(Proto)modello di trascrizione automatica per i manoscritti danteschi in Littera Textualis (XIV-XV sec.) con Transkribus

Serena Malatesta
Università degli Studi di Padova - Università degli Studi della Campania "Luigi Vanvitelli"

La crescente quantità di manoscritti digitalizzati e messi a disposizione dalle biblioteche offre la possibilità di usufruire di contenuti di altissima qualità che velocizzano lo studio filologico sui codici. La tecnologia del riconoscimento automatico della scrittura a mano (*Handwritten Text Recognition*, da qui in avanti HTR) consente di tradurre velocemente materiale manoscritto in *machine readable form*, essenziale nei progetti che hanno una vocazione digitale[1].

La scelta di creare un modello HTR per la *Littera Textualis* nei manoscritti della *Commedia* nasce nell'ambito del progetto LiMINA - Lost in Manuscripts, Ideas, Notes, Acknowledgments[2] entro cui si intende realizzare uno spin-off di trascrizione automatica dei codici danteschi, datati o databili entro la cosiddetta antica vulgata. Durante una prima fase di sperimentazione, si è provato a razionalizzare strumenti e pratiche vòlte a trasformare un oggetto digitalizzato (*digitized object*), come file immagine con estensione .jpeg, .png, o .tiff, in un *electronic encoded text* ovvero un testo codificato elettronicamente convertibile in un *electronic text file*, in formato .txt, .doc, .pdf o .xml.

Nell'ambito del *Automatic Text Recognition* (ATR) si predilige l'HTR a discapito dell'OCR (*Optical Character Recognition*), poiché quest'ultimo ha il limite di poter essere applicato, sulla base dell'analisi della forma dei caratteri di un alfabeto preimpostato, senza tener conto degli aspetti linguistici delle parole o delle frasi; ottenendo buoni risultati solo su testi stampati, preferibilmente da immagini di

[1] A titolo d'esempio, faccio riferimento ai lavori di Béatrice Daille — Amir Hazem — Christopher Kermorvant — Martin Maarand — Marie-Laurence Bonhomme — Dominique Stutzmann — Jacob Currie — Christine Jacquin, *Transcription automatique et segmentation thématique de livres d'heures manuscrits*, «Traitement Automatique des Langues», 60 (2019), n. 3, p. 13-36 <https://aclanthology.org/2019.tal-3.2.pdf>, Vera I. Schwarz-Ricci, *Handwritten Text Recognition per registri notarili (secc. XV-XVI): una sperimentazione*, «Umanistica Digitale», 13 (2022), p. 171-181 <https://umanisticadigitale.unibo.it/article/view/14926> e 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)*, Humanidades Digitales y estudios literarios hispánicos, 2022 (Historias Fingidas, Número Especial 1), p. 67-125 <https://historiasfingidas.dlls.univr.it/article/view/1190>.

[2] Sul progetto nato in collaborazione con l'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), recentemente finanziato da un PRIN (Cofin2023), si rimanda al volume miscellaneo *Voci d'inchiostro. Per uno studio dei* Limina *nei manoscritti della Commedia*, a cura di C. Perna, E. Tonello, Padova: libreriauniversitaria.it, 2023 (Storie e Linguaggi, 9).

buona qualità e bi-tonali, dove i singoli caratteri sono facilmente isolabili. L'HTR, invece, riconosce la scrittura nel suo contesto, poiché, grazie all'apprendimento automatico basato sull'addestramento di reti neuronali artificiali (ANNs), utilizza un modello di linguistica statistica sulla base di n-grammi[3] per riconoscere le sequenze di parole. Un altro punto favorevole è che nella maggior parte dei casi non è necessario un pretrattamento delle immagini[4].

Come software di HTR, si è optato per l'utilizzo della piattaforma *Transkribus*[5], rilasciato nel 2015 come parte del progetto *tranScriptorium*, sviluppato dal gruppo DEA (Digitalisierung & Elektronische Archivierung) e finanziato dal programma Horizon2020 e READ (Recognition and Enrichment of Archival Documents), oggi gestita dalla READ-COOP[6], che allo stato attuale si mostra come una delle più complete piattaforme per la trascrizione del testo potenziate dall'AI, con la possibilità di ottenere diversi formati di output, tra cui un file .xml con codifica TEI[7].

Work Flow

La complessità del riconoscimento delle scritture manoscritte medievali risiede nella specificità della forma delle lettere, delle varianti grafiche e nell'uso delle abbreviazioni: per queste ragioni è necessario addestrare sistemi specifici basati sulla grafia e il contenuto testuale.

La community di READcoop dispone già di 129 modelli pubblici per Transkribus creati per varie tipologie di scrittura manoscritta, ma nessuna di queste è stata ritenuta adattabile ai manoscritti della *Commedia* in *littera textualis*. Si è optato, dunque, per l'addestramento di un *Model Data* di riconoscimento specifico.

