisogno di addestramento personalizzato.

I dati e i documenti caricati dagli utenti su Transkribus sono privati, il che riduce le restrizioni imposte dal copyright[3], e vengono archiviati nella piattaforma in modalità centralizzata, in modo da costituire un ampio bacino che può essere utilizzato per tutti i processi di apprendimento automatico.
La piattaforma permette inoltre di condividere tra gli utenti collezioni, documenti e prototipi di HTR per l'addestramento di modelli più ampi: «users of Transkribus are therefore not only consumers but also producers of their own model(s) and data which can then also be re-used by and shared with other users»[4].
Ad oggi, le tecnologie di riconoscimento di testi scritti a mano (HTR) hanno raggiunto un elevato grado di maturità e sono in grado di produrre trascrizioni precise dalle immagini di manoscritti storici. Grazie all'intelligenza artificiale e all'applicazione delle reti neurali profonde, infatti, le tecniche di HTR hanno registrato notevoli progressi nell'ultimo decennio, e vengono sempre più impiegate da biblioteche e archivi, per accelerare la trascrizione di fonti primarie e facilitare la ricerca e l'analisi di testi storici[5].
Questo strumento è stato applicato per ottenere le trascrizioni del materiale del *Panegirico di Plinio a Trajano* contenuto in Alfieri 6, il manoscritto conservato presso la Biblioteca Medicea Laurenziana di Firenze (Fondo Alfieri, [Cat. Sala Studio 6][6]). Trattasi della più antica attestazione del testo a noi giunta: l'opera fu ideata e stesa da Vittorio Alfieri[7] a Pisa tra l'inverno e l'estate del 1785, poi rivista durante il soggiorno a Martinsbourg, nel 1786, quindi stampata una prima volta a Parigi nell'aprile del 1787 per i tipi di François-Denis Pierres, e nuovamente, dopo un'ulteriore revisione, nell'autunno del 1789, da Didot, questa volta assieme a due altre opere, l'ode *Parigi sbastigliato* e la favoletta *Le mosche e l'Api*. Alfieri 6 ospita – oltre al *Panegirico* alle cc. 29*r*-38*r* – i materiali relativi alla *Tirannide*, recante ancora il titolo *Del Tiranno e Della Tirannide*

[3] «All documents uploaded to Transkribus are private by default. They are stored on the servers of READ-COOP SCE (i.e. the company that develops and maintains the software). The servers are all located in Innsbruck, Austria, in a GDPR-compliant manner, and the data may be processed according to the *terms & conditions* on the READ-COOP SCE website» Cfr. *Transkribus Help Center*, <https://help.transkribus.org/uploading-files-to-transkribus>.

[4] S. Colutto *et al*., *Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents*, cit., p. 463.

[5] Joe Nockels — Paul Gooding — Sarah Ames — Melissa Terras, *Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research*, «Archival Science», 22 (2022), p. 367-392: 368.

[6] Sul fondo dell'autore conservato alla Laurenziana cfr. Simone Casini, *Il fondo di manoscritti alfieriani della Laurenziana. Appunti per una storia interna*, in: *Alfieri a Siena e dintorni*. *Omaggio a Lovanio Rossi*. *Atti della Giornata di Studi, Colle di Val d'Elsa (22 settembre 2001)*, a cura di A. Fabrizi, Roma: Domograf, 2007, p. 175-193.

[7] Per una panoramica su Alfieri, l'opera e il relativo contesto storico-culturale, si vedano almeno Vittore Branca, *Sbastigliamenti alfieriani fra miti solari e fede palingenetica, delirio pindarico e autobiografia poetica*, «Italica» 68 (1991), n. 4, p. 401-418; Guido Santato, *Le Mosche sul Panegirico: Alfieri "sbastigliato"*, «Lettere italiane», 46 (1992), n. 1, p. 57-92; Giuseppe Rando, *Alfieri europeo: le "sacrosante leggi". Scritti politici e morali – Tragedie – Commedie*, Soveria Mannelli: Rubbettino, 2007; *Quand Alfieri écrivait en français: Alfieri et la culture française*, a cura di C. Del Vento, G. Santato, Paris: Bibliothèque Mazarine, 2003; Christian Del Vento, *Il Principe e il Panegirico. Alfieri tra Machiavelli e De Lolme*, «Seicento & Settecento», 1 (2006), p. 149-170; Laura Sannia Nowé, U*na* institutio principis *moderna: il* Panegirico di Plinio a Trajano *di Vittorio Alfieri*, in: *Studi dedicati a Gennaro Barbarisi*, a cura di C. Berra, M. Mari, Milano: Cuem 2007, p. 489-526.

(cc. 3*r*-26*r*); il dialogo *La Virtù Sconosciuta* (cc. 41*r*-50*r*); il trattato *Del Principe e Delle Lettere* (cc. 51*r*-89*r*); le *Prose diverse pel Misogallo. Prose cinque* (cc. 91*r*-112*r*). L'opera nasce, secondo quanto narrato nell'autobiografia[8], dalla lettura del panegirico che Plinio il Giovane recitò all'imperatore Traiano in senato nel 100 d.C.; e si configura come un'orazione il cui nucleo centrale ruota attorno alla richiesta di Plinio rivolta a Traiano di deporre volontariamente il potere assoluto e ristabilire la libertà repubblicana a Roma.
I vantaggi offerti dalla piattaforma Transkribus per lo studio di manoscritti sono molteplici: dopo una prima fase di preparazione del materiale si riducono drasticamente i tempi di trascrizione; è possibile aggiungere dei metadati ai testi così da arricchire le informazioni all'interno dei file; i documenti sono inoltre esportabili in diversi formati, tra cui in xml, aspetto fondamentale per permettere una successiva codifica secondo le linee guida TEI.

Flusso di lavoro per il ms. Alfieri 6

Il flusso di lavoro per la trascrizione del manoscritto si è composto di tre fasi: 1) creazione di una collezione e caricamento del materiale; 2) analisi del layout, selezione e verifica di un modello, *text recognition*; 3) operazioni di postproduzione, in particolare controllo e correzione delle trascrizioni ottenute.
La scelta di trascrivere Alfieri 6 con Transkribus non è dovuta tanto alla complessità grafica del manoscritto – la scrittura di Vittorio Alfieri, infatti, risulta piuttosto regolare e leggibile, al punto da non richiedere competenze paleografiche specialistiche – quanto alla possibilità di sfruttare questo caso come punto di partenza per la creazione di un modello HTR dedicato e altamente specifico. L'obiettivo è quello di sviluppare un modello in grado di riconoscere con precisione la mano di Alfieri, che potrà essere progressivamente raffinato e potenziato con l'inserimento di nuovo materiale autografo. In questo modo, il lavoro non solo ottimizza la trascrizione di Alfieri 6, ma pone le basi per una risorsa condivisibile e riutilizzabile da altri studiosi, favorendo una più ampia diffusione e valorizzazione digitale del *corpus* alfieriano.
Dopo la creazione delle collezioni sono stati caricati i file contenenti le carte del manoscritto. La piattaforma permette il caricamento di file in diversi formati: sono accettati sia file di immagini nei formati jpeg e png, con una dimensione massima di 10 MB ciascuno, che file in pdf fino a 3.000 pagine e con una dimensione massima di 200 MB. Per Alfieri 6 è stato caricato un unico file pdf, ogni pagina del documento è stata poi estratta e caricata automaticamente come singola. La buona qualità delle immagini non ha reso necessario alcun trattamento prima della trascrizione automatica.

[8] «In queste semiletture avea scorse le lettere di Plinio il Minore [...]. Finite l'epistole, impresi di leggere il Panegirico a Traiano, opera che mi era nota per fama, ma di cui non avea mai letta parola. Inoltratomi per alcune pagine, e non vi ritrovando quell'uomo stesso dell'epistole, e molto meno un amico di Tacito [...], io sentii nel mio intimo un certo tal moto d'indegnazione; e tosto, [...] impugnata con ira la penna, ad alta voce gridando dissi a me stesso: "Plinio mio, se eri tu davvero e l'amico, e l'emulo, e l'ammiratore di Tacito, ecco come avresti dovuto parlare a Traiano". E senza più aspettar, né riflettere, scrissi d'impeto, quasi forsennato, così come la penna buttava, circa quattro gran pagine del mio minutissimo scritto» (Vittorio Alfieri, *Vita*, in: *Opere*, Introduzione e scelta di Mario Fubini, testo e commento a cura di Arnaldo Di Benedetto, Milano-Napoli: Ricciardi, 1977, v. I, p. 251).

