Trascrivere, analizzare e pubblicare documenti d'archivio: Transkribus e l'intelligenza artificiale al servizio dello storico

Sara Mansutti[1], Salvatore Spina[2]
[1] READ-COOP SCE, Austria - s.mansutti@readcoop.eu
[2] Università degli Studi di Catania, Italia - salvatore.spina@unict.it

ABSTRACT

Il workshop ha l'obiettivo di riflettere sulle opportunità che l'Intelligenza Artificiale offre al ricercatore per lo studio di testi manoscritti e a stampa. Negli ultimi anni, l'uso dell'Intelligenza Artificiale si è fatto strada nell'ambito della ricerca umanistica e dei Beni Culturali, e uno dei maggiori campi di applicazione è stato quello della trascrizione automatica dei manoscritti. Utilizzando la piattaforma Transkribus – sviluppata nell'ambito del progetto europeo Horizon 2020 "READ" e oggi mantenuta e implementata dalla società cooperativa europea READ-COOP SCE –, i partecipanti impareranno ad allenare un modello di riconoscimento automatico del testo manoscritto (Handwritten Text Recognition), processare automaticamente layout complessi (es. tabelle e registri), arricchire manualmente o automaticamente il layout e il testo dei documenti tramite tag. Verranno, inoltre, mostrate varie possibilità per esportare e pubblicare online le trascrizioni, e per analizzarle al di fuori di Transkribus.

PAROLE CHIAVE

Intelligenza Artificiale, Handwritten Text Recognition, Layout Analysis, Transkribus

1. OBIETTIVI DEL WORKSHOP

L'obiettivo principale del workshop è mostrare le potenzialità d'applicazione dell'Intelligenza Artificiale nella ricerca storica e archivistica, in particolare per la trascrizione di documenti manoscritti o stampe antiche. Il workshop inizierà con un'introduzione teorica sull'impatto che l'Intelligenza Artificiale potrebbe avere per lo storico e quali nuove possibilità di ricerca potrebbe aprire. Seguirà poi una parte più pratica in cui, avvalendosi della piattaforma Transkribus, verrà mostrato passo passo come addestrare e applicare modelli di Handwritten Text Recognition e analisi del layout. Infine, verranno presentate diverse possibilità per esportare le trascrizioni generate automaticamente da Transkribus, pubblicarle o analizzarle con altri strumenti. Al termine del workshop, i partecipanti avranno familiarizzato con la piattaforma Transkribus, acquisito gli strumenti per allenare un modello di IA e conosciuto diverse opzioni per pubblicare e analizzare le trascrizioni. Avranno inoltre acquisito una visione più ampia delle possibilità dell'IA in campo storico.

2. CONTENUTO DEL WORKSHOP

Dati, Big Data, Textual Analysis, Web Semantico, sono solo alcuni dei concetti fondanti del *digital turn*. Tutto diventa processabile, e la maggior parte degli oggetti si 'virtualizzano', portando l'esperienza di essi verso significati non più fisici. Le parole diventano 'enti'-dati equiparabili alle «astrazioni di second'ordine»[1] (i numeri), e come tali possono essere analizzate, attraverso algoritmi in grado di ampliare la nostra conoscenza. Ma se ciò è vero per edizioni testuali a stampa (convertibile in testo digitale leggibile dalla macchina grazie all'OCR) e per i testi "nativi digitali", il discorso cambia quando ci riferiamo all'enorme patrimonio archivistico, costituito da manoscritti di diverse epoche e da variegate strutture linguistiche e di linguaggio, su cui è impossibile effettuare analisi computazionali. Così, già dagli anni Cinquanta, emerse la necessità di sviluppare tecnologie in grado di "leggere" tali documenti.[2] Negli ultimi anni, sono stati finanziati vari progetti per lo sviluppo di software e piattaforme che rispondessero proprio a queste necessità: tra essi, possiamo

[1] Cartesio, *Meditazioni metafisiche*; Cartesio, *Discorso sul metodo*.
[2] Dimond, «Devices for Reading Handwritten Characters»; Dunley, «The National Archives - Machines Reading the Archive»; Fischer et al., «Automatic Transcription of Handwritten Medieval Documents»; Sivasankari e Victor, «Handwritten Text Recognition»; Seaward e Kallio, «Transkribus».

menzionare tranScriptorium,[3] Transkribus[4] e eScriptorium.[5, 6] Grazie ad essi, l'IA riesce, infatti, a rendere i documenti archivistici dei testi digitali accessibili, sia all'utente del web, che all'analista-umanista che vuole operare su di essi. Parlare, oggi, di digitalizzazione delle carte archivistiche non può dunque prescindere dal "trasformarle in testi *machine-readable*"; ossia superare la semplice raccolta fotografica, che, seppur ha fatto sì che numerosissime fonti d'archivio venissero immesse nella rete Internet, ostacola un approccio metodologico fondato sull'uso di tecnologie informatiche.