Il modello è stato addestrato su materiale che è stato acquisito in alta risoluzione dall'URL sorgente del IIIF (International Image Interoperability Framework) manifest[8]. Sono state trascritte per il tag di *Ground Truth*[9] 21 carte da un corpus di 4 manoscritti che per ora si configurano come il campione rappresentativo di diverse realizzazioni di *littera textualis*: Firenze, Biblioteca Medicea Laurenziana, Ashb. 827 (CNMD\0000250010); Città del Vaticano, Biblioteca Apostolica Vaticana, Barb. lat. 4117 (CNMD\0000278058) e Barb. lat. 4112 (CNMD\0000278056); Paris, Bibliothèque de l'Arsenal, 8530 (CNMD\0000276990). Si tratta di codici con *mise en page* simile: struttura a piena pagina a colonna unica, iniziali di terzina sporgenti con datazione che oscilla tra sec. XIV ex e XV in[10]. Questa premessa è necessaria al fine di allenare il modello all'analisi del layout del documento (DLA, Document Layout Analysis) verso una segmentazione del testo per terzine (*text region*) e versi (*base line*), poiché ad ogni porzione di immagine corrisponderà il testo.

[3] Sul concetto di n-gram, si veda William B. Cavnar — John M. Trenkle, *N-Gram-Based Text Categorization*, «Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval», Las Vegas, 1994, <https://dsacl3-2019.github.io/materials/CavnarTrenkle.pdf>.

[4] Un confronto comparativo tra i due sistemi è presente in Christopher Kermorvant, *Automatic Text Recognition - The convergence between OCR and HTR technologies*, 2022 <https://teklia.com/blog/202212-atr/> e Felix Dietrich, *OCR vs. HTR or "What is AI, actually?"*, 2021 <https://readcoop.eu/insights/ocr-vs-htr/>.

[5] https://transkribus.eu/.

[6] https://readcoop.eu/it/.

[7] Ad oggi il migliore standard per le Digital Humanities: <https://tei-c.org/>.

[8] https://iiif.io/community/consortium/.

[9] Ground Truth è un termine utilizzato nel Machine Learning. In Transkribus, viene utilizzato per indicare le immagini e le corrispondenti trascrizioni utilizzate per addestrare l'intelligenza artificiale. Le trascrizioni devono essere il più possibile accurate, perché qualsiasi errore nella Ground Truth addestrerà il modello ad apprendere erroneamente.

[10] Per le descrizioni dei manoscritti si rimanda a Manus Online (<https://manus.iccu.sbn.it/>).

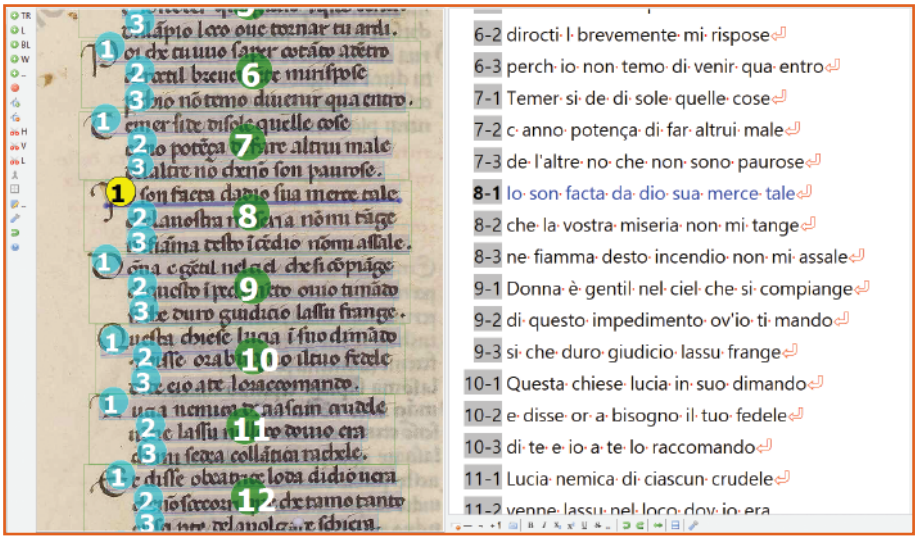

Figura 1. *Interfaccia di Transkribus: a sinistra l'immagine della c.3r del manoscritto Barb.lat.4117 con segmentazione del layout e a destra la trascrizione manuale.*

La trascrizione e la segmentazione delle carte è stata totalmente manuale: sulla base delle linee guida della piattaforma, si è resa una trascrizione diplomatica, con lo scioglimento della *scriptio continua*. Si è scelto altresì di trascrivere le lettere come appaiono nel documento nei casi di i/j, I/J, z/ç ma di rendere secondo l'uso moderno le opposizioni u/v e U/V; la lettera S sia normale che lunga è resa "s".