È inoltre possibile editare le collezioni create, inserendo metadati descrittivi sulle base delle regole xml TEI[9]. Le informazioni aggiunte possono essere poi esportate, assieme ai contenuti della collezione, in vari formati[10], tra cui xml codificato in TEI e ALTO (Analyzed Layout and Text Object[11]). La pagina esportata nel formato xml registra le informazioni relative alle caratteristiche dell'immagine, oltre che alla struttura del layout e al contenuto della pagina[12]; la pagina esportata in ALTO presenta uno schema XML che dettaglia i metadati tecnici per descrivere il layout e il contenuto delle risorse testuali fisiche, e viene spesso utilizzato in combinazione con METS.
Dopo aver creato ed editato la collezione, l'analisi si è concentrata sulle singole pagine. È stato quindi effettuato un censimento dei modelli presenti nella piattaforma per testi manoscritti in lingua italiana. I modelli proposti sono tre: *Transkribus Italian Handwriting M1*, *Italian Administrative Hands, 1550-1700* e *Ligorio 0.3 Pyl*. *Ligorio 0.3 Pyl* (CER[13] 2.7%), basato sulla grafia dell'antiquario italiano Pirro Ligorio (Napoli 1513 – Ferrara 1583)[14]. *Italian Administrative Hands, 1550-1700* (CER del 12.2%) è costituito da documenti in lingua italiana provenienti dagli archivi di Stato di Milano, Venezia, Firenze, Pisa e Genova: «the training set represents a spectrum of humanistic, italic and cursive hands characteristic of administrative records, employed by secretaries and newswriters»[15].
L'eccessiva specificità dei modelli appena descritti ha vincolato la scelta al modello *Transkribus Italian Handwriting M1* per un primo test di trascrizione. Il modello, che presenta un tasso di errore dei caratteri (CER) del 6.7%, è addestrato su testi di grafia italiana compresi tra il XVI e il XIX secolo. Tuttavia, «this is a generic model trained on a diverse dataset. Such models provide good results without the need for any extra training work. However, the best results can usually be achieved by training a special

[9] XML, acronimo di eXtensible Markup Language è un metalinguaggio creato e gestito dal World Wide Web Consortium (W3C); è una semplificazione e adattamento di SGML, da cui è nato nel 1998, e permette di definire la grammatica di diversi linguaggi specifici derivati. È un linguaggio di marcatura, ossia un linguaggio dotato di una semantica e di una sintassi specifiche in grado di mostrare all'utente tutte le annotazioni di caratterizzazione. I testi codificati in XML possono essere poi utilizzati sia per creare *output* orientati alla visualizzazione, sia come base di dati per effettuare delle elaborazioni automatiche. Nell'ambito dell'informatica umanistica, il consorzio TEI (Text Encoding Initiative, <https://tei-c.org>) ha messo a punto, a partire dal linguaggio XML, un vasto e complesso schema di codifica, proponendo una DTD (Document Type Definiton) per tutti i fenomeni dei testi umanistici. La sua finalità «è di definire uno standard di codifica, specificamente orientato alla gestione dei dati umanistico-letterari, e realizzare una normalizzazione dei formati di memorizzazione dell'informazione testuale, al fine di consentire l'interscambio dei documenti» (Francesca Tomasi, *Metodologie informatiche e discipline umanistiche*, Roma: Carocci, 2008, p. 132).

[10] Oltre a xml e ALTO è possibile esportare le immagini in formato jpeg e png o come file METS (Metadata NEcoding and Trasmission Standards) o creare documenti pdf, docx e xlsx.

[11] ALTO è uno standard sviluppato per la descrizione dei testi OCR e delle informazioni di *layout* delle pagine digitalizzate.

[12] https://gitlab.com/readcoop/transkribus/TranskribusCore/-/blob/master/src/main/resources/xsd/pagecontent_extension.xsd.

[13] In Transkribus, CER sta per Character Error Rate, ovvero tasso di errore sui caratteri. Si tratta di una metrica fondamentale per valutare l'accuratezza di un modello HTR (Handwritten Text Recognition). Cfr. <https://help.transkribus.org/it/tasso-di-errore-del-carattere-e-curva-di-apprendimento>.

[14] *Ligorio 0.3 Pyl*, in *Public AI Models in Transkribus*, <https://readcoop.eu/model/ligorio-0-3/>.

[15] *Italian Administrative Hands, 1550-1700*, in *Public AI Models in Transkribus*, <https://readcoop.eu/model/italian-administrative-hands-1550-1700/>.

model for homogenous material, e. g. texts written by the same person or from a narrow historical period»[16].
Il modello si è mostrato inefficace nel trascrivere i testi anche nei casi in cui le immagini sottoposte a text recognition presentavano poche righe di scrittura e un'interlinea ben distanziata, come nel seguente esempio:

