Strumenti digitali per la trascrizione e la lemmatizzazione di testi in italiano antico

Emiliano Degl'Innocenti[1], Alessia Spadi[2], Federica Spinelli[3], Lucia Francalanci[4], Michela Perino[5], Irene Falini[6], Francesco Coradeschi[7], Francesco Pinna[8]

[1]CNR Istituto Opera del Vocabolario Italiano, Italia - emiliano.deglinnocenti@cnr.it

[2]CNR Istituto Opera del Vocabolario Italiano, Italia - alessia.spadi@cnr.it

[3]CNR Istituto Opera del Vocabolario Italiano, Italia - federica.spinelli@cnr.it

[4]CNR Istituto Opera del Vocabolario Italiano, Italia - lucia.francalanci@cnr.it

[5]CNR Istituto Opera del Vocabolario Italiano, Italia - michela.perino@cnr.it

[6]CNR Istituto Opera del Vocabolario Italiano, Italia - irene.falini@cnr.it

[7]CNR Istituto Opera del Vocabolario Italiano, Italia - francesco.coradeschi@cnr.it

[8] CNR Istituto Opera del Vocabolario Italiano, Italia - francesco.pinna@cnr.it

ABSTRACT

Il contributo si focalizza sullo sviluppo e sull'uso di metodologie per supportare e potenziare la ricerca nel contesto delle discipline umanistiche e del patrimonio culturale, con particolare riferimento all'ambito della filologia digitale. Partendo dal caso di studio del Fondo Datini dell'Archivio di Stato di Prato, l'obiettivo è lo sviluppo di nuovi strumenti digitali, nonché l'integrazione ed il potenziamento di strumenti esistenti, finalizzati allo studio del carteggio privato e commerciale del mercante pratese Francesco di Marco Datini. Lo scopo di questo progetto pilota è l'ampliamento, nel contesto del cluster H2IOSC [4], degli obiettivi raggiunti nell'ambito del progetto RESTORE (smaRt accESs TO digital heRitage and mEmory) in riferimento al trattamento di lettere edite che costituiscono il corpus lemmatizzato Archivio Datini realizzato dall'Istituto Opera del Vocabolario Italiano (OVI-CNR). L'implementazione di tali strumenti consentirà di facilitare la ricostruzione di una parte significativa della storia delle città d'Europa e dei porti del Mediterraneo del XIV secolo, evidenziandone sia le dinamiche della vita quotidiana, sia le specificità territoriali, sociopolitiche e commerciali.

PAROLE CHIAVE

HTR (Handwritten Text Recognition); lemmatizzazione; machine learning; filologia digitale; italiano antico.

1. INTRODUZIONE

Il progetto di ricerca descritto prende avvio dai risultati raggiunti dal progetto RESTORE[1] [3] (smaRt accESs TO digital heRitage and mEmory), incentrato sulla figura del mercante pratese Francesco di Marco Datini, della sua famiglia e del suo entourage di collaboratori. Partendo dalla dimensione locale è possibile ricostruire una parte rilevante della storia delle città dell'Europa e dei porti del Mediterraneo del XIV secolo, con le loro dinamiche sociolinguistiche e le loro peculiarità sociali ed economiche. Le due sezioni principali del Fondo Datini dell'Archivio di Stato di Prato sono infatti costituite dal carteggio privato tra Francesco e i suoi cari (tra cui la moglie Margherita Bandini e l'amico Lapo Mazzei) e dall'imponente carteggio commerciale che testimonia la fervida attività dei fondaci aziendali del mercante, situati in tutto il Mediterraneo.

Nello specifico, il punto di partenza di questo studio è il corpus (parzialmente) lemmatizzato Archivio Datini[2] – realizzato dall'Istituto Opera del Vocabolario Italiano del Consiglio Nazionale delle Ricerche (OVI-CNR) –, grazie al quale ci si propone di sviluppare una piattaforma di strumenti digitali integrati per supportare la ricerca nelle scienze umanistiche, focalizzandosi su alcune funzionalità chiave, quali la trascrizione e la lemmatizzazione automatica dei testi in italiano antico (con focus sulle scritture mercantesca e cancelleresca). L'obiettivo è fornire agli studiosi un ambiente avanzato per analizzare e interpretare i testi, che includa servizi integrati e interoperabili con strumenti già a disposizione della comunità dei ricercatori (quale ad esempio il TLIO).

Il contesto di riferimento è dato dagli studi condotti dal team riunito attorno al nodo italiano dell'Infrastruttura di Ricerca DARIAH[3] (Digital Research Infrastructure for the Arts and Humanities), con sede presso l'OVI[4], attualmente impegnato

[1] http://restore.ovi.cnr.it/

[2] Corpus lemmatizzato del carteggio Datini: http://aspweb.ovi.cnr.it/(S(acenhe55wjva14ulrxas2oyw))/CatForm01.aspx

[3] Nodo Italiano dell'Infrastruttura di Ricerca DARIAH: https://dariah.cnr.it/

[4] Istituto Opera del Vocabolario Italiano: http://www.ovi.cnr.it/

nel progetto H2IOSC[5] [4] (Humanities and cultural Heritage Italian Open Science Cloud), finanziato dal Piano Nazionale di Ripresa e Resilienza italiano (PNRR), che mira a creare un cluster partecipato dai nodi nazionