DALLA BIBLIOTECA FISICA ALLA BIBLIOTECA DIGITALE

Creazione e applicazione di un modello di riconoscimento del testo HTR per la trascrizione di postillati gioniani

Virginia MELOTTO

ABSTRACT • ***From the Physical to the Digital Library. Creation and Application of a HTR Model for the Transcription of Giono's Annotated Books.*** As part of the author's archive, Giono's physical library is a source of extratextual information that should be taken into account for the interpretation of his late novels, particularly in regards to his sociopolitical views. The volumes containing the reading marks represents, in fact, an intertextual context for the novels, and the presence of numerous political works contrasts with the authorial image associated with a disengagement starting from the end of Second World War. Within the digitalization pipeline, which aims at the publication of the digital edition on the web of a selected number of political texts, the present article focusses on the extraction of machine-readable text from the image files, describing how the transcription process is carried out automatically by the creation and application of a HTR model with Transkribus. We will provide a description of the ground-truth material inserted, the parameters set and the training of the model, the results of multiple trainings as well as examples of the transcriptions. The resulting model is ready to be used for future transcriptions, enabling the efficient digitalization of a great number of volumes from the author's library as well as other documents from his archive.

KEYWORDS • Giono; Digital Library; Transcription; HTR; Marginalia.

1. La biblioteca politica di Jean Giono

L'esegesi di un'opera letteraria novecentesca può trovare conferma alle ipotesi di partenza attingendo a elementi extratestuali che alimentano la produzione di senso. I documenti d'archivio, materiali di natura eterogenea prodotti da un autore o da un'autrice o da lui/lei fruiti nel contesto di redazione dei testi primari, possono rappresentare in questo senso una fonte di informazione indispensabile al processo interpretativo.[1] Nel caso di Jean Giono (1895-1970), la presenza di nu-

[1] Come sottolinea Trevisan, "l'uso di materiali d'archivio ai fini della ricostruzione di una biografia o dell'interpretazione di un'opera ha segnato un forte mutamento nell'indagine critica e nel processo di rilettura dei protagonisti della storia intellettuale, modificando l'approccio metodologico in campo sia archivistico che letterario." Trevisan, M. (2015), *Autoritratti all'inchiostro*, in Albonico, S., Scaffai, N., (dir.), *L'autore e il suo archivio*, Milano, Officina libraria, p. 16.

merosi testi politici letti e annotati a partire dal Secondo dopoguerra collide con l'immagine autoriale di uno scrittore che si disinteressa della politica dopo la sconfitta della lotta pacifista.[2] La presenza di tali volumi, insieme alle considerazioni emerse da recenti studi condotti sull'opera romanzesca alla luce del pensiero sociopolitico dell'autore[3] invita a un approfondimento mirato e a una messa in relazione della biblioteca e dei romanzi successivi al '45, per rivelare nuovi aspetti dell'opera e riflettere sull'immagine autoriale. L'interesse per la biblioteca d'autore, intesa come "parte integrante del fondo archivistico di una personalità letteraria"[4] e quindi reale,[5] trova conferma nelle ricerche più recenti in questo campo.[6] I volumi che compongono la biblioteca possono costituire, infatti, un contesto intertestuale[7] in cui l'opera letteraria si colloca e il loro contenuto può offrire informazioni sul profilo intellettuale dell'autore, in particolare grazie ai segni di lettura verbali e non verbali. Per via di questi segni, che sono da osservare nel contesto in cui sono stati apposti,[8] lo studio di un postillato richiede la consultazione dell'esemplare specifico.

Sottolineata l'importanza di tali documenti, restano i problemi legati alla loro accessibilità, trattandosi di esemplari fisici consultabili unicamente nella sede in cui sono conservati. Nel caso di Giono, i volumi sono consultabili dai ricercatori grazie alla disponibilità dell'*Association des Amis de Jean Giono*.[9] La digitalizzazione e la possibilità di creare edizioni digitali di documenti

[2] Periodizzando le letture politiche, Mény sottolinea questo scarto. Mény, J., *Les lectures politiques de Jean Giono*, articolo di prossima pubblicazione, trasmesso dall'autore, che ricordiamo con affetto.

[3] Labouret, D. (1995), *L'Écriture polémique de Giono*, in Sacotte, M. (dir.), *Giono l'enchanteur*, Paris, Grasset. Schaelchli, É. (2013), *Pour une révolution à hauteur d'hommes, Neuvy-en-Champagne, Le passager clandestin*; Id. (2016), *Jean Giono, le non-lieu imaginaire de la guerre, une lecture de l'œuvre de Giono à la lumière de la «Lettre aux Paysans sur la Pauvreté et la Paix»*, Paris, Eurédit; Melotto, V., *Il sangue dei camaleonti e la follia dei convulsionari. L'engagement di Jean Giono in* Promenade de la mort, in «Studi Francesi», in corso di stampa.

[4] Leonardi, L. (2015), *La funzione-Contini nella cultura del novecento: notizie dal suo archivio*, in Albonico, S., Scaffai, N., cit. p. 57.

[5] La distinzione tra biblioteca reale (fisica o materiale) e virtuale è tratta da Ferrer, D., D'Iorio, P. (2001), *Bibliothèques d'écrivains*, Paris, Cnrs éditions, «Textes et manuscrits».

[6] Lo studio delle biblioteche d'autore ha riscosso negli ultimi anni un notevole interesse in Europa, dove sono stati condotti studi sui volumi e sui segni di lettura, nonché dell'impiego dei dati estratti per l'analisi testuale di opere letterarie, in particolare tra Settecento e Novecento. Citiamo, a titolo di esempio, i contributi del convegno «L'Autore e il suo Archivio», Université de Lausanne, novembre 2013 o la giornata di studi «Manzoni e altri grandi postillatori tra Sette e Ottocento», Università di Parma, 16 avril 2018. Albonico, S., Scaffai, N. (2015), cit., Raboni, G. (cur.) (2018), *Manzoni*
e altri grandi postillatori tra Sette e Ottocento in «Prassi Ecdotiche della modernità letteraria», 3/2018, pp. 5-365.

[7] Come sottolinea Suleiman, in particolare nel quadro dell'analisi dei tratti strutturali e modali del romanzo a tesi, il legame intertestuale può legittimare l'interpretazione di un'opera romanzesca. Data l'assenza di un intertesto dottrinale, i romanzi gioniani successivi al 1945 non sono riconducibili al genere del romanzo a tesi così come definito dalla studiosa. Essi intrattengono, tuttavia, un legame con la biblioteca politica che interviene nella produzione di significato, avendo nutrito le riflessioni dell'autore in fase di redazione. In questo senso, essi costituiscono un contesto intertestuale funzionale all'interpretazione. Suleiman S. R. (2018), *Le Roman à thèse ou l'autorité fictive*, Paris, Classiques Garnier, «Classiques de la littérature» (1983), p. 55 e pp. 65-66.

[8] Nel caso dei marginalia, ad esempio, i commenti dell'autore si riferiscono spesso a uno specifico passaggio, pagina o capitolo. Il riferimento può essere esplicitato dalla presenza di altri segni non verbali come tratti, frecce, sottolineature o asterischi.

cartacei offrono in questo senso vantaggi evidenti: accesso ai volumi a distanza con una riproduzione fedele delle singole pagine con i relativi segni di lettura, nonché possibilità di effettuare ricerche mirate per parole chiave, funzione resa possibile dalla presenza di testo in formato *machine-readable* che è stato precedentemente trascritto. Il percorso che sfocia nella produzione di un'edizione digitale prevede diverse tappe: acquisizione delle immagini, trascrizione del testo, marcatura, visualizzazione, produzione e pubblicazione sul web. Il presente articolo si focalizza sulla fase di trascrizione di alcuni postillati della sezione politica della biblioteca gioniana, attualmente conservati a Manosque, proponendo un approccio che sfrutta un sistema di riconoscimento automatico della scrittura manuale o *handwritten text recognition* (HTR) per la trascrizione dei postillati tramite modello di riconoscimento creato *ad hoc*. L'ambiente di lavoro scelto è *Transkribus*,[10] piattaforma sempre più diffusa nel campo delle Humanities sviluppata a partire dal 2013 dall'Università di Innsbruck nel quadro dei progetti EU *tranScriptorium* e *READ* (*Recognition and Enrichment of Archival Documents*)[11] che offre agli utenti la possibilità di allenare le reti neurali basate sul *machine learning* applicandole ai documenti al fine di generare la trascrizione automatica e di renderli ricercabili.[12] Il passaggio dai tradizionali sistemi di OCR ai sistemi di HTR costituisce un importante passo avanti nel processo evolutivo dei sistemi di trascrizione automatica del testo, iniziato nel secondo Ottocento e nel quale la nascita e l'impiego di sistemi OCR sempre più sofisticati ha contribuito a ottimizzare innumerevoli operazioni in ambito aziendale e governativo, sia in Europa sia negli Stati Uniti, in particolare a partire dagli anni Cinquanta del Novecento e, successivamente, estendendo il proprio campo applicativo grazie alla possibilità di produrre macchine performanti a costi sempre meno elevati.[13] Se per lungo tempo la trascrizione automatica di testi scritti a mano è stata considerata impossibile[14] e negli anni Novanta i progetti di digitalizzazione di testi nell'ambito delle *Humanities* erano ancora legati ai sistemi OCR,[15] a partire dagli anni '10 del XXI secolo è stato possibile applicare l'HTR alle *Humanities* grazie al-