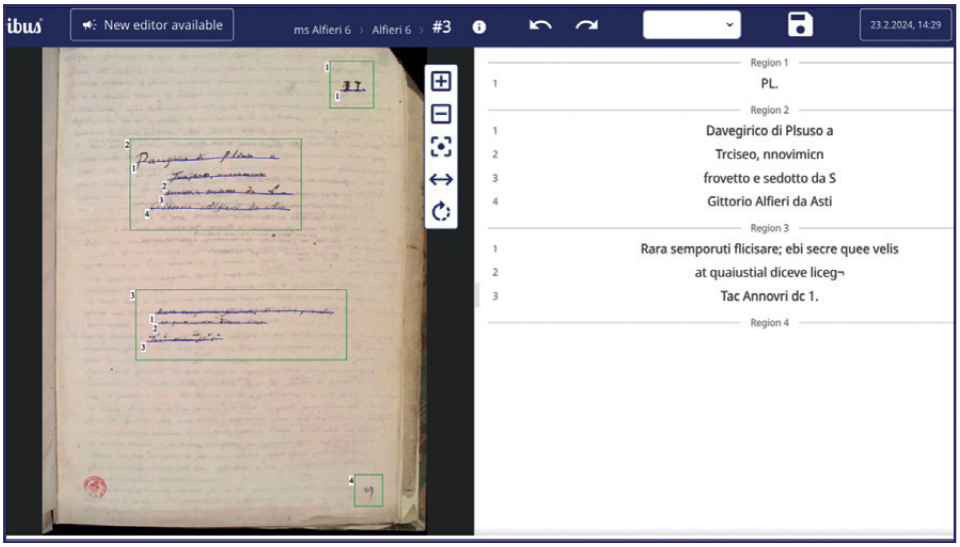

Figura 1. *Esempio di trascrizione di Alfieri 6 con modello Transkribus Italian Handwriting M1 della c. 29r. Firenze, Biblioteca Medicea Laurenziana, Ms. Alfieri 6, c. 29r. Su concessione del MiC. È vietata ogni ulteriore riproduzione con qualsiasi mezzo*

Le cause del mancato successo nel processo di trascrizione automatica possono probabilmente essere attribuite a una serie di fattori legati alla grafia dell'autore. Pur non essendo particolarmente difficile da leggere per un occhio umano, la scrittura di Alfieri presenta una serie di variabilità formali che la rendono meno adatta a una lettura automatica standardizzata. Non si tratta infatti di una grafia "standard" nel senso tecnico richiesto dai modelli HTR[17], ovvero caratterizzata da tratti ripetitivi, regolarità formale e coerenza morfologica tra le lettere. Al contrario, Alfieri mostra una scrittura soggetta a oscillazioni grafiche, abbreviazioni idiosincratiche e modifiche successive sul testo (come cancellature, sovrascritture, integrazioni a margine), tutti elementi che introducono rumore visivo e rendono più complesso l'addestramento del modello. Inoltre, la stessa disposizione del testo all'interno

[16] *Transkribus Italian Handwriting M1*, in *Public AI Models in Transkribus*, <https://readcoop.eu/model/italian-general-model/>.

[17] La "grafia standard" in Transkribus è una forma normalizzata di scrittura che fornisce il modello di riferimento per la trascrizione dei testi. Non esiste una sola "grafia standard": «There does not exist a general model for all the handwritings [...]. When choosing a text model, you need to consider the following: 1) the type of material, handwritten or printed; 2) the language; 2) the period; 3) the type of script; 4) the Character Error Rate (CER)» (Cfr. <https://help.transkribus.org/choosing-a-model>).

del manoscritto – talvolta disomogenea, con variazioni nella spaziatura e allineamenti irregolari – può compromettere l'efficacia del *layout analysis*, fase preliminare fondamentale per una corretta segmentazione delle righe e delle parole.
Si è pertanto deciso di seguire le indicazioni fornite da Transkribus, addestrando un modello privato, composto da materiale omogeneo, e stabilendo *Transkribus Italian Handwriting M1* come modello di base. Sebbene una delle *key-features* di Transkribus sia quella di permettere la creazione di modelli personalizzati e privati, generarne uno senza una base di riferimento è un'operazione molto complessa e poco significativa se non si ha un elevato numero di carte su cui addestrare il sistema: i modelli pubblici esistenti possono quindi essere utilizzati come punti di partenza per ridurre la quantità di nuovi dati necessari.
Il *corpus* di riferimento individuato è quello contenente i testi del *Panegirico di Plinio a Trajano* (cc. 29*r*-38*r*) e della *Virtù Sconosciuta* (cc. 41*r*-50*r*). Per avere un adeguato numero di carte e di parole su cui addestrare il modello, infatti, il solo materiale del *Panegirico* non è quantitativamente sufficiente[18]: si è resa pertanto necessaria l'aggiunta di materiale per ottenere un numero valido di dati su cui addestrare il modello. La scelta è ricaduta sulle carte della *Virtù Sconosciuta* per diverse ragioni. Anzitutto, le stesure del *Panegirico* e della *Virtù* contenute in Alfieri 6 risalgono a periodi contigui[19] e presentano delle grafie del tutto simili, sebbene il dialogo sia caratterizzato da un numero maggiore di correzioni e di interventi dell'autore; le opere presentano inoltre analoghe strutture di impaginazione. La relativa omogeneità nella grafia e nell'impaginazione delle pagine rende il materiale adeguato alla produzione di un modello di grafia dell'autore. Del materiale scelto, una parte delle carte del *Panegirico* è stata utilizzata per costruire la *ground truth*[20] e la restante è stata sottoposta alla *text recognition* mediante il modello personalizzato; le carte della *Virtù Sconosciuta*, invece, hanno fornito la restante parte della *ground truth* per allenare il modello.
Sono quindi state selezionate le carte per formare la *ground truth* per un totale di 13 sulle 45 disponibili[21]. L'analisi del layout è stata eseguita automaticamente, prendendo come modello *Universal Lines* e modificando eventuali imprecisioni su regioni di testo e *baseline*. I risultati dell'analisi automatica del layout si sono rivelati soddisfacenti, e il numero di interventi manuali nella segmentazione del testo è stato minimo. In questo caso, all'effi-

