Umanistica Digitale – ISSN: 2532–8816 – n. 23, 2026: Special Issue
F. Boschetti, F. Boldrer – *Filologia virgiliana e codici digitalizzati: esperienza d'uso di servizi. H2IOSC / CLARIN–IT fra ricerca e didattica*
DOI: https://doi.org/10.60923/issn.2532-8816/23841

Filologia virgiliana e codici digitalizzati: esperienza d'uso di servizi

H2IOSC / CLARIN–IT fra ricerca e didattica

Federico Boschetti

Institute for Computational Linguistics "A. Zampolli"
federico.boschetti@ilc.cnr.it

Francesca Boldrer

Università di Macerata
francesca.boldrer@unimc.it

Abstract

The seminar and laboratory activities involving students from the University of Macerata are presented, focusing on the application of Layout Analysis, Handwritten Text Recognition, and manual correction techniques to samples extracted from two Virgilian manuscripts. The collaborative framework among the University, the CNR, and Research Infrastructures is therefore discussed and highlighted.

Keywords: DL2; Latin Philology; CLARIN–IT; H2IOSC; Vergilius.

Sono qui presentate le attività seminariali e di laboratorio che hanno coinvolto gli studenti dell'Università di Macerata, con un focus sull'applicazione di tecniche di Layout Analysis, Handwritten Text Recognition e correzione manuale su campioni estratti da due manoscritti virgiliani. Viene quindi discusso e messo in evidenza il quadro collaborativo tra l'Università, il CNR e le Infrastrutture di Ricerca.

Parole chiave: DL2; Filologia latina; CLARIN–IT; H2IOSC; Virgilio.

1. Descrizione dell'attività – di Francesca Boldrer

Il caso di studio proposto riguarda l'attività di *Digital humanities* svolta presso l'Università di Macerata nel seminario–laboratorio "Teoria e pratica del riconoscimento del testo delle immagini digitali di manoscritti: *Virgilio*", volto a creare una sinergia tra competenze digitali e linguistico–filologiche latine[1].

Scopo dell'attività è stato sia un aggiornamento metodologico riguardante le recenti tecnologie digitali che supportano le discipline umanistiche, in particolare quelle funzionali alla trascrizione digitalizzata di manoscritti latini (tramite la piattaforma *eScriptorium*), sia l'applicazione a passi virgiliani come integrazione allo studio filologico e critico–testuale. Alla spiegazione teorica ha fatto seguito un approccio laboratoriale e sperimentale incentrato sulle due diverse scritture presenti in due codici celebri per la costituzione del testo virgiliano e significativi per lo studio della sua trasmissione in età tardoantica e medievale. Si tratta di codici redatti rispettivamente – in ordine di utilizzo – in scrittura "minuscola carolina", diffusa tra la fine dell'VIII e il IX secolo[2] (cod. *Bernensis* 172), per la quale è stato utilizzato un modello di riconoscimento già collaudato, e in "capitale rustica" o "libraria"[3] (cod. Vat. lat. 3867), scrittura in uso a Roma tra il I sec. a.C. e il VI sec. d.C., il cui riconoscimento testuale è tuttora oggetto di perfezionamento.

Il codice membranaceo *Bernensis* 172 (siglato come cod. **a** nelle edizioni virgiliane)[4], è un manoscritto medievale miscellaneo conservato presso la Burgerbibliothek di Berna, datato al sec. IX[5]. Contiene sia le *Vitae Vergilianae* che le opere di Virgilio (*Bucoliche, Georgiche* e parte dell'*Eneide* fino al V libro compreso)[6], nonché gli *Scholia Bernensia*. Dal punto di vista paleografico **a** presenta una scrittura "minuscola carolina" chiara e regolare, definita "eccellente" ("ausgezeichnet") nella scheda catalografica della Burgerbibliothek, funzionale al lavoro di recupero e di trasmissione dei testi classici svolto durante la rinascenza carolingia per rendere più agevole la lettura dei testi. Contiene correzioni o varianti di una seconda mano quasi coeva.