[9] Rispetto ad altre testimonianze, la nostra situazione è, sotto questo aspetto, molto positiva, potendo contare sulla disponibilità di un personale dedicato come S. Gest, mediatrice presso il *Centre Giono* e membro del consiglio d'amministrazione dell'*Association des Amis de Jean Giono*. Ricordiamo che nel 2015 la biblioteca di Jean Giono è stata ceduta da Sylvie Giono alla città di Manosque ed è stata messa a disposizione dei ricercatori appartenendo, insieme alla casa, all'Associazione. Negli scorsi anni, inoltre, sono stati inoltre mossi i primi passi per un progetto di digitalizzazione, totale o parziale dei volumi.

[10] Sito web di riferimento : https://readcoop.eu/transkribus/?sc=Transkribus. Ultimo accesso: 25 aprile 2023 (valido per tutte le pagine web successive). Per una panoramica della letteratura esistente su Transkribus vedere Nockels, J., Gooding, P. et al. (2022), *Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research*, in «Archival Science» 3/22, pp. 367-392.

[11] http://transkriptorium.com/; https://cordis.europa.eu/project/id/674943.

[12] Colutto, S. et al. (2017), *Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents*, International Conference on Document Analysis and Recognition (ICDAR), 4, p. 19.

[13] Per una storia dei dispositivi OCR dagli esordi ai primi anni Ottanta del Novecento, vedere Schantz, H. F. (1982), *The history of OCR: optical character recognition*, *Recognition Technologies Users Association*.

[14] Il fatto che negli ultimi decenni del Novecento uno dei problemi che si tentava di risolvere era la mancanza di accuratezza dei sistemi di fronte a caratteri e materiali irregolari, per il quale una soluzione proposta era la standardizzazione dei caratteri, mostra quanto si fosse lontani dal concepire sistemi di trascrizione automatica della scrittura manuale. Vedere Schantz, H. F. (1982), cit. p. 38.

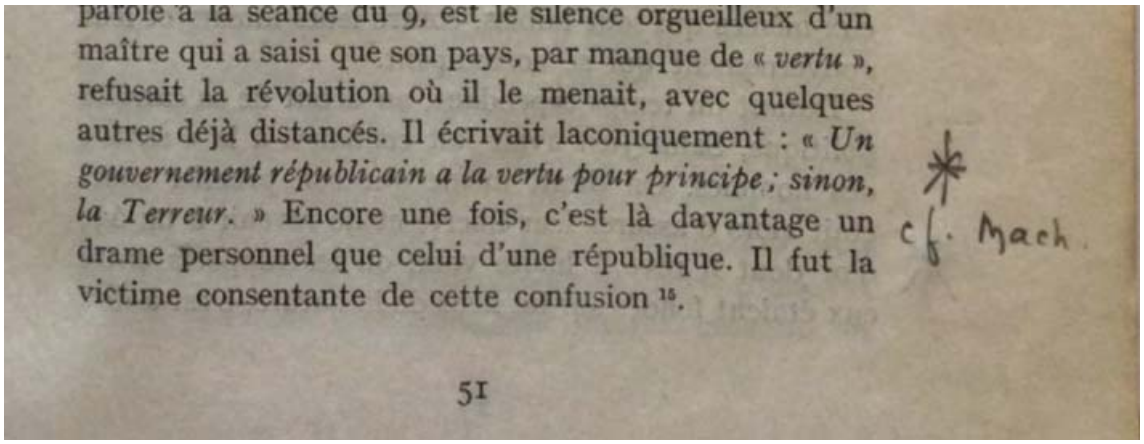
[15] Hodel, T. et al. (2021), General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example in «Journal of Open Humanities Data», 7/13, p. 2.

l'implementazione del *deep learning*, in particolare con l'architettura della rete neurale basata su celle (*cell-based neural network architecture*) chiamata *long-short memory LSTM*.[16]

2. Descrizione del *corpus*

Come anticipato, l'obiettivo del percorso è la pubblicazione online di una porzione della biblioteca politica gioniana, creando edizioni digitali dei singoli postillati. I volumi selezionati per il corpus sono testi a stampa pubblicati tra il 1945 e il 1970. Le pagine presentano i segni di usura dovuti all'utilizzo e al passaggio del tempo, come macchie, lacerazioni e segni di inchiostro, oltre al naturale ingiallimento causato dall'ossidazione. Sul piano contenutistico, le letture includono gli scritti di Machiavelli, Saint-Just, Hobbes,[17] nonché una ricca letteratura critica, tra saggi storico-politici,[18] antologie ragionate,[19] biografie più o meno romanzate e saggi biografici,[20] studi critici,[21] romanzi a sfondo politico.[22] Nell'insieme, i temi trattati sono molteplici e includono riflessioni sulla natura umana, sul concetto di governo nelle sue varie forme, sulla struttura della società, sulla gestione dei conflitti, sulla violenza, sul terrore organizzato e sulla rivolta/rivoluzione. Una certa attenzione è rivolta alla formazione della società moderna, in particolare con il saggio di P. Hazard.[23] Migliaia di pagine, tra le quali Giono si muove lasciando preziose tracce del suo passaggio. I segni di lettura gioniani sono suddivisibili in interni ed esterni. I segni interni ai volumi possono essere verbali (marginalia) o non verbali (sottolineature, riquadri, tratti, asterischi, frecce, disegni, orecchie, segnalibri), mentre i segni esterni sono costituiti da estrazioni o citazioni riportate più o meno fedelmente nei quaderni di lavoro. Nel presente articolo ci concentreremo sui segni verbali, rimandando a prossime pubblicazioni per un lavoro più completo sulle altre tipologie, le quali presentano, sole o in combinazione, diversi livelli gerarchici da cui emerge un'architettura propria dell'autore. I marginalia gioniani contengono riflessioni e commenti suscitati dalla lettura di singoli paragrafi o porzioni più ampie di testo. Lo scrittore inserisce commenti di vario tipo sui contenuti, corregge errori di grammatica, produce rimandi extratestuali ad altre letture, come nell'esempio in Fig. 1., o extraletterari, con richiami alla realtà storico-politica a lui coeva, menzionando correnti e partiti politici, come mostrato in Fig. 2 e Fig. 3. Solitamente gli appunti sono brevi e presentano un impiego diffuso di abbreviazioni.