Sono già numerosi gli archivi che utilizzano l'Handwritten Text Recognition (HTR) per rendere il loro patrimonio cartaceo digitalizzato maggiormente accessibile sia ai ricercatori e sia al pubblico generale, ampliando così le possibilità di studiare e scoprire il contenuto dei documenti manoscritti. Un esempio è offerto dal Gemeente Amsterdam Stadsarchief (Archivio della città di Amsterdam)[7] che dal 2016 sta lavorando al progetto *Alle Amsterdamse Akten* per la digitalizzazione, trascrizione e indicizzazione di tutti gli archivi notarili storici della città. Nel 2021, benché solo 1.5% dell'intera collezione di 20 milioni di immagini fosse stato trascritto tramite IA, vennero alla luce due menzioni di pagamenti fatti a Rembrandt per un ritratto da lui dipinto, fino ad allora sconosciuto.[8] Sebbene fin dal 2018 l'HTR sia stato recepito come un potenziale rivoluzionario nel ridefinire l'approccio dei ricercatori alle collezioni archivistiche,[9] non esistono ancora studi che ne abbiano misurato l'impatto. Un sondaggio fatto tra gli utenti di Transkribus, nel marzo-aprile 2019, ha evidenziato come le possibilità offerte dalla piattaforma – e dall'Handwritten Text Recognition in generale – abbiano incentivato la digitalizzazione e orientato la scelta dei materiali da digitalizzare verso documenti manoscritti piuttosto che fotografie o mappe. Il 33% dei partecipanti al sondaggio, inoltre, ha affermato che, senza l'accesso all'HTR, le trascrizioni dei documenti non sarebbero mai state effettuate, mentre il 40% ha sostenuto che le trascrizioni sarebbero state portate a termine ma in un tempo molto maggiore.[10]

Presentate dunque le possibilità che l'IA offre sia allo storico sia agli archivi, la parte pratica del workshop si focalizzerà sulla piattaforma Transkribus, sviluppata nell'ambito del progetto europeo Horizon 2020 "READ" e oggi mantenuta e implementata dalla società cooperativa europea READ-COOP SCE. L'applicazione web Transkribus non richiede alcuna installazione, rendendola quindi accessibile sia agli storici che sono interessati a processare i loro documenti, senza dover imparare nuove competenze informatiche, sia agli studenti e a coloro che si approcciano per la prima volta alle Digital Humanities. Il primo aspetto fondamentale da trattare sarà addestramento di modelli per la trascrizione automatica dei documenti e l'analisi del layout.
Questa è la fase più importante nell'utilizzo dell'IA, quella in cui il ricercatore deve impegnare tempo ed energie, per poi poter processare efficacemente le collezioni che vuole studiare ed ottenere risultati soddisfacenti. Uno dei maggiori vantaggi della piattaforma Transkribus è proprio la possibilità di addestrare modelli di IA specifici per i propri documenti, senza alcuna competenza richiesta in linguaggi di programmazioni o machine-learning. I modelli HTR sono quelli più conosciuti, ma è anche possibile allenarne altri per riconoscere automaticamente layout complessi (es. cartoline, moduli, mappe, annotazioni marginali…), per estrarre dati da tabelle e registri, e per taggare automaticamente sia elementi del layout, sia parole del testo – ad esempio, le abbreviazioni. I partecipanti impareranno a conoscere tutte queste diverse tipologie di modelli addestrabili all'interno di Transkribus e a comprendere come combinarle per processare i loro documenti. A partire da una serie di esempi, verrà mostrato come preparare e annotare le pagine di Ground Truth, come fare il training, quali risultati aspettarsi e come valutare l'efficacia di un modello.