Nell'ambito degli studi virgiliani questo codice è ritenuto tra i principali *recentiores* del IX sec. in quanto apografo del cod. *Vaticanus Latinus* 3867, detto *Romanus* (noto come **R**), uno dei tre più

[1] L'attività di *Digital humanities*, suddivisa in due sessioni per la durata complessiva di 8 ore, è stata organizzata nei giorni 10–11 marzo 2025 da Francesca Boldrer, professoressa di Lingua e letteratura latina (Università di Macerata), nell'ambito dell'insegnamento di Filologia latina per studenti del corso di laurea magistrale (interclasse LM–14, LM–15), e aperta anche a dottorandi UniMc (dottorati di ricerca in "Umanesimo e tecnologie" e "Diritto e Innovazione"). Relatore è stato il dott. Federico Boschetti, ricercatore del CNR–ILC e responsabile del progetto Arcipelago DH presso il Venice Centre for Digital and Public Humanities, VeDPH (Università Ca' Foscari Venezia).

[2] Vd. ad es. [14] ("La nuova scrittura comune: la minuscola carolina").

[3] Vd. [14] ("La capitale romana nell'uso librario ['rustica']").

[4] Nel *conspectus codicum*; vd. ad es. l'edizione virgiliana di Geymonat 2008²[7].

[5] Verso la metà del secolo; vd. Munk Olsen (MO) B.12 [7]. Cfr. la scheda catalografica con bibliografia e la riproduzione digitale nel sito della Burgerbibliothek di Berna e–codices – Virtual Manuscript Library of Switzerland (accessibile anche attraverso l'archivio digitale di poesia latina *MQDQ*, s.v. "Vergilius, eclogae, testimoni, cod. **a**"). Cfr. inoltre [8], *ad cod.* 172.

[6] Per la parte mancante (*Aen.* libri VI–XII) il cod. *Bernensis* 172 è completato dal cod. *Parisinus lat.* 7929. Alcune lacune sono presenti anche in Verg. *buc.* I e *Aen.* V. Vd. [7].

antichi e importanti, sia in particolare per il testo di Virgilio[7], sia in generale rispetto a tutti i codici latini esistenti, ma in parte lacunoso[8]. Il cod. *Bernensis* 172 appare dunque prezioso per integrarlo nelle parti mancanti.

Il cod. **R**, conservato presso la Biblioteca Apostolica Vaticana, è appunto il secondo manoscritto preso in esame nel seminario–laboratorio per il riconoscimento del testo delle immagini digitali di manoscritti. Si tratta di un codice membranaceo datato tra il V e il VI secolo[9], famoso anche per le sue 19 miniature[10]. Scritto, come anticipato, in "capitale rustica", presenta interventi di più mani, tra cui la prima risalente alla fine del VI sec. d.C., mentre le successive (dalla seconda alla quinta) di età carolina o più recente[11].

Prima di utilizzare i codici digitalizzati sopra indicati, in una prima fase teorico–scientifica si è discussa l'importanza delle infrastrutture di ricerca europee per la produzione, il mantenimento e la fruizione di risorse e servizi pensati per gli studi umanistici. In particolare, attraverso l'analisi dell'infrastruttura CLARIN[12], si è visto come esistano famiglie di risorse[13] molto diversificate per la ricerca e lo studio di testi in latino, che vanno dai *corpora* testuali, alle treebank (dati morfosintattici strutturati secondo la grammatica di dipendenza), ai lessici e alle risorse lessico–semantiche come le wordnet (basi di dati che permettono di effettuare ricerche sia semasiologiche, partendo dalle parole per arrivare ai concetti, che onomasiologiche, partendo dai concetti per arrivare alle parole)[14]. L'attenzione si è poi concentrata su quello che è uno dei principali obiettivi della filologia digitale nell'ambito delle *digital humanities*, ovvero la creazione di edizioni scientifiche digitali a partire dai facsimili digitalizzati di manoscritti o di edizioni a stampa per arrivare ai testi *machine actionable*.

Ha quindi fatto seguito l'esposizione e dimostrazione delle diverse fasi necessarie per estrarre il testo dalle immagini digitali di manoscritti delle opere di Virgilio tramite la piattaforma eScriptorium. Tale strumento ha come obiettivo la creazione dell'edizione 'diplomatica' di manoscritti attraverso la segmentazione e il riconoscimento del testo per aree e per linee. A partire da immagini digitali del manoscritto, eScriptorium elabora una fedele trascrizione del

[7] Assieme a due altri codici virgiliani, ovvero, come noto, il *Mediceus Laurentianus* lat. Plut. XXXIX, 1 (**M**) e il *Vaticanus Palatinus* lat. 1631 (**P**).

[8] Mancano *buc.* 7,1–10,9; *georg.* 2,1–215; *Aen.* 2,73–3,684; 4,217–5,36; 11,757–792; 12,651–686; 759–830; 939–952. Vd. [7].

[9] La datazione tra V e VI è indicata da Reynolds 1983, p. 434; propende invece per la metà del VI sec. [7]; cfr. Lowe 1934–1971 (CLA), I 19. Vd. la scheda catalografica con bibliografia presente nel sito della Biblioteca Apostolica Vaticana (https://digi.vatlib.it/mss/detail/Vat.lat.3867) con riproduzione digitale in Vat.lat.3867 | DigiVatLib (consultabile anche attraverso *MQDQ*; vd. *supra* n. 5). Sulle proposte di datazione (e localizzazione di **R**) cfr. anche Manoscritti di IV–VI sec. ("codices antiquiores") | Manuscripta Vergiliana.

[10] Tra cui tre ritratti di Virgilio. Vd. [20], [19].

[11] Vd. [7].

[12] https://www.clarin.eu

[13] Per le famiglie di risorse linguistiche e testuali si veda: https://www.clarin.eu/resource–families

[14] Le risorse letterarie disponibili possono essere costituite da singoli testi o da *corpora*, ossia collezioni organizzate in base a criteri linguistici o letterari, accessibili tramite il Virtual Language Observatory (VLO): https://vlo.clarin.eu.

testo, applicando un modello che permette di riconoscere i caratteri di varie scritture e di mantenere abbreviazioni ed eventuali corruttele ed errori.

Come illustrato nel corso del seminario, il modello per leggere una varietà di scritture si crea sulla base dei campioni che i ricercatori sottopongono a eScriptorium, ovvero delle trascrizioni fatte dagli studiosi sulla base delle immagini digitali del manoscritto. La trascrizione fatta da eScriptorium è tanto più accurata quanti più numerosi sono i campioni forniti al programma. Attualmente eScriptorium ottiene buone prestazioni sulle scritture latine "carolina" e "mercantesca", mentre la "capitale rustica" e l'"onciale" necessitano di modelli *ad hoc*. L'obiettivo è dunque innanzitutto quello di ottenere riproduzioni fedeli graficamente all'originale con la possibilità di effettuare ricerche linguistiche e testuali (varianti, frequenza lessicale, caratteristiche grammaticali e stilistiche). Inoltre, ciò può fornire la base di ulteriori lavori filologici, quali la creazione di edizioni critiche.

Nella seconda parte laboratoriale si è quindi sperimentato il processo di caricamento delle immagini digitalizzate dei manoscritti per giungere alla segmentazione e alla trascrizione automatica e manuale del testo di un *folium* del cod. *Bernensis* 172 (**a**). Operativamente, una volta distribuite le sezioni da esaminare, è stato possibile iniziare un lavoro di segmentazione del testo, esplorando le tecniche utilizzate dal copista nella composizione della pagina, le particolarità grafiche ed eventuali anomalie della scrittura. A tal proposito, alcune pagine presentavano singolarità da evidenziare tramite la segmentazione manuale, quali annotazioni nelle aree marginali o sovrapposizioni di righe che escono dallo spazio occupato dal testo principale.

Nel caso del secondo codice virgiliano preso in esame, il Vat. lat. 3867 (**R**), il lavoro consisteva nella verifica della trascrizione automatica realizzata da eScriptorium. Effettuata la segmentazione delle linee è necessario fare il riconoscimento delle diverse aree testuali, da indicare poi come testo unico o, se presenti, come colonna di sinistra e di destra. In ultimo si associa il testo alla rispettiva area ottenendo un testo principale con eventuali varianti o glosse marginali. Fatto ciò, il testo va numerato, ordinato e sottoposto a eScriptorium, che effettua la trascrizione riga per riga. Come hanno rilevato i partecipanti, la scrittura nell'originale era in genere precisa e facilmente leggibile, eppure la trascrizione presentava errori in varie parti del testo, come nel caso della sostituzione della lettera G alla C; inoltre alcune lettere erase e sovrascritte nel codice non erano state lette correttamente.