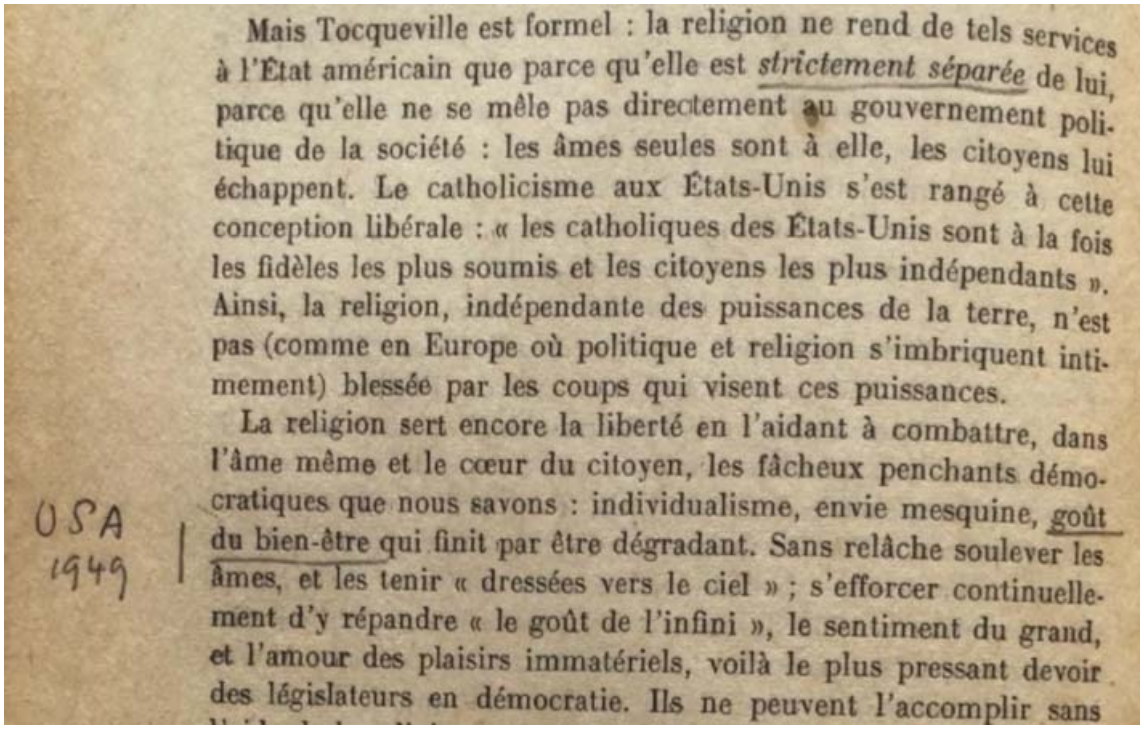
[16] Ibid.
[17] Opere complete di Machiavelli e Saint-Just, *Leviathan* di Hobbes.
[18] Hazard, P. (1935), *La Crise de la conscience européenne (1680-1715)*, Paris, Boivin et C[ie]; Burnham, J. (1949), *Les Machiavéliens défenseurs de la liberté*, Paris, Calmann-Lévy, «Liberté de l'esprit»; D'Astorg, B. (1945), *Introduction au monde de la Terreur*, Paris, Éditions du Seuil, «Pierres vives».
[19] Vellay, C. (1946), *Saint-Just, théoricien de la Révolution*, Monaco, Éditions L. Jaspard, «Les Grands témoins».
[20] Dubreton, J. (1913), *La Disgrâce de Nicolas Machiavel, Florence: 1469-1527*, Paris, Mercure de France; Prezzolini, G. (1929), *Vie de Nicolas Machiavel Florentin*, Paris, Librairie Plon; Gignoux, C.- J. (1947), *Saint-Just*, Paris, La Table Ronde; Ollivier, A. (1954), *Saint-Just et la force des choses*, Paris, Gallimard.
[21] Poirel, V. (1869), *Essai sur les Discours de Machiave*l. *Avec les considérations de Guicciardini*, Paris, Librairie Internationale; Benoist, C. (1907), *Le Machiavélisme* I, «Avant Machiavel», Paris, Librairie Plon; Benoist, C. (1936), *Le Machiavélisme*, III, «Après Machiavel», Paris, Librairie Plon; Chevallier, J.-J. (1949), *Les grandes œuvres politiques, de Machiavel à nos jours*, Paris, Armand Colin, «Sciences politiques»; Burnham, J. (1949), *Les Machiavéliens défenseurs de la liberté*, Paris, Calmann-Lévy, «Liberté de l'esprit».
[22] Wells, H.-G. (1924), *Le Nouveau Machiavel*, Paris, Albin Michel, «Les Maîtres de la littérature étrangère».
[23] Hazard, P. (1935), cit.


parole à la séance du 9, est le silence orgueilleux d'un maître qui a saisi que son pays, par manque de « *vertu* », refusait la révolution où il le menait, avec quelques autres déjà distancés. Il écrivait laconiquement : « *Un gouvernement républicain a la vertu pour principe ; sinon, la Terreur.* » Encore une fois, c'est là davantage un drame personnel que celui d'une république. Il fut la victime consentante de cette confusion [15].

cf. Mach

51

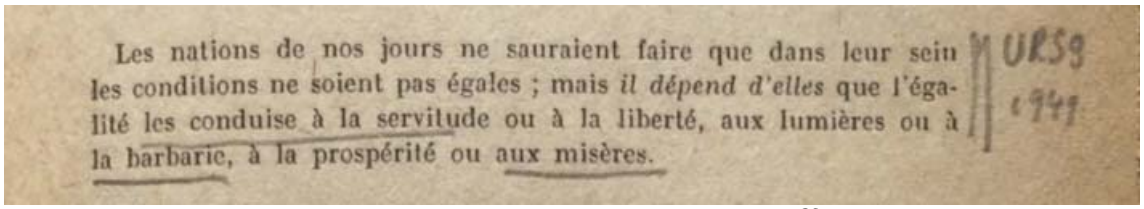
Fig. 1. Primo esempio di marginalia gioniani[24]


Mais Tocqueville est formel : la religion ne rend de tels services à l'État américain que parce qu'elle est *strictement séparée* de lui, parce qu'elle ne se mêle pas directement au gouvernement politique de la société : les âmes seules sont à elle, les citoyens lui échappent. Le catholicisme aux États-Unis s'est rangé à cette conception libérale : « les catholiques des États-Unis sont à la fois les fidèles les plus soumis et les citoyens les plus indépendants ». Ainsi, la religion, indépendante des puissances de la terre, n'est pas (comme en Europe où politique et religion s'imbriquent intimement) blessée par les coups qui visent ces puissances.

La religion sert encore la liberté en l'aidant à combattre, dans l'âme même et le cœur du citoyen, les fâcheux penchants démocratiques que nous savons : individualisme, envie mesquine, goût du bien-être qui finit par être dégradant. Sans relâche soulever les âmes, et les tenir « dressées vers le ciel » ; s'efforcer continuellement d'y répandre « le goût de l'infini », le sentiment du grand, et l'amour des plaisirs immatériels, voilà le plus pressant devoir des législateurs en démocratie. Ils ne peuvent l'accomplir sans

USA 1949

Fig. 2. Secondo esempio di marginalia gioniani [25]


Les nations de nos jours ne sauraient faire que dans leur sein les conditions ne soient pas égales ; mais *il dépend d'elles* que l'égalité les conduise à la servitude ou à la liberté, aux lumières ou à la barbarie, à la prospérité ou aux misères.

URSS 1949

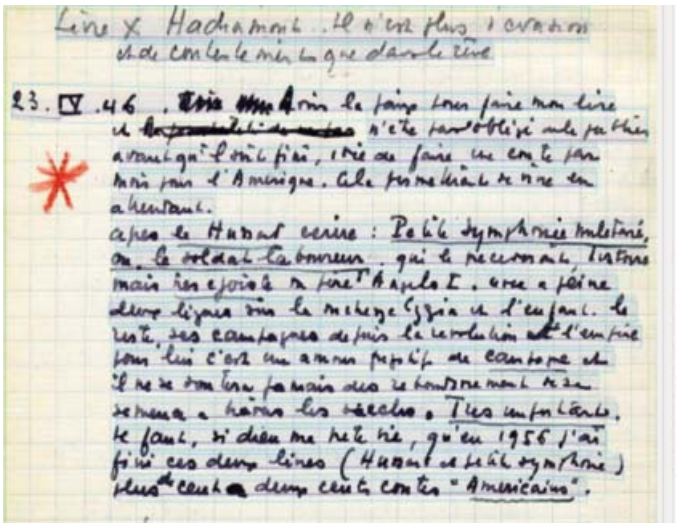
Fig. 3. Terzo esempio di marginalia gioniani [26]

[24] D'Astorg, B. (1945), cit. p. 51.
[25] Chevallier, J.-J. (1949), cit., p. 248.
[26] Ivi, p. 249.