Risultati preliminari

L'addestramento del modello è stato avviato su base PyLaia[11] con set di validazione del 5% (opzione costretta dall'esiguità delle carte trascritte). Il modello restituisce un tasso d'errore sui caratteri (o CER, che misura il numero di parole errate nella trascrizione fornita dal sistema rispetto alla trascrizione umana) del 1.10% sul corpus di addestramento (*Training set*) e 7.41% sul *dataset* di validazione (*Validation set*) sulla base di 250 *epochs*[12].

Nell'applicazione del modello, si osservano (come in Fig. 3) delle imperfezioni per quanto riguarda la DLA e il riconoscimento delle terzine; una leggera confusione di determinate lettere (u/a, e/c, e/r, s/l, u/v); talora una difficoltà di riconoscimento della "r uncinata" a seguito di applicazione della legge di Meyer e l'errata separazione delle parole. A discapito delle previsioni, lo scioglimento delle abbreviazioni dà buoni risultati, anche se eterogenei in base alla frequenza e alla posizione. Le ipotesi di lettura sono migliorabili ma in questa fase sembrano valide.

A tal proposito si è preferito utilizzare la definizione di "protomodello", poiché chi scrive

[11] PyLaia è un toolkit flessibile e open source, che è utilizzato per condurre una vasta gamma di esperimenti, compresi l'addestramento e l'inferenza su modelli di reti neurali profonde convoluzionali e ricorrenti. PyLaia rappresenta l'evoluzione di Laia, scritto in Lua, basato su PyTorch. <https://github.com/jpuigcerver/PyLaia>.

[12] Un'epoca (*epoch*) rappresenta un ciclo completo in cui il modello impara dai dati di addestramento. Durante ciascuna epoca, il modello viene esposto a tutti i campioni di addestramento e i parametri vengono aggiornati in base all'errore calcolato rispetto alle previsioni desiderate. Il numero ottimale di epoche può variare a seconda delle richieste, delle dimensioni del dataset e dell'architettura del modello.

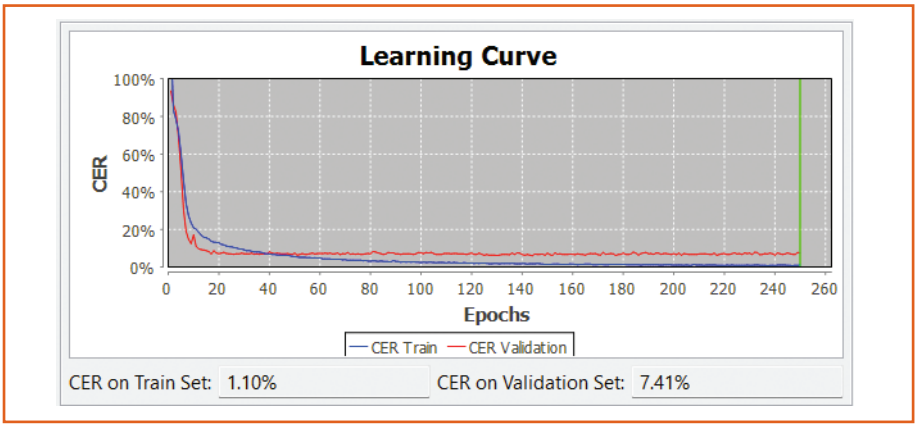

Figura 2. *Curva di apprendimento del modello.*

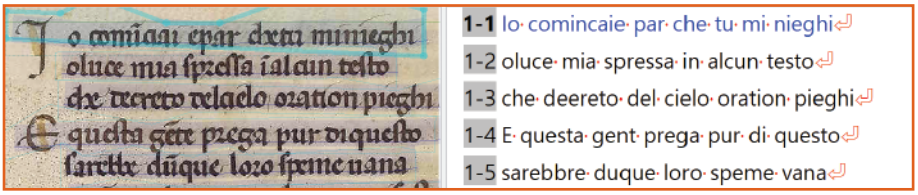

Figura 3. *Applicazione del modello: a sinistra la c.54r del medesimo codice e a destra la trascrizione automatica.*

non ritiene sufficiente la quantità di materiale usato per la *ground truth*. L'allenamento si è basato su un totale di 855 linee per un ammontare di 5.688 parole, mentre le linee guida auspicano una forbice compresa tra le 5.000 e le 15.000 parole. Sarà, dunque, certamente necessario un numero maggiore di trascrizioni per il *training set* per garantire risultati sempre più precisi, ma allo stato attuale sembrerebbe che il modello goda di una base già solida. Con l'intenzione di creare un corpus di dati aperto e a disposizione della ricerca, una volta raggiunti gli standard previsti dalle linee guida, sarà prevista la pubblicazione del modello tra quelli pubblici per Transkribus[13].

[13] https://readcoop.eu/transkribus/public-models/.

L'ultima consultazione dei siti web è avvenuta nel mese di dicembre 2023