Nella prima attività pratica è stato interessante, per i partecipanti, poter osservare direttamente il testo originale dei manoscritti, accessibile nei databases delle rispettive biblioteche in cui sono conservati, e confrontarlo parola per parola con quello stampato in edizioni cartacee o riprodotto in banche–dati prese come riferimento (ad es. nell'archivio digitale di poesia latina *Musisque Deoque*). In una seconda fase si è aggiunta l'analisi più accurata della scrittura "capitale rustica" per correggere la trascrizione automatica prodotta da eScriptorium. Tale attività seminariale, che ha permesso ai partecipanti di interagire con testi classici, non senza varie difficoltà da superare nella decifrazione – non tanto nella comprensione della grafia dei copisti, quanto nel caso di aree sbiadite e consunte –, ha contribuito potenzialmente alla creazione di un nuovo modello di riconoscimento applicabile all'intero manoscritto.

La partecipazione attiva e le valutazioni a fine attività hanno mostrato come l'esperienza sia stata molto interessante ed istruttiva. Come hanno osservato gli studenti e i dottorandi partecipanti nelle relazioni finali[15], le due sessioni hanno permesso un arricchimento delle competenze,

[15] Tra cui particolarmente dettagliata quella della dottoranda Giovanna Migliorelli (dottorato di ricerca in "Diritto e Innovazione", UniMc).

coniugando un adeguato background iniziale sull'argomento ad aspetti pratici. Compatibilmente con il tempo a disposizione è stata fornita una panoramica utile a orientarsi tra i vari strumenti e le varie infrastrutture, nella consapevolezza dei loro principi fondanti, e i partecipanti sono stati messi in condizione di gestire un workflow completo, dal reperimento delle risorse agli strumenti di trascrizione automatica e manuale fino alla diffusione dei risultati; inoltre, sono stati forniti spunti interessanti sul backend kraken per eventuali approfondimenti tecnici.

In questa esperienza di seminario–laboratorio di *Digital Humanities* si è data preminenza all'aspetto tecnologico–digitale, data la novità dei mezzi proposti per indagini di Filologia latina. Ciò ha richiesto un'ampia introduzione metodologica e permesso solo alcuni saggi esemplificativi, sostanzialmente indipendenti dal contenuto dell'opera virgiliana utilizzata. In futuro vi è l'intenzione di rinnovare l'esperienza puntando a un'applicazione diffusa su testi letterari più estesi e unitari per senso e forma, con analisi linguistiche che potranno avvalersi del perfezionamento dei metodi di riconoscimento anche delle scritture storiche più antiche.

In conclusione, l'incontro ha evidenziato l'importanza della collaborazione interdisciplinare tra classicisti e specialisti di *digital humanities* per la valorizzazione del patrimonio culturale, e ha mostrato come anche gli studi classici possano avvalersi delle tecnologie più avanzate e come queste possano fornire strumenti utili per supportare il *durus labor* e stimolare l'*ingenium* del filologo.