3. Creazione e applicazione del modello di trascrizione

In una prima fase del lavoro, i volumi del corpus sono stati descritti in maniera strutturata tramite schede postillato secondo un modello adattato,[27] scansionati pagina per pagina producendo file .pdf e .jpg. I file sono stati successivamente ottimizzati e ridotti a 300 dpi come consigliato da Read-Coop[28] e infine importati nel programma. Come anticipato, la trascrizione automatica prevede l'applicazione di un modello, che può essere precedentemente creato e allenato fino al raggiungimento di una percentuale di errore sufficientemente ridotta, normalmente al di sotto del 5%.[29] Per la creazione del modello è necessaria la compilazione di un certo numero di pagine *ground-truth*, ovvero trascrizioni corrette, suddivise in linee e aree di testo allineate all'immagine. Queste pagine vengono poi suddivise in set di allenamento e di validazione per il controllo dell'accuratezza (*training/validation set*). Data la natura ibrida dei postillati, che contengono principalmente testo a stampa e, in quantità più limitata, la scrittura manuale presente nei marginalia, è stato necessario fornire al modello campioni di entrambe le tipologie,[30] creando una collezione *ad hoc*. Per velocizzare l'operazione, si è deciso di utilizzare, per i testi a stampa, trascrizioni precedentemente ottenute con OCR di una parte dei postillati che compongono il corpus, mentre per i testi manoscritti si è ricorso alle trascrizioni dei quaderni di lavoro di Giono, già pubblicate nella «Revue Giono» a cura di C. Morzewski. [31] Quest'ultima risorsa ha semplificato notevolmente il lavoro, che si è limitato al rispristino del testo originale, ad esempio, ove in presenza di abbreviature sciolte dai curatori o di commenti descrittivi inseriti all'interno del testo tra parentesi quadre. Al termine del *layout analysis*, funzione che individua automaticamente aree di testo e linee nel file immagine,[32] al quale ha fatto seguito una correzione relativamente rapida, abbiamo associato il testo trascritto, linea per linea, al testo presente nelle fotoriproduzioni fornite dall'*Association des Amis de Jean Giono*, come visibile nell'esempio in Fig. 4.

[27] Ghirardi, S. (2008), *Le postille manzoniane al* Dictionnaire des proverbes français *di Pierre de la Mésangère*, «Prassi Ecdotiche» 3/2008, pp. 227-230.
[28] https://readcoop.eu/terms-and-conditions/.
[29] Si considera buono un tasso inferiore al 10%, molto buono un tasso inferiore al 5% ed eccellente un tasso inferiore al 2.5%. Hodel, T. et al. (2021), cit., p. 2. Osservando altri progetti di trascrizione che impiegano un modello di trascrizione HTR, il grado di accuratezza accettabile si colloca al di sotto del 5% di CER. https://readcoop.eu/category/success-stories/.
[30] Ricordiamo che, per la produzione delle pagine *ground truth* sono consigliate tra le 5.000 e le 15.000 parole manoscritte e circa 1.500 parole per i testi a stampa. https://readcoop.eu/glossary/model-training/.
[31] Morzewski, C. (cur.) (2008), *Journal de Giono – 1946*, in «Revue Giono» 1/2007, pp. 39-74; Id. (cur.), *Journal de Giono – 1946-1949* in «Revue Giono» 2/2008, pp. 59-90.
[32] https://readcoop.eu/transkribus/wiki/layout-analysis/.



1-1 Livre X Hadramaout. Il n'est plus d'évasion
1-2 et de contentement que dans le rêve.
1-3 23.IV.46. Avoir la paix pour faire mon livre
1-4 et n'être pas obligé de le publier
1-5 avant qu'il soit fini. Idée de faire un conte par
1-6 mois pour l'Amérique. Cela permettrait de vivre en
1-7 attendant.
1-8 Après le Hussard écrire : Petite symphonie militaire
1-9 ou Le Soldat Laboureur qui le précéderait, histoire
1-10 mais très égoïste du père d'Angelo I. Avec à peine
1-11 deux lignes sur la duchesse Ezzia et l'enfant. Le
1-12 reste, ses campagnes depuis la Révolution et l'Empire
1-13 pour lui c'est un amour fugitif de campagne et
1-14 il ne se doutera jamais des rebondissements de sa
1-15 semence à travers les siècles. Très important.
1-16 Il faut, si dieu me prête vie, qu'en 1956 j'ai
1-17 fini ces deux livres (Hussard et Petite symphonie)
1-18 plus de cent à deux cents contes « américains ».

Fig. 4. Frammento manoscritto con suddivisione in linee e relativa trascrizione.

Così composto, il modello è stato allenato e testato quattro volte, aggiungendo via via nuovo materiale, fino al raggiungimento di un rapporto soddisfacente tra la percentuale di errore e il *validation set*. Presentiamo in Fig. 5. i parametri del modello prima dell'allenamento.

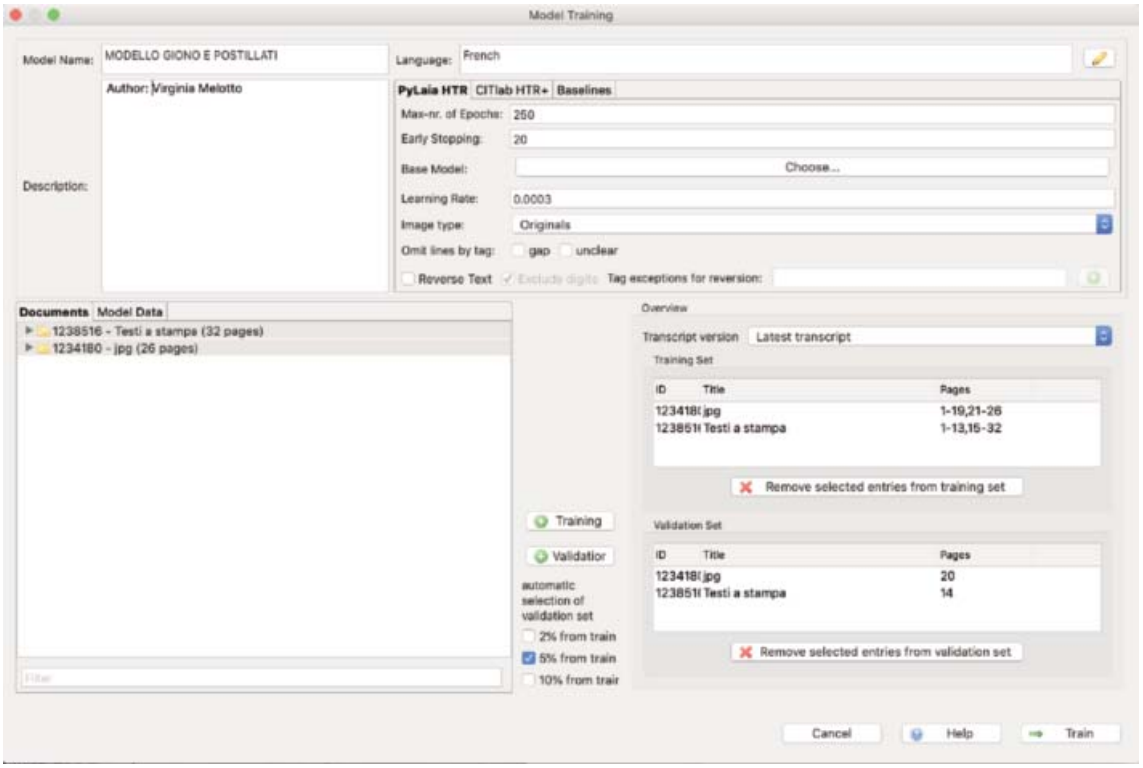


Fig. 5. Parametri del modello (primo allenamento).

Nella Tab. 1. sono visibili i risultati dei vari allenamenti accompagnati dai dati relativi al materiale aggiunto di volta in volta.

Allenamento	Materiale testuale inserito	Risultati ottenuti
Primo allenamento	manoscritti: 5.062 parole a stampa: 8.033 parole	138 epochs were trained, the final CER on the training / validation set was: 1.8000000000000003% / 9.6%
Secondo allenamento	manoscritti: + ca. 7000 parole tot. man. ca. 12.000 parole a stampa: + 9680 parole tot. stamp. ca. 17.700 parole	112 epochs were trained, the final CER on the training / validation set was: 1.7000000000000002% / 4.8%
Terzo allenamento	a stampa: + ca. 25.000 parole tot. stamp. ca. 42.700 parole	74 epochs were trained, the final CER on the training / validation set was: 1.4% / 4.6%
Quarto allenamento	manoscritti: + 4.774 parole tot. man. ca. 17.720 parole a stampa: + 88.800 parole totale: ca. 134.000 parole	79 epochs were trained, the final CER on the training / validation set was: 1.3% / 2.1%

Tab. 1. Dati relativi alla composizione e all'allenamento del modello.