[3] Sánchez et al., 'Handwritten Text Recognition for Historical Documents in the Transcriptorium Project'.
[4] Kahle et al., «Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents»; Muehlberger et al., «Transforming Scholarship in the Archives through Handwritten Text Recognition».
[5] Kiessling et al., 'EScriptorium: An Open Source Platform for Historical Document Analysis'.
[6] Vari articoli hanno analizzato le differenze tra i software e sistemi di riconoscimento HTR; in particolare i due qui citati mettono a confronto Transkribus ed eScriptorium: Huff and Stöbener, 'Projekt OCR-BW'; Maarand et al., 'A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian'. Il seguente resoconto della giornata di studio tenutasi presso la Bibliothèque nationale de France il 9 maggio 2022 offre invece una visione su come le due piattaforme si stanno sviluppando: Gautier et al., 'Compte-rendu de la journée d'étude « Point HTR 2022 » Transkribus / eScriptorium'.
[7] https://transkribus.eu/r/amsterdam-city-archives/#/
[8] Ponte e Schmitz, «Rembrandt Paints Master Carpenter Jacob Wesselsz Wiltingh.»
[9] Dunley, «The National Archives - Machines Reading the Archive».
[10] Terras, «Inviting AI into the Archives».

Una volta terminato il training, i modelli possono essere applicati per trascrivere nuove pagine dalle caratteristiche simili a quelle usate durante l'addestramento. Le trascrizioni così generate – eventualmente, ma non necessariamente corrette manualmente –, possono poi essere utilizzate per diversi scopi: il principale è l'esportazione dei testi in formati digitali che possano garantire, da un lato, la diffusione – e quindi l'accesso e la consultazione – attraverso il Web; da un altro lato, l'analisi testuale attraverso strumenti quali Voyant Tool, Recogito, e altri ancora che si fondano sull'uso di particolari formati, come il semplice ".txt", o i più complessi "TEI-XML". Transkribus consente, inoltre, di esportare le collezioni in PDF ricercabili che, da un lato, garantiscono un approccio "analogico" alla fonte, e, dall'altro, consentono una prima dinamicità dello studio di essa, ossia la ricerca nel corpo del testo. I PDF di Transkribus possono, successivamente, diventare dei "Flipbook" consultabili attraverso portali e siti Web (es. https://www.biscariepistolography.it), rispondendo, in tal modo, ad uno dei principi cardine dell'era digitale: la diffusione capillare dei prodotti della ricerca. Tra le varie opzioni, è possibile anche esportare le trascrizioni in formato TEI-XML. Lo standard TEI è uno strumento importante per la conservazione e l'interoperabilità delle edizioni digitali, nonché per facilitare la loro pubblicazione online tramite tools come EVT e TEI Publisher. Per chi volesse invece pubblicare immagini e trascrizioni direttamente da Transkribus, senza scaricarle e ricaricarle su un altro server, la piattaforma Read&Search facilita questo processo, permettendo di pubblicare le proprie collezioni online, rendendole accessibili a tutti o previa registrazione.

L'IA, come mostra Transkribus, rappresenta, quindi, una possibilità innovativa per l'Umanistica Digitale, con particolare riferimento alle progettualità archivistiche e storiche, oltre che filologiche e linguistiche. La trascrizione automatica garantirebbe, infatti, un accesso a fonti digitali sempre più numerose e, soprattutto, accelererebbe i lavori di digitalizzazione dei fondi archivistici. Dal suo canto, lo storico – così come il linguista ed il filologo – avrebbe la possibilità di lavorare su più fonti relative al suo progetto di ricerca. Ad esempio, per descrivere la pandemia di peste del 1630, gli storici hanno lavorato su una documentazione che rappresentava non più di tre giorni della vicenda. L'acquisizione digitale di una documentazione più corposa, durante il progetto "Venice Time Machine" (Venezia-Losanna),[11] ha consentito, invece, uno studio su tre interi anni dell'evento pestilenziale.[12]

Ottenute le trascrizioni, l'umanista può, infine, avvalersi di ulteriori strumenti per soddisfare le necessità della sua ricerca e analizzare grandi quantità di testi, andando oltre il *close reading* e l'interpretazione. Tra questi, sicuramente Keyphrase Digger[13] rappresenta un valido esempio: un tool in grado di estrapolare "frasi-chiave" da corpi di testi (quali potrebbero essere le lettere di un fondo "Corrispondenza" di una casata nobiliare), con la finalità di mostrare – in una rappresentazione statistico-linguistica – delle informazioni utili allo storico per approfondire aspetti della sua ricerca che sarebbero emersi, magari, dopo un lungo tempo d'analisi. Oppure, la nota – oggi – ChatGPT (OpenAI),[14] i cui modelli di training consentono a questa IA di correggere i testi esportati da Transkribus, in diverse lingue moderne, così che lo storico possa lavorare su documenti la cui struttura linguistica non presenta errori che impedirebbero ad altri tools di operare efficacemente.

3. STRUTTURA

Il workshop (due ore) è stato pensato secondo questa struttura:

1. Lo scoglio dei testi manoscritti e le opportunità dell'Intelligenza Artificiale per lo storico (20 minuti)
2. L'impatto dell'Intelligenza Artificiale negli archivi (20 minuti)
3. Come addestrare modelli di Intelligenza Artificiale con Transkribus (40 minuti)
4. Come esportare le trascrizioni e pubblicarle (20 minuti)
5. Come analizzare le trascrizioni al di fuori di Transkribus (10 minuti)
6. Domande (10 minuti)

Non è richiesta nessuna competenza o conoscenza pregressa per partecipare al workshop. Chiediamo ai partecipanti di venire con il proprio laptop e di registrarsi alla piattaforma Transkribus.

---

[11] Kaplan, «The Venice Time Machine»; Archeomatica, «Venice Time Machine»; Kaplan e di Lenardo, «Big Data of the Past».
[12] Lazzari et al., «A Digital Reconstruction of the 1630–1631 Large Plague Outbreak in Venice».
[13] Moretti, Sprugnoli, e Tonelli, «Digging in the Dirt».
[14] Biswas, «ChatGPT and the Future of Medical Writing»; Jiao et al., «Is ChatGPT A Good Translator?»; Pavlik, «Collaborating With ChatGPT»; Sobania et al., «An Analysis of the Automatic Bug Fixing Performance of ChatGPT»; Zhai, «ChatGPT User Experience»; Alshater, «Exploring the Role of Artificial Intelligence in Enhancing Academic Performance».

BIBLIOGRAFIA

[1] Alshater, Muneer. «Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT». *SSRN Electronic Journal*, 2022. https://doi.org/10.2139/ssrn.4312358.

[2] Archeomatica, Redazione. «Venice Time Machine: quando i Big Data sposano la cultura». Archeomatica - Cultural Heritage Technology, 29 giugno 2017. https://www.archeomatica.it/ict-beni-culturali/venice-time-machine-quando-i-big-data-sposano-la-cultura.

[3] Biswas, Som. «ChatGPT and the Future of Medical Writing». *Radiology*, 2 febbraio 2023, 223312. https://doi.org/10.1148/radiol.223312.

[4] Cartesio, Renato. *Discorso sul metodo*. Armando Editore, 1999.

[5] ———. *Meditazioni metafisiche*. Armando Editore, 2003.

[6] Dimond, T. L. «Devices for Reading Handwritten Characters». In *Papers and Discussions Presented at the December 9-13, 1957, Eastern Joint Computer Conference: Computers with Deadlines to Meet on XX - IRE-ACM-AIEE '57 (Eastern)*, 232–37. Washington, D.C.: ACM Press, 1958. https://doi.org/10.1145/1457720.1457765.

[7] Dunley, Richard. «Machines Reading the Archive: Handwritten Text Recognition Software». Text. The National Archives blog. The National Archives, 19 marzo 2018. https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/.

[8] Fischer, Andreas, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, e Michael Stolz. «Automatic Transcription of Handwritten Medieval Documents». In *2009 15th International Conference on Virtual Systems and Multimedia*, 137–42. Vienna, Austria: IEEE, 2009. https://doi.org/10.1109/VSMM.2009.26.

[9] Gautier, Dassonneville, Adèle Huguet, Marie-Laure Massot, Agnès Tricoche, Marie Carlin, Jean-Philippe Moreux, and Rostaing Aurélia. 'Compte-rendu de la journée d'étude « Point HTR 2022 » Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d'archives'. Report, 9 June 2022. https://hal.science/hal-03692413.

[10] Huff, Dorothee, and Kristina Stöbener. 'Projekt OCR-BW: Automatische Texterkennung von Handschriften'. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 9, no. 4 (29 November 2022): 1–19. https://doi.org/10.5282/o-bib/5885.