[18] Le linee guida Transkribus raccomandando almeno 10.000 parole per addestrare il modello di un manoscritto attribuito ad una sola mano di scrittura. «In the case of handwritten documents, our advice is to train the model on at least 10,000 words for each hand» (Cfr. *Transkribus Help Center*, <https://help.transkribus.org/data-preparation>).

[19] Stando alle annotazioni poste da Alfieri nel manoscritto, la stesura del *Panegirico* risale al marzo 1785, quella della *Virtù Sconosciuta* al gennaio 1786. Nel *Rendimento di conti* l'autore annota: «1785. Nel Marzo, ideato, e scritto d'un fiato, il *Panegirico*»; «1786. Nel Gennaio […] steso il dialogo della *Virtù Sconosciuta*» (V. Alfieri, *Rendimento di conti da darsi al Tribunal d'Apollo*, in *Opere*, v. I, cit., p. 434).

[20] La *ground truth* è la trascrizione manuale corretta di un testo manoscritto, usata come base di confronto per addestrare e valutare i modelli HTR. L'insieme di dati della *ground truth* è cruciale, in quanto i modelli di machine learning operano attraverso la replicazione statistica dei pattern forniti in fase di addestramento. Di conseguenza, la qualità e l'affidabilità della *ground truth* influenzano in maniera determinante le prestazioni e l'accuratezza del modello risultante. Cfr. anche <https://blog.transkribus.org/en/what-is-ground-truth>.

[21] Nello specifico, le carte selezionate dal *Panegirico* per costituire la *ground truth* sono: cc. 30*r*-30*v*; 31*r*-31*v*; 32*r*; 33*v*; 35*r*; 36*v*. Per la *Virtù Sconosciuta*, le carte selezionate sono: cc. 44*v*; 45*r*-*v*; 46*r*; 47*r*.

cienza di Transkribus si somma il fatto che il manoscritto oggetto di studio è una con tutta probabilità una copia in pulito della prima stesura, non presentando segni di incertezza nella struttura e nella composizione, così come nella scrittura, minuta, regolare e densa[22].
I testi delle carte scelte per la *ground truth* sono stati quindi sottoposti ad una *text recognition* mediante il modello pubblico *Transkribus Italian Handwriting M1*; le trascrizioni ottenute sono state poi corrette manualmente. La *ground truth* «is crucial for the training of probabilistic HTR models since it is the basis of their learning, that is, the training sample. For this reason, GT [*ground truth*] must be as precise as possible to obtain useful results»[23].
Si è poi proceduto all'addestramento del modello: per il *training set* sono state selezionate le 13 pagine del documento contenenti la *ground truth*, il *validation set* è invece formato dal 10% del set delle carte di training. Per il riconoscimento automatico della scrittura manoscritta è stato scelto il motore PyLaia, noto per la sua accuratezza nei compiti di Handwritten Text Recognition. Come *base model* è stato utilizzato *Transkribus Italian Handwriting M1*, già addestrato su una varietà di scritture italiane. A partire da questo, è stato avviato il processo di generazione di un modello personalizzato. In condizioni di carico server regolari, l'addestramento di un *training set* di 13 pagine con PyLaia ha richiesto circa due ore. Al termine dell'elaborazione, il sistema ha inviato una notifica di completamento tramite email.
Terminato l'addestramento, il software ha prodotto un modello con CER del 13.40%. Nello specifico, il modello presenta un CER del 13% sul *training set* e di 13.8% sul *validation set*.