Come è visibile nella Tab. 1., i risultati soddisfacenti a livelli di CER sono stati raggiunti a circa 17720 parole per i testi manoscritti e 88.800 parole per i testi a stampa. I testi manoscritti sono effettivamente più lenti da inserire e per questo motivo sono in numero inferiore rispetto ai testi a stampa. Osservando a confronto le curve di apprendimento del primo e dell'ultimo allenamento (Fig. 6. e 7.), notiamo un notevole avvicinamento del CER *train* e del CER *validation*.[33]

[33] Come indicato da *Read-Coop*, il grafico relativo alla curva di apprendimento mostra una linea blu che corrisponde la progressione dell'allenamento e una linea rossa che rappresenta la progressione delle valutazioni sul set di validazione. L'asse delle ordinate corrisponde al numero di *epochs* o numero di volte in cui i dati di allenamento sono valutati, mentre l'asse delle ascisse rappresenta l'accuratezza in base al CER. Il programma si allena prima sul set di allenamento e successivamente su pagine del set di valutazione. https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/.

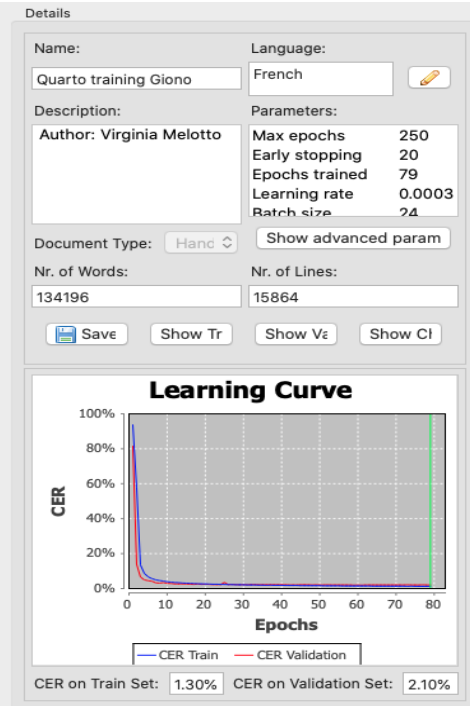


Fig. 6. Dati relativi al primo training.

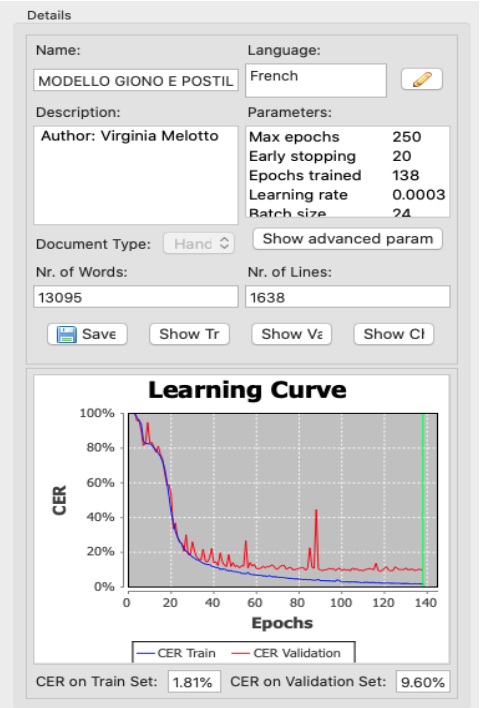


Fig. 7. Dati relativi al quarto training.

A conferma dei dati, al maggiore allenamento del modello e alla riduzione del tasso di errore corrisponde una maggiore accuratezza nella trascrizione, come emerge in Fig. 8. in cui proponiamo un confronto tra la trascrizione dopo il primo allenamento e dopo il quarto.

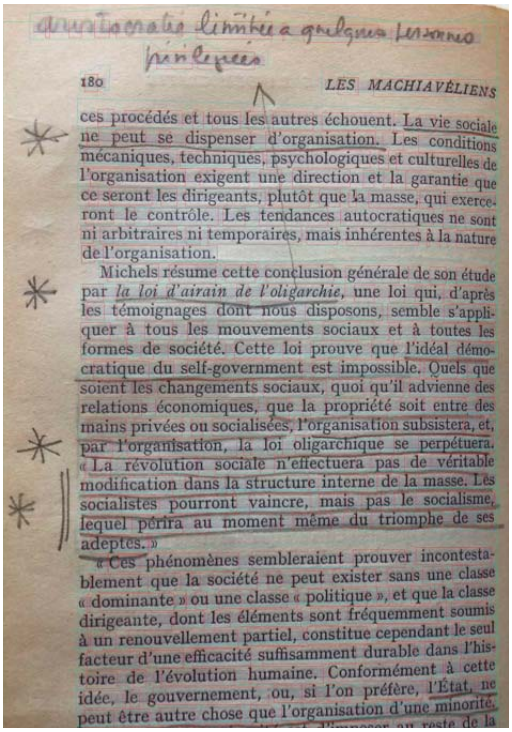


Fig. 8. Prove di trascrizione di un frammento di postillato.[34]

Come si nota in Fig. 8., gli errori presenti sul testo a stampa, già presenti in numero limitato nella prima prova di trascrizione, si riducono ulteriormente nell'ultima. La differenza maggiore emerge, però, nella trascrizione dei marginalia (linee 1-1 e 1-2). A titolo di esempio, possiamo osservare la linea 1-1:

Aristocratie limitée à quelques personnes

I 14 ila cr atis l'isiibiieà quelqires persorma –

e le dr acorâtés limoitée à quelques personnes

A livello di parola, la correttezza si aggira attorno allo 0%-50%, con 1 elemento corretto su 5 nel primo allenamento e 3 su 5 nel quarto, come visibile nella Tab. 2., dove possiamo constatare anche che il sistema ha aggiunto l'accento sulla preposizione *à*, assente nell'originale.

[34] A sinistra il file immagine del postillato, al centro la trascrizione prodotta dopo il primo allenamento e a destra la trascrizione ottenuta dopo l'ultimo allenamento.

Ground truth	Trascrizione dopo il primo allenamento	Trascrizione dopo il quarto allenamento
Aristocratie	I 14 ila cr atis	e le dr acorâtés
limitée	l'isiibiie	limoitée
A	à	à
quelques	quelqires	quelques
personnes	persorma	personnes

Tab. 2. Errori a livello di parola nella linea 1.1.

Calcolando la percentuale di errore CER sull'intera riga,[35] otteniamo i valori presenti nella Tab. 3.

Ground truth	Aristocratie	limitée à	quelques	personnes	CER: 59%
Trascrizione dopo primo allenamento	I 14 ila cr atis	l'isiibiieà	quelqires	persorma	

Ground truth	Aristocratie	limitée à	quelques	personnes	CER: 37%
Trascrizione dopo quarto allenamento	e le dr acorâtés	limoitée à	quelques	personnes	

Tab. 3. Calcolo CER per la linea 1.1.

Com'è possibile intuire dai dati relativi al CER, i risultati dell'applicazione del modello presentano attualmente casi in cui è necessario intervenire maggiormente per la correzione e casi in

[35] Calcolo effettuato applicando la formula impiegata da Read-Coop: CER = [(i + s + d) / n]*100, dove *n* è il numero totale di caratteri, *i* il numero minimo di inserimenti, *s* il numero minimo di sostituzioni e d il numero minimo di cancellazioni.
CER primo allenamento:
N. tot. caratteri 41
Cancellazioni (d) 8
Inserimenti (i) 2
Sostituzioni (s) 14
CER = [(2 + 14 + 8) / 41]*100 = 59%
CER quarto allenamento:
N. tot. caratteri 41
Cancellazioni (d) 5
Inserimenti (i) 0
Sostituzioni (s) 10
CER = [(0 + 10 + 5) / 41]*100 = 37%
Per la formula, vedere https://readcoop.eu/glossary/character-error-rate-cer/
Per il conteggio degli errori come somma di ogni inserimento, sostituzione e cancellazione vedere Levenshtein, V. I. (1966), *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals*, in «Soviet

cui, al contrario, il sistema fornisce una trascrizione corretta al 100%.[36] Riportiamo alcuni esempi di questo tipo (Fig. 9./10./11.), seguiti da una tabella (Tab. 4.) contenente i numeri relativi alle parole corrette, con una distinzione tra testo a stampa e manoscritto.