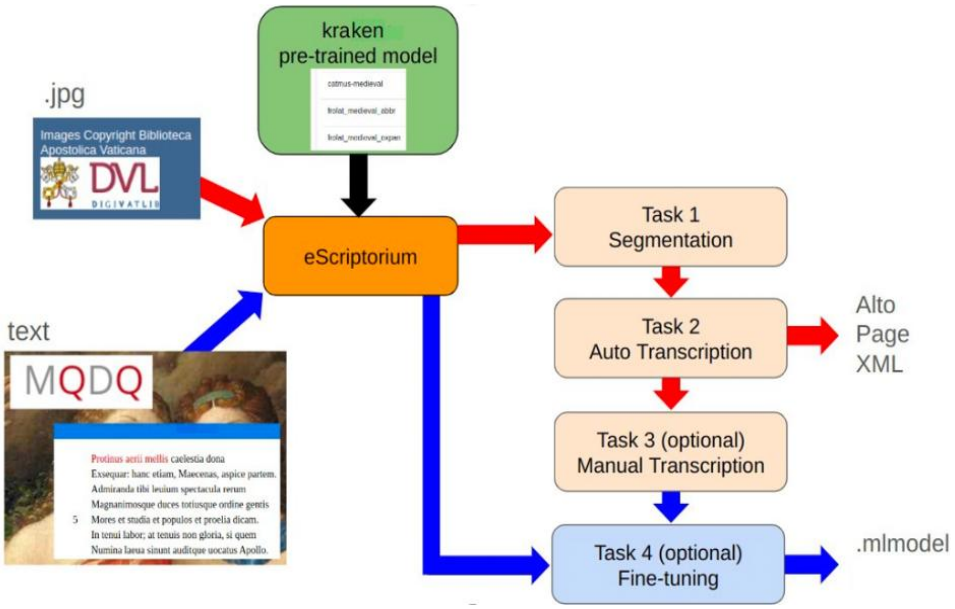

Fig. 1 dati in ingresso e in uscita in un sistema di HTR

2. Il contesto di collaborazione in cui è nata l'attività didattica – di Federico Boschetti

L'attività didattica si è svolta in stretta collaborazione fra l'Università di Macerata, il CNR–Istituto di Linguistica Computazionale "A. Zampolli" e CLARIN–IT[16], il consorzio italiano dei centri afferenti all'infrastruttura di ricerca europea CLARIN per le risorse linguistiche.

Grazie ai finanziamenti di NextGenerationEU & PNRR "Italia Domani" – Missione 4 – Componente 2 – Linea di Investimento 3.1 – Azione 3.1.1 i consorzi dei nodi italiani delle quattro principali infrastrutture di ricerca per le scienze umane, E–RIHS–IT[17], DARIAH–IT[18], CLARIN–IT e OPERAS–IT[19] si sono federate dal 2022 in H2IOSC[20] (Humanities and cultural Heritage Italian Open Science Cloud) per mettere a disposizione della comunità scientifica di ambito umanistico risorse e servizi web interoperabili.

Le quattro infrastrutture sono infatti complementari e coprono ambiti diversi del sapere: E–RIHS è prevalentemente focalizzata sulla digitalizzazione dei beni culturali materiali; DARIAH sulla creazione di edizioni scientifiche digitali; CLARIN sull'analisi linguistica con strumenti computazionali e OPERAS sulla promozione della scienza aperta, pur essendoci ampie e proficue zone di sovrapposizione.

H2IOSC ha offerto quindi l'opportunità alle quattro infrastrutture di ricerca di ideare un flusso di lavoro continuo che va dalle immagini digitali dei manoscritti messe a disposizione tramite il protocollo IIIF, alla digitalizzazione del documento e del testo a partire da tali immagini tramite l'applicazione di tecnologie di Layout Analysis e di Automatic Text Recognition (ATR), all'analisi linguistica dei documenti risultanti tramite UDPipe[21], fino all'esposizione dei metadati di tutti i prodotti digitali per garantire i principi FAIR a sostegno della Scienza Aperta.

Nel corso del 2025 H2IOSC ha pubblicato quattro bandi di Transnational and National Access[22] (TNA/NA) ai propri servizi rivolti a università ed enti di ricerca per cominciare a valutare l'interesse dei ricercatori e l'impatto sulla comunità scientifica degli strumenti offerti. Ciascuna infrastruttura ha contribuito al TNA/NA con una selezione di risorse che meglio caratterizzano le attività di ciascuna e con incontri mirati alla formazione all'uso di tali risorse.

Fra gli strumenti e i servizi messi a disposizione da CLARIN–IT, oltre a quelli di analisi linguistica, si trova anche eScriptorium[23], piattaforma online e insieme di web services sviluppati da INRIA[24] per l'ATR applicato a testi a stampa (ICR: Intelligent Character Recognition) oppure a manoscritti (HTR: Handwritten Text Recognition). All'interno del TNA/NA eScriptorium,

[16] https://clarin–it.it

[17] https://www.e–rihs.it

[18] http://stdl.cnr.it/it/dariah

[19] https://operas–eu.org/operas–national–nodes

[20] https://www.h2iosc.cnr.it

[21] https://lindat.mff.cuni.cz/en/udpipe)

[22] https://www.h2iosc.cnr.it/tna–na–calls

[23] https://gitlab.com/scripta/escriptorium

[24] https://www.inria.fr

con un'istanza temporaneamente ospitata da D4Science[25] che sta per essere trasferita sui server del datacenter di H2IOSC, ha riscosso notevole interesse, con più di cento utenti distribuiti su una dozzina di progetti fortemente diversificati per lingua (greco bizantino, latino, italiano antico), obiettivi (didattici, di ricerca, di sperimentazione tecnica), numero di componenti (dal singolo a classi di venti allievi), distribuzione geografica (italiana, europea e in alcuni casi extraeuropea).

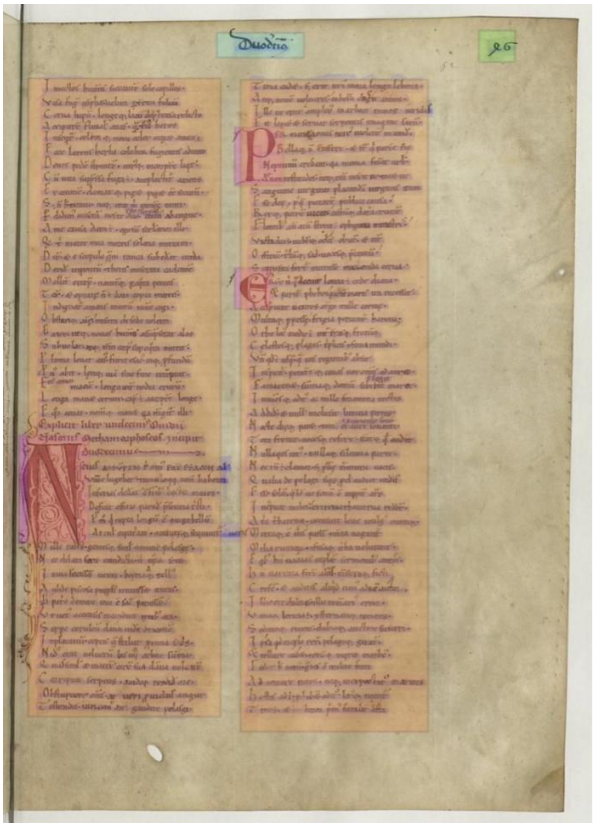

Fig. 2 Layout analysis e classificazione delle aree di testo

La scelta di eScriptorium

Nel panorama degli strumenti di HTR attualmente i protagonisti sono Transkribus[26] ed eScriptorium. La scelta è caduta su eScriptorium perché più aderente ai paradigmi della Scienza aperta, che richiede non solo apertura per quanto riguarda i prodotti finali della ricerca, ma anche in tutte le fasi del processo di produzione, rendendo tutte le fasi tracciabili e riproducibili.

eScriptorium, distribuito con licenza aperta MIT, usa kraken[27] come motore di riconoscimento automatico del testo, a sua volta distribuito con licenza aperta Apache 2.0. I modelli di ATR

[25] https://www.d4science.org

[26] https://www.transkribus.org

[27] https://kraken.re/7.0/index.html, https://github.com/mittagessen/kraken

sono distribuiti su Zenodo[28]. Inoltre — e questo costituisce una delle novità maggiori rispetto a iniziative analoghe — i metadati e spesso anche i dati della *ground truth* sono resi disponibili in modo aperto, tramite HTR–United[29]. La *ground truth* è costituita dalla mappatura del testo trascritto manualmente o corretto accuratamente sull'immagine della pagina del manoscritto o dell'edizione a stampa. Tale mappatura si ottiene fornendo le coordinate delle linee di testo in formato XML–ALTO o XML–Page.

Infine, è degno di nota il vocabolario controllato SegmOnto,[30] che fornisce le linee guida per classificare le aree di testo, facilitando la conversione dai formati per ATR come XML–ALTO e XML–Page in XML–TEI.