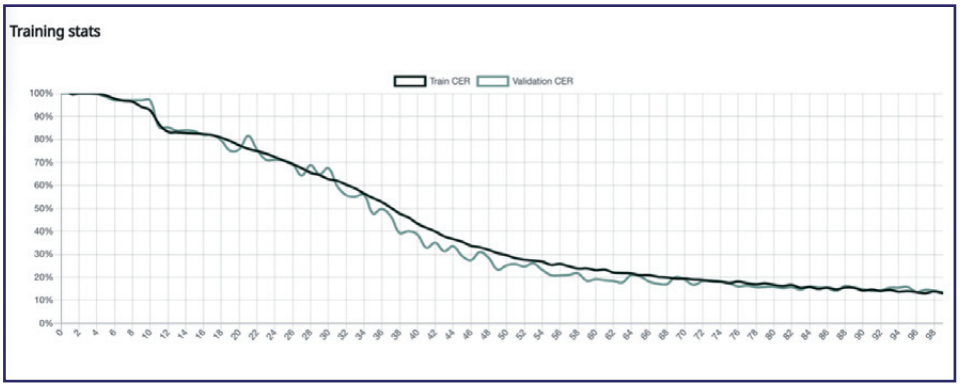

Figura 2. *Learning curve del modello* Alfieri

È stata quindi generata la trascrizione automatica di una delle carte (c. 32*v*, Fig. 4) con il modello appena creato, il risultato è poi stato confrontato con quello precedentemente ottenuto dal modello pubblico *Italian Handwriting M1*. Il modello addestrato sulle carte si è mostrato molto più performante, nonostante il CER più elevato: si propone di se-

[22] V. Alfieri, *Panegirico di Plinio a Trajano. Parigi sbastigliato. Le Mosche e l'Api*, cit., p. 18.
[23] Pepita Raventós Pajares — Celio Hernández Tornero — Simon Martin Meritxell, *AI and archive. Handwritten Text Recognition Applied to Patrimonial Holdings: An Example of 10 diaries written by Spanish Republican Teachers in 1932*, «IEEE Computer Society» (2022), p. 2572-2577: 2573-2574.

guito la stessa sezione di testo trascritta prima con il modello pubblico, poi con quello addestrato, infine il testo corretto.

Text recognition della c. 32*v* con il modello *Italian Handwriting M1*:
9 Dai meviti postri ragioni dicootanera per cui indorro il debbi a vessistno liberrà, av
10 con lieve cn pare posga e porratintina sa, e da baste quali sola; il desidevarla
11 sersomente csi dogtij e que dboto a meritarla questo parvesi sere E nsa credere gia
12 o nel su libemi altra intrasache dobbesir mopre a Frasaro, noe alle logg si
13 c poli gi diserore; ma che cessando egli porsono nella proveri daltos porenso al

Text recognition della c. 32*v* con il modello *Alfieri* (Fig. 2):
9 Dai meriti nostri cagioni dimostrarti per cui indurre ti debbi a restitaione libertà, non è
10 cosi lieve; ma pure prima, e potentisima sia, e da bastar quasi sola; il desiderarla
11 andantemente noi tutti; e gran titolo a meritarla questo parer ti dere. E non credere già,
12 che de nnel dir libertà altro intenda, che d'obbedir sempre a Trajano; sioè alle leggi, di
13 sui eggli, è difensore; ma, che, assando egli, possono nella pertena daloro potente al

Testo corretto manualmente (in grassetto le parole corrette rispetto al modello *Alfieri*):
9 Dei meriti nostri cagioni dimostrarti per cui indurre ti debbi a **restituire** libertà, non è
10 **così** lieve; ma pure prima, e **potentissima** sia, e da bastar quasi sola; il desiderarla
11 **ardentemente** noi tutti; e gran titolo a meritarla questo parer ti **deve**. E non credere già,
12 che io **nel** dir libertà altro intenda, che d'obbedir sempre a Trajano; **cioè** alle leggi, di
13 cui **egli** è difensore; ma, che, **cessando** egli, possono nella **persona d'altro** potente al

Pur rilevando un significativo miglioramento rispetto alla trascrizione con il modello pubblico M1, si è deciso di addestrare un nuovo modello aumentando il numero di fogli di *ground truth* da 13 a 24: il maggior numero di dati a disposizione per l'addestramento dovrebbe infatti fornire esiti più soddisfacenti rispetto a quelli fino ad ora ottenuti.
Definita la nuova *ground truth*, è stato allenato il modello con il motore PyLaia. Per il *training set* sono state selezionate le 24 pagine del documento contenenti le *ground truth*, il *validation set* è stato creato automaticamente sul 10% delle carte del training. È stato poi avviato il lavoro e, terminato l'allenamento, il software ha prodotto un modello con CER del 6.9%; in particolare il *train* CER è del 7.4%, il *validation* CER è del 6.9%:

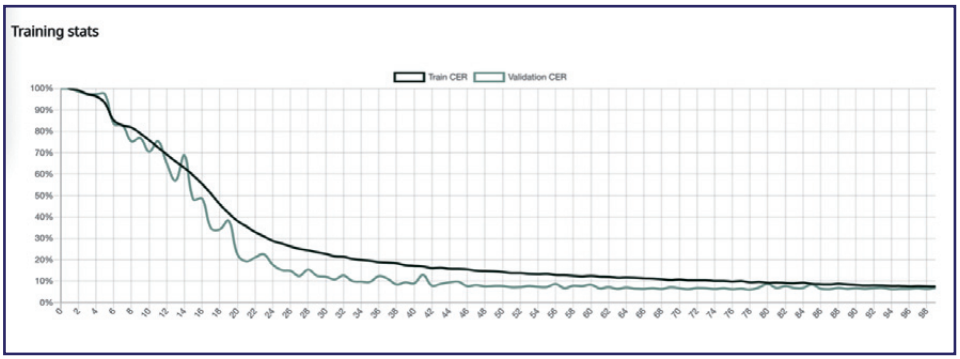

Figura 3. *Learning curve del modello Alfieri 2*

Conclusioni

Stando a quanto disponibile nelle linee guida di Transkribus, «results with a CER of 10% or below can be considered very efficient for automated transcription»[24]. Il valore del set di convalida, inoltre, è il più significativo tra i due, poiché mostra come il modello si comporta nelle pagine su cui non è stato addestrato.
Si propone di seguito la trascrizione con il modello *Alfieri 2* (Fig. 3) della sezione di testo della c. 32*v*. precedentemente analizzata:

9 Dai meriti nostri cagioni dimostrarti per cui indurre ti debbi a restituiue libertà, non e
10 così lieve; ma prure prima, e potentissima sia, e da bestar quasi sola; il desiderarla
11 ordentemente noi tutti; e gran titolo a meritarla questo parer ti deve. e non credere già
12 che de nel dir libertà altro intenda, che d'obbedir sempre a Trajano; cioè alle leggi, di
13 cue eueli è difersore; ma, che, cessando egli, possono nella persena d'aloro potente al

Lo studio condotto sul manoscritto Alfieri 6 conferma l'efficacia di Transkribus per la trascrizione automatica di testi manoscritti, evidenziando al contempo le potenzialità e i limiti delle tecnologie HTR applicate a materiali storici.
L'aumento della *ground truth* e l'addestramento di un secondo modello su un numero di dati più elevato[25] ha prodotto una trascrizione migliore rispetto alla prima: l'incremento da 13 a 24 pagine di training ha prodotto un miglioramento sostanziale delle prestazioni del modello, confermando le raccomandazioni delle linee guida Transkribus. Questo aspetto assume particolare rilevanza per i progetti di digitalizzazione di *corpus* letterari,

[24] Cfr. *Character error rate and learning curve*, <https://help.transkribus.org/character-error-rate-and-learning-curve>.
[25] La *ground truth* del modello *Alfieri* ammonta a 13 pagine con 6.071 parole, quella del modello *Alfieri 2* comprende 24 pagine e 12.148 parole.

suggerendo la necessità di pianificare investimenti adeguati nella fase di preparazione dei dati di addestramento.

Tuttavia persistono alcuni problemi, come la confusione sistematica di alcune lettere: è il caso della *q* e della *g*, della *a* e della *e*, o della *a* e della *o*, che l'autore trascrive in maniera quasi identica. Alla linea # 10, ad esempio, la *a* della parola "bastar" viene identificata erroneamente come *e;* alla linea # 11 si verifica uno scambio della *a* con la *o*, per cui la trascrizione riporta la parola "ordentemente" al posto di "ardentemente".

Per superare questi limiti, è possibile integrare le trascrizioni HTR con strategie di post-correzione automatizzata: la trascrizione prodotta dal modello verrebbe confrontata con dizionari storici o *corpora* dell'italiano settecentesco per identificare parole improbabili o errori tipici; successivamente, algoritmi di correzione (basati su regole, *fuzzy matching* o modelli linguistici statistici) suggerirebbero alternative coerenti, che potrebbero essere approvate o modificate dall'operatore umano; in questo modo si migliorerebbe la qualità delle trascrizioni, riducendo gli errori e senza richiedere correzioni manuali parola per parola.

Inoltre, la creazione di un modello specifico per la mano di Alfieri rappresenta un contributo metodologico che va oltre il singolo caso studio: il modello sviluppato costituisce infatti una risorsa riutilizzabile per future ricerche sul *corpus* alfieriano, il quale potrebbe essere implementato e aggiornato con l'aggiunta di nuovo materiale dell'autore, realizzando uno strumento sempre più efficiente e condivisibile con altri studiosi, favorendo al contempo dinamiche collaborative.

L'approccio adottato potrà infine servire da modello per progetti simili su altri autori, contribuendo alla creazione di un patrimonio digitale di strumenti specializzati per la trascrizione automatica di testi letterari storici.