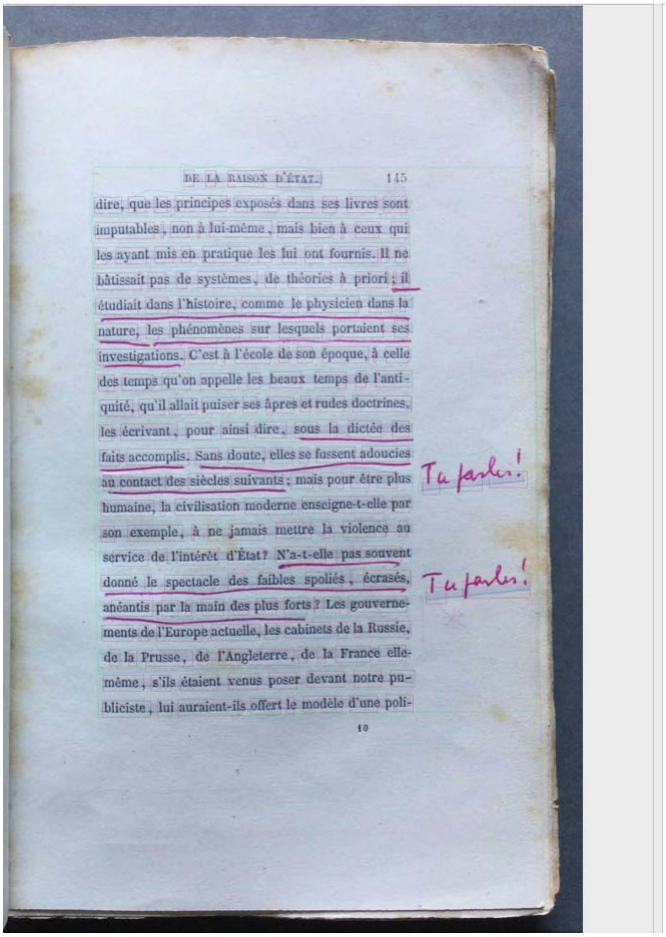


Fig. 9. Esempio di trascrizione 1 (ultima versione del modello).

Phys. Dokl.», 10/8, pp. 707-710, su cui si basa anche il calcolo dell'accuratezza a livello di carattere dei sistemi OCR presentato da Rice, S. V., Kanai, J., Nartker, T. A. (1993), *An evaluation of OCR accuracy*, in Grover, K. O., Goetz, J. P. (dir.), *Information Science Research Institute. 1993 Annual Research Report*, Las Vegas, NV, University of Las Vegas, p. 11. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.7878&rep=rep1&type=pdf#page=9.

[36] In generale si notano difficoltà di riconoscimento maggiori nei casi in cui i marginalia sono apposti a matita e nei casi in cui essi sono inclinati rispetto al testo, mentre nei casi di utilizzo della penna e con postille parallele al testo si ottengono risultati migliori.

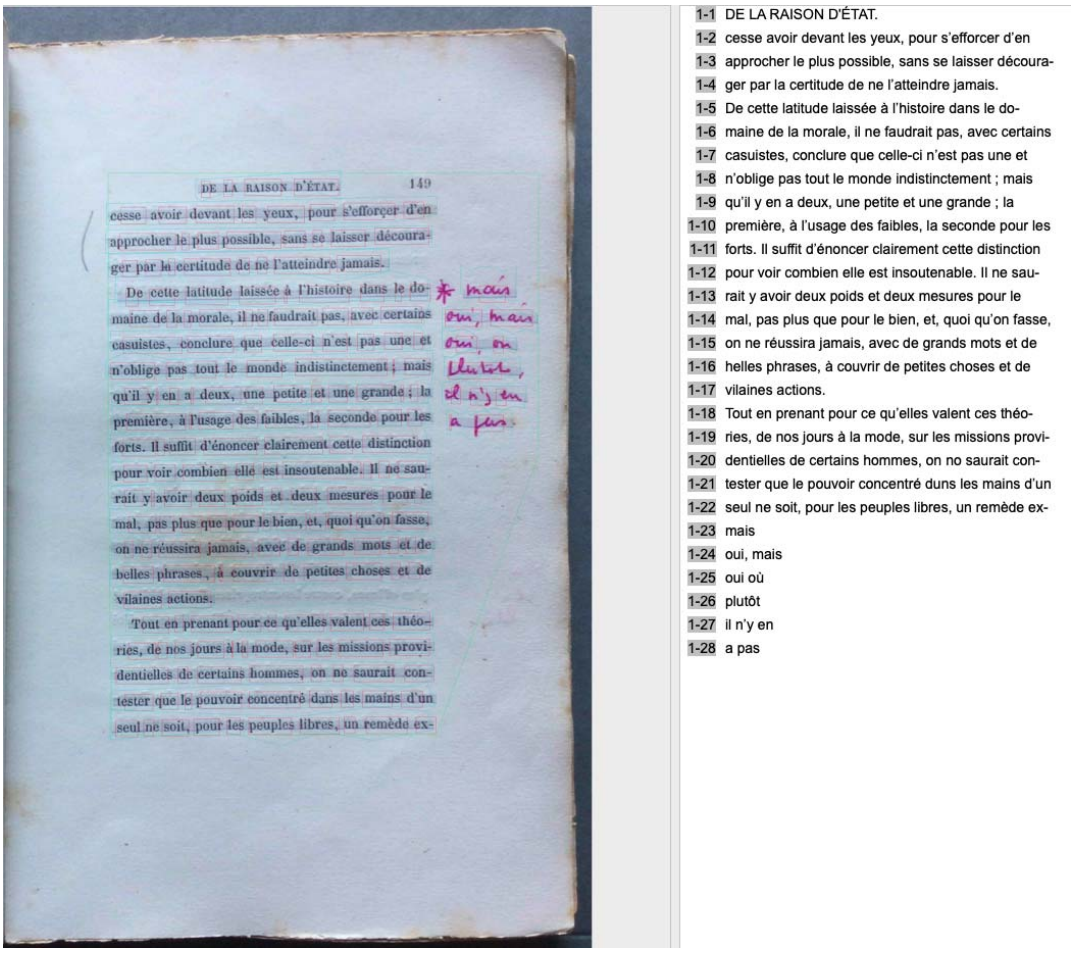


Fig. 10. Esempio di trascrizione 2 (ultima versione del modello).

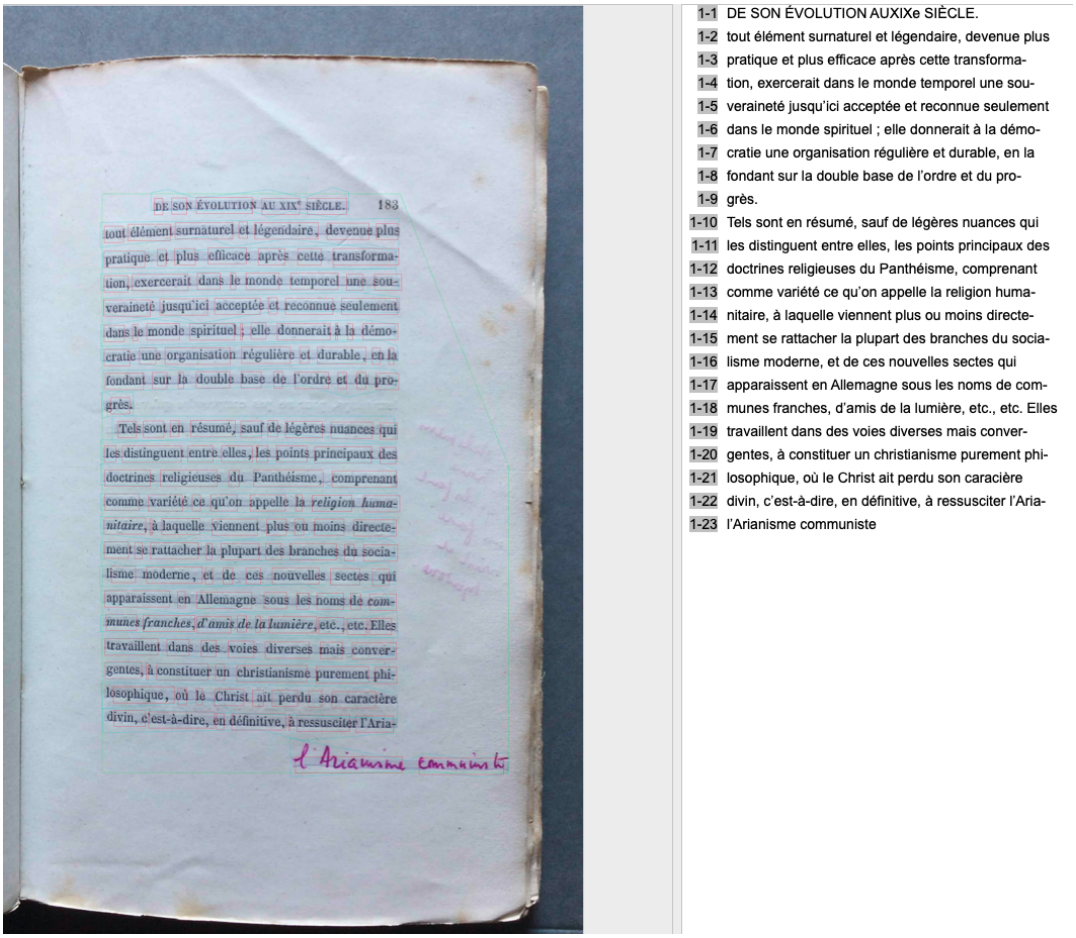


Fig. 11. Esempio di trascrizione 3 (ultima versione del modello).