[11] Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, e Zhaopeng Tu. «Is ChatGPT A Good Translator? A Preliminary Study», 2023. https://doi.org/10.48550/ARXIV.2301.08745.

[12] Kahle, Philip, Sebastian Colutto, Gunter Hackl, e Gunter Muhlberger. «Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents». In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 19–24. Kyoto: IEEE, 2017. https://doi.org/10.1109/ICDAR.2017.307.

[13] Kaplan, Frédéric. «The Venice Time Machine». In *Proceedings of the 2015 ACM Symposium on Document Engineering*, 73–73. Lausanne Switzerland: ACM, 2015. https://doi.org/10.1145/2682571.2797071.

[14] Kaplan, Frédéric, e Isabella di Lenardo. «Big Data of the Past». *Frontiers in Digital Humanities* 4 (29 maggio 2017): 12. https://doi.org/10.3389/fdigh.2017.00012.

[15] Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 'EScriptorium: An Open Source Platform for Historical Document Analysis'. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2:19–19, 2019. https://doi.org/10.1109/ICDARW.2019.10032.

[16] Lazzari, Gianrocco, Giovanni Colavizza, Fabio Bortoluzzi, Davide Drago, Andrea Erboso, Francesca Zugno, Frédéric Kaplan, e Marcel Salathé. «A Digital Reconstruction of the 1630–1631 Large Plague Outbreak in Venice». *Scientific Reports* 10, fasc. 1 (20 ottobre 2020): 17849. https://doi.org/10.1038/s41598-020-74775-6.

[17] Maarand, Martin, Yngvil Beyer, Andre Kåsen, Knut T. Fosseide, and Christopher Kermorvant. 'A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian'. In *Document Analysis Systems*, edited by Seiichi Uchida, Elisa Barney, and Véronique Eglin, 399–413. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-06555-2_27.

[18] Moretti, Giovanni, Rachele Sprugnoli, e Sara Tonelli. «Digging in the Dirt: Extracting Keyphrases from Texts with KD». In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-It 2015*, a cura di Cristina Bosco, Sara Tonelli, e Fabio Massimo Zanzotto, 198–203. Accademia University Press, 2015. https://doi.org/10.4000/books.aaccademia.1518.

[19] Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, et al. «Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study». *Journal of Documentation* 75, fasc. 5 (9 settembre 2019): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

[20] Pavlik, John V. «Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education». *Journalism & Mass Communication Educator* 78, fasc. 1 (marzo 2023): 84–93. https://doi.org/10.1177/10776958221149577.

[21] Ponte, Mark, e Erik Schmitz. «Rembrandt Paints Master Carpenter Jacob Wesselsz Wiltingh.» *Kroniek van Het Rembrandthuis* 2021, fasc. 1 (2021): 18–29. https://doi.org/10.48296/KvhR2021.02.

[22] Sánchez, Joan Andreu, Vicent Bosch, Verónica Romero, Katrien Depuydt, and Jesse de Does. 'Handwritten Text Recognition for Historical Documents in the Transcriptorium Project'. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 111–17. DATeCH '14. New York, NY, USA: Association for Computing Machinery, 2014. https://doi.org/10.1145/2595188.2595193.

[23] Seaward, Louise, e Maria Kallio. «Transkribus: Handwritten Text Recognition Technology for Historical Documents», 2017. https://dh-abstracts.library.cmu.edu/works/4193.

[24] Sivasankari, M., e S. P. Victor. «Handwritten Text Recognition», 2014. https://www.semanticscholar.org/paper/Handwritten-Text-Recognition-Sivasankari-Victor/513c052d1dd787367bf3343d3ad7ed2e4efdb7da.

[25] Sobania, Dominik, Martin Briesch, Carol Hanna, e Justyna Petke. «An Analysis of the Automatic Bug Fixing Performance of ChatGPT», 2023. https://doi.org/10.48550/ARXIV.2301.08653.

[26] Terras, Melissa. «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*, a cura di Lise Jaillant, 179–204. Bielefeld: Bielefeld University Press, 2022. https://doi.org/10.1515/9783839455845-008.

[27] Zhai, Xiaoming. «ChatGPT User Experience: Implications for Education». *SSRN Electronic Journal*, 2022. https://doi.org/10.2139/ssrn.4312418.