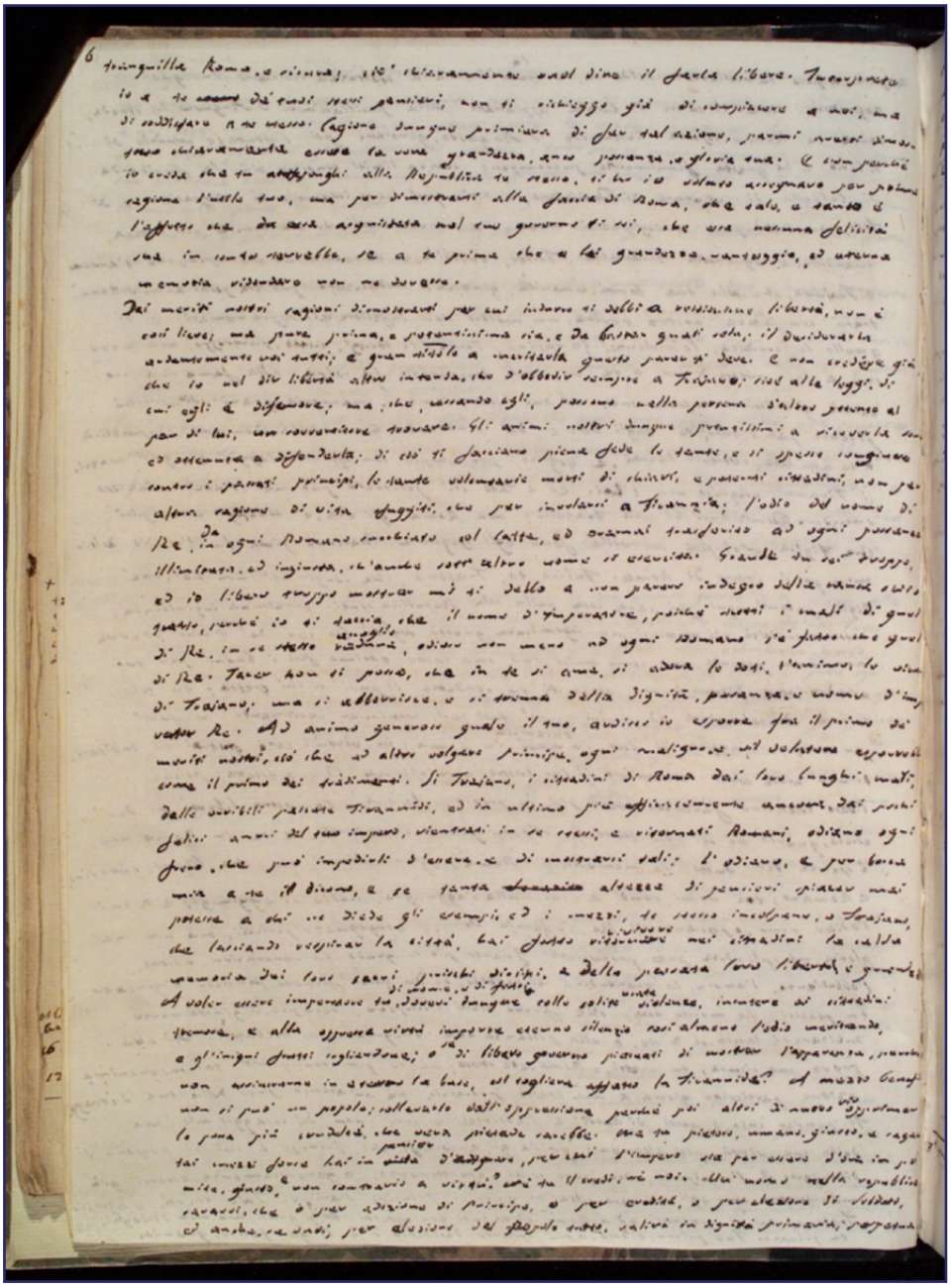

Figura 4. *Firenze, Biblioteca Medicea Laurenziana, ms. Alfieri 6, c. 32v. Su concessione del MiC. È vietata ogni ulteriore riproduzione con qualsiasi mezzo*

The study explores the use of the Transkribus platform for the automatic transcription of manuscripts, using texts included in the Alfieri 6 manuscript (Florence, Biblioteca Medicea Laurenziana). After an initial test phase with the generic Italian Handwriting M1 model, the results of which proved unsatisfactory, a customised model was created from the M1 model, which obtained a CER of 13.40%. A second training, expanding the ground truth set, reduced the CER to 6.9%, with more accurate transcriptions. Despite the progress, some difficulties in distinguishing between similar letters remain, but the study confirms the effectiveness of Transkribus for transcribing handwritten texts, suggesting the importance of adequate data to improve the quality of automatic transcription.

L'ultima consultazione dei siti web è avvenuta nel mese di dicembre 2025.