Pagina Postillato	Parole a stampa totali	Parole a stampa corrette	Parole manoscritte totali	Parole manoscritte corrette
145 (Fig. 9.)	192	187	4	2
149 (Fig. 10.)	195	194	12	12
183 (Fig. 11.)	162	161	3	3

Tab. 4. Parole totali e parole corrette.

Sebbene ancora migliorabili, i risultati appaiono soddisfacenti e confermano l'utilità del modello già allo stadio attuale, giustificando l'introduzione di altro materiale testuale, in particolare manoscritto, per ulteriori allenamenti. È importante sottolineare che in un modello ibrido come il nostro, contenente materiale manoscritto e testi a stampa, la percentuale di testi manoscritti tende ad alzare il CER.[37]

In ogni caso, è bene ricordare che, a prescindere dalla quantità di dati *ground truth* inseriti,

[37] Come sottolineato da Muehlberger et al., i modelli contenenti materiale manoscritto possono mirare a un

alcune irregolarità presenti nella scrittura a mano comporteranno comunque una percentuale di errore.[38] L'obiettivo è quindi quello di ridurre al minimo questa percentuale raggiungendo il CER più basso possibile,[39] al fine di ridurre al minimo gli interventi di correzione e velocizzare il lavoro di trascrizione.

4. Conclusioni

Il presente lavoro ha voluto proporre una possibile strategia per agevolare la trascrizione di documenti utili all'esegesi dell'opera gioniana, nel quadro di un progetto di digitalizzazione. Come in altre fasi del percorso che sfocerà nella pubblicazione online di edizioni digitali, l'impiego delle tecnologie è funzionale allo svolgimento del lavoro filologico. Grazie alle funzioni di creazione e allenamento di un modello di riconoscimento personalizzato, nel nostro caso adattato alla natura ibrida dei postillati, Transkribus consente di processare pagine contenenti scrittura manuale e testo a stampa, creando un collegamento tra linee/aree di testo e immagine che è utile nella successiva fase di marcatura perché riportato nei file xml esportati dal programma. Seppur non utilizzabile al di fuori dell'ambiente in cui è stato creato, il modello HTR è comunque condivisibile con altri utenti che utilizzano il software. La sua creazione è quindi funzionale non solo al lavoro che proponiamo, ma anche nell'ottica di un progetto di digitalizzazione più ampio o volto a processare altre porzioni della biblioteca gioniana.[40] Inoltre, il modello di trascrizione attinge a dati *ground truth* che vengono salvati in cartelle separate all'interno del programma e che possono essere quindi riutilizzati per la creazione di nuovi modelli, ad esempio per la scrittura manuale di Giono per la trascrizione di materiale manoscritto di tipo diverso.[41] La trascrizione costituisce una tappa importante nel percorso di digitalizzazione di fonti disponibili unicamente in formato cartaceo. L'accesso alle edizioni digitali consentirà la diffusione delle letture politiche coeve alla redazione dei romanzi del Secondo dopoguerra, nonché l'osservazione dei marginalia gioniani nel loro contesto di produzione. Se è vero che la trascrizione di documenti tramite modello è applicabile a diversi materiali d'archivio e biblioteche d'autore, infatti, la sua utilità nel caso specifico di Giono, in particolare nell'ambito sociopolitico, si rivela fondamentale per lo studio dei romanzi redatti a partire dalla Liberazione.

Lo studio dei postillati e dei segni di lettura potrà gettare una nuova luce sull'immagine au-

CER al di sotto del 5%, mentre i modelli allenati su testi a stampa raggiungono risultati migliori, con un CER tra l'1 e il 2%. Muehlberger, G. et al. (2019), *Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*
, in «Journal of Documentation», 75(5)/2019, p. 962.

[38] Come affermato dagli autori, l'esperienza mostra che, con un CER inferiore al 3%, i sistemi di trascrizione attualmente esistenti non sono in grado di trattare correttamente i dati a causa dell'irregolarità di alcune scritture manuali, a prescindere dalla quantità di materiale inserito per l'allenamento. Hodel, T. et al. (2021), cit., p. 2.

[39] https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/.

[40] Ricordiamo che la biblioteca del *Paraïs* accoglie ca. 8500 volumi. Mény, J. (2022), *Giono en sa «Librairie»*, «Instinct Nomade», 10/2022, pp. 88-89.

[41] Pensiamo, in particolare, alla corrispondenza e ai manoscritti presenti nell'archivio del *Centre Giono*, del fondo degli *Archives départementales* e del *Paraïs*. Per una panoramica dei fondi e della distribuzione dei documenti vedere Ursch, J. *Les archives de Jean Giono en Haute-Provence*, in Bertrand, M., Not, A. (dir.) (2018), *Patrimoines gioniens*, Aix-en-Provence, Presses Universitaires de Provence, pp. 11-27.

toriale gioniana del Secondo dopoguerra e sta consentendo di elaborare nuovi possibili percorsi di lettura dei romanzi di tale periodo, aprendo a considerazioni sul concetto di engagement e sulla postura dello scrittore in questo ambito. Già in un prima fase della nostra ricerca, infatti, è emerso un contatto tra le letture politiche e alcuni romanzi del secondo periodo, come *Les Âmes fortes* (1949) e *Le Moulin de Pologne* (1952). Dallo studio di questi testi emergono, inserite all'interno della narrazione, riflessioni che vanno ben oltre la singolarità delle storie esemplari e che aprono a considerazioni di carattere etico e politico provenienti da alcuni dei testi di Machiavelli e nell'opera di Paul Hazard.[42] Se il ruolo di Machiavelli nella formazione del pensiero gioniano postbellico è noto, lo studio dei testi e dei segni di lettura, unitamente ai contenuti dei saggi redatti dall'autore nel quadro del progetto di pubblicazione del pensatore fiorentino nella «Bibliothèque de la Pléiade» da parte di Gallimard,[43] rivela nuove aree tematiche e nodi concettuali che si riversano nella narrazione romanzesca. Il personaggio di Thérèse in *Les Âmes fortes*, ad esempio, incarna la figura di Machiavelli, in quanto osservatrice della società nei vari strati che la compongono. Il sistema in cui opera il personaggio esclude o supera, come avviene nella politica secondo Machiavelli, la dicotomia bene/male, dominio della morale, concetto che si ritrova, evidenziato da segni non verbali, in Chevallier.[44] Per Machiavelli, la postura di osservatore è funzionale all'elaborazione della teoria politica, mentre, per il personaggio in questione, è funzionale all'adozione di una serie di comportamenti volti al raggiungimento degli scopi prefissati. A supporto di tale interpretazione interviene l'impiego da parte di Giono del motivo della locanda come osservatorio privilegiato, a cui si aggiungono i commenti contenuti nei saggi, visibili come un'esplicitazione di concetti che, seppur rielaborati, trovano nei testi della biblioteca una base di partenza. Nel *Moulin de Pologne* si evidenzia, tra le altre, una riflessione sulla manipolazione della popolazione per governare e ottenere consensi. In questo quadro, la presenza di un'allusione alla paura di morire provocata dal passaggio di una cometa rimanda chiaramente al testo di Paul Hazard, nel quale si fa riferimento all'episodio realmente accaduto per descrivere l'ignoranza e la superstizione che caratterizzavano la società dell'epoca.[45] Nel romanzo, il narratore omodiegetico si interessa alla vicenda della morte dei de M… de la Commanderie in un incidente ferroviario. L'episodio causa la paura irrazionale degli abitanti del circondario dei confronti della famiglia Costa, sulla quale peserebbe una maledizione che potrebbe danneggiarli. Il narratore riporta allora una frase scritta da un abitante: «Je crains la mort apportée par un astre !».[46] Tale frase compare anche in uno dei quaderni di lavoro di Giono,[47] associata all'apparizione della cometa di Halley nel 1910. Ora, una delle letture coeve alla redazione è il testo di Paul Hazard, *Crise de la conscience européenne*, che include un capitolo intitolato "La négation du miracle, les comètes, les oracles et les sorciers", nel quale si rievocano le reazioni di terrore della popolazione al passaggio di una cometa nel 1680 e la conseguente accensione del dibattito tra atei, razionalisti e credenti sulla legittimità di tali credenze. L'associazione alla cometa di Halley nei carnet è indice di una volontà di attualizzazione del discorso di Hazard e, inserita nel tessuto romanzesco del *Moulin de Pologne*, invita a un esame più attento della società come massa inerte, incapace di razionalità e