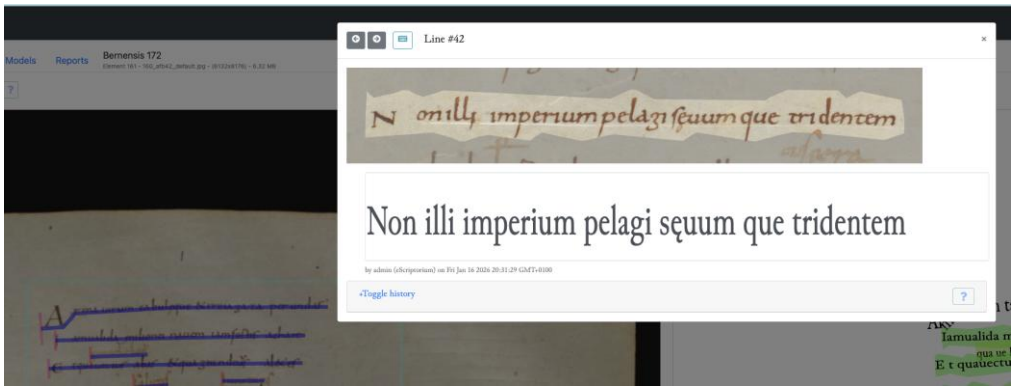

Fig. 3 Maschera per la correzione manuale (Verg. Aen. 1,138)

3. Conclusione

In conclusione, questa esperienza ha dimostrato che l'applicazione dell'ATR a campioni estratti da manoscritti latini può essere adottata a fini didattici sia come palestra per integrare lo studio della letteratura con elementi di codicologia e di paleografia a partire da casi concreti, sia per introdurre gli studenti al mondo delle infrastrutture di ricerca per gli studi umanistici.

[28] https://zenodo.org/communities/ocr_models

[29] https://htr–united.github.io

[30] https://segmonto.github.io

References

[1] Balbo, Andrea. 2023. Insegnare latino. Sentieri di ricerca per una didattica ragionevole. UTET (in part.: 198–200).

[2] Boldrer, Francesca. 1991. "Virgilio, Georg. 2,332" [germina/gramina]. Materiali e discussioni per l'analisi dei testi classici 27: 145–157.

[3] Boldrer, Francesca. 1997. "Il ritorno di Orfeo (Verg. Georg. 4,509)". In Miscillo flamine. Studi in onore di Carmelo Rapisarda; Dipartimento di Scienze filologiche e storiche, Università di Trento: 83–99.

[4] Burdick, Anne, and Drucker, Johanna, and Lunenfeld, Peter, and Presner, Todd, and Schnapp, Jeffrey. 2014. Umanistica digitale. Mondadori.

[5] Capaccioni, Andrea. 2022. Umanistica digitale. Tra transizione tecnologica e tradizione. Maggioli editore.

[6] Ciotti, Fabio. 2023. Digital Humanities. Metodi, strumenti, saperi. Carocci.

[7] Geymonat, Mario. 2008². P. Vergili Maronis Opera. Edizioni di Storia e Letteratura.

[8] Hagen, Hermann. 1974. Catalogus codicum Bernensium (Bibliotheca Bongarsiana). Georg Olms.

[9] Lazzari, Marco. 2014. Informatica umanistica. Connect (in part.: 156–162).

[10] Lowe, E.A. 1934–1971. Codices Latini Antiquiores, A Paleographical guide to Latin Manuscripts Prior to the Ninth Century (CLA), I–XII, Oxonii.

[11] Marassi, Massimo, and Scotti Muth. 2024. Umanesimo e digitalizzazione. Vita e pensiero.

[12] Munk Olsen, B. 1985. L'étude des auteurs classiques latins aux XIe et XIIe siècles, Catalogues des manuscrits classiques latins copiés du Ixe au XIIe siècle, II, Pariis, 673–826 (cfr. III.1, Addenda et corrigenda, ibidem 1989, 138–153).

[13] Pace, Rosaria. 2015. Digital Humanities, una prospettiva didattica. Carocci.

[14] Petrucci, Armando. 1989. Breve storia della scrittura latina. Bagatto Libri (in part.: 51-55 e 109-119).

[15] Reynolds, L.D. 1983. Texts and transmission. A survey of the Latin Classics. Clarendon Press.

[16] Spinazzè, Linda. 2015. Filologia digitale. Dalla ricerca alla didattica. Tangram Edizioni Scientifiche.

[17] Stok, Fabio. 2016. I classici dal papiro a Internet. Carocci (in part.: 251–255).

[18] Venuti, Martina (coordinatrice dal 2019). Musisque deoque https://www.mqdq.it

[19] Walther, Ingo F. and Wolf, Norbert. 2005. Codices Illustres: The world's most famous illuminated manuscripts, 400 to 1600. Taschen.

[20] Weitzmann, Kurt. 1977. Late Antique and Early Christian Book Illumination. George Braziller (in part.: 11, 52–59).