[42] Hazard, P. (1935), cit.
[43] Giono, J. (1986), *De Homère à Machiavel*, Paris, Gallimard «Cahiers Giono 4».
[44] "Par delà le bien et le mal; bien et mal non pas niés, mais cantonnés dans leur domaine propre, et expulsés du domaine politique." Chevallier, J.-J. (1949), cit., p. 21.
[45] Hazard, P. (1935), cit.
[46] Giono, J. (1980), *Œuvres romanesques complètes*, Paris, Gallimard, «Bibliothèque de la Pléiade», p. 671.
[47] Morzewski, C. (cur.) (2008), *Journal de Giono – 1946* in «Revue Giono» 1/2007, p. 72.

quindi manipolabile. Se qualsiasi pensiero di natura politica, per concretizzarsi con successo, deve ancorarsi a una conoscenza della società con il quale ogni sistema di governo va messo a confronto,[48] l'idea che la Francia di fine Seicento, dell'Ottocento,[49] di inizio Novecento e, data la continuità, perché no, di metà Novecento, sia facilmente manipolabile, rivela dei problemi di fondo nel tessuto sociale che potrebbero favorire l'affermarsi di un potere tirannico. Questi primi esempi invitano a riflettere sulla postura di Giono romanziere nei confronti di tematiche di carattere sociale, etico e politico e delle modalità che adotta per affrontarle nel tessuto narrativo. Nella torre d'avorio del *Paraïs*, al riparo dagli attacchi esterni, Giono si ricostruisce dopo gli eventi traumatici e le accuse subite, ritrovando progressivamente la statura letteraria che merita. Tra gli scaffali della biblioteca dove coltiva il proprio pensiero e la scrivania dove scatena l'immaginazione creatrice, l'autore afferma la propria individualità e il proprio diritto alla libertà, nonché il rifiuto dei dogmi letterari imposti dall'*intellighenzia* dell'epoca, Sartre in testa.

BIBLIOGRAFIA

A. Fonti bibliografiche

Albonico, S., Scaffai, N. (2015), (dir.), *L'autore e il suo archivio*, Milano, Officina libraria.

Castellin, L. G. (2022), *Sotto un cielo vuoto. Il realismo politico nella storia del pensiero internazionale*. Firenze, Le Monnier, «Università».

Colutto, S. et al. (2017), *Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents*, International Conference on Document Analysis and Recognition (ICDAR), 4 2017.

Ferrer, D., D'Iorio, P. (2001), *Bibliothèques d'écrivains*, Paris, Cnrs éditions, «Textes et manuscrits».

Ghirardi, S. (2008), *Le postille manzoniane al* Dictionnaire des proverbes français *di Pierre de la Mésangère*, «Prassi Ecdotiche» 3 2008, pp. 227-230.

Giono, J. (1980), *Œuvres romanesques complètes*, Paris, Gallimard, «Bibliothèque de la Pléiade».

– (1986), *De Homère à Machiavel*, Paris, Gallimard «Cahiers Giono 4».

Hodel, T. et al. (2021), *General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example* in «Journal of Open Humanities Data», 7/13, pp. 1-10. Disponibile online all'indirizzo https://boris.unibe.ch/157474/3/46-597-1-PB.pdf.

Labouret, D. (1995), *L'Écriture polémique de Giono*, in Sacotte, M. (dir.), *Giono l'enchanteur*, Paris, Grasset.

Levenshtein, V. I. (1966), *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals*, in «Soviet Phys. Dokl.», 10/8, pp. 707-710.

Melotto, V., *Il sangue dei camaleonti e la follia dei convulsionari. L'engagement di Jean Giono in Promenade de la mort*, in «Studi Francesi», accettato in data 3 febbraio 2022. In corso di stampa.

Mény, J., *Les lectures politiques de Jean Giono*, articolo di prossima pubblicazione.

- (2022), *Giono en sa «Librairie»*, «Instinct Nomade», 10/2022, pp. 88-89.

Morzewski, C. (cur.) (2008), *Journal de Giono – 1946* in «Revue Giono» 1/2007, pp. 39-74; Id. (cur.), *Journal de Giono – 1946-1949* in «Revue Giono» 2/2008 pp. 59-90.

[48] Nota è, sotto questo aspetto, la visione concreta della politica da parte di Machiavelli. La conoscenza della realtà fattuale come metodo di studio della politica e come base per l'elaborazione di teorie, si concretizza, come noto, nel realismo politico del pensatore fiorentino. Castellin, L. G. (2022), *Sotto un cielo vuoto. Il realismo politico nella storia del pensiero internazionale*. Firenze, Le Monnier, «Università».

[49] L'incidente in cui perdono la vita i de M… de la Commanderie si ispirerebbe a quello, realmente accaduto nel 1842, noto col nome di catastrofe di Versailles. Giono, J. (1980), cit., p. 1284.

Muehlberger, G. et al. (2019), *Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*, in «Journal of Documentation», 75(5).
Nockels, J. et al. (2022), *Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research*, Arch Sci 22, pp. 367-392.
Raboni, G. (cur.) (2018), *Manzoni e altri grandi postillatori tra Sette e Ottocento* in «Prassi Ecdotiche della modernità letteraria», 3/2018, pp. 5-365.
Rice, S. V., Kanai, J., Nartker, T. A. (1993), *An evaluation of OCR accuracy*, in Grover, K. O., Goetz, J. P. (dir.), *Information Science Research Institute. 1993 Annual Research Report*, Las Vegas, NV, University of Las Vegas. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.7878&rep=rep1&type=pdf#page=9.
Schaelchli, É. (2013), *Pour une révolution à hauteur d'hommes*, Neuvy-en-Champagne, Le passager clandestin.
- (2016), *Jean Giono, le non-lieu imaginaire de la guerre, une lecture de l'œuvre de Giono à la lumière de la «Lettre aux Paysans sur la Pauvreté et la Paix»*, Paris, Eurédit.
Schantz, H. F. (1982), *The history of OCR: optical character recognition*, Recognition Technologies Users Association.
Suleiman S. R. (2018), *Le Roman à thèse ou l'autorité fictive*, Paris, Classiques Garnier «Classiques de la littérature» (1983).
Ursch, J. (2018), *Les archives de Jean Giono en Haute-Provence*, in Bertrand, M., Not, A. (dir.) *Patrimoines gioniens*, Aix-en-Provence, Presses Universitaires de Provence, pp. 11-27.

B. Siti web

https://readcoop.eu/transkribus/?sc=Transkribus – Transkribus website
http://transkriptorium.com/ – Progetto TranScriptorium
https://cordis.europa.eu/project/id/674943 – Progetto READ
https://readcoop.eu/terms-and-conditions/ – Riferimento per riduzione risoluzione immagine a 300dpi
https://readcoop.eu/category/success-stories/ – Elenco e descrizione di progetti di digitalizzazione che hanno previsto l'impiego di Transkribus.
https://readcoop.eu/glossary/model-training/ – Riferimento al quantitativo di materiale ground truth consigliato per l'allenamento del modello.
https://readcoop.eu/transkribus/wiki/layout-analysis/ – Riferimento alla funzione di layout analysis
https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/ – Sull'allenamento del modello con Transkribus e l'interpretazione dei dati.
https://readcoop.eu/glossary/character-error-rate-cer/ – Formula per il calcolo della percentuale d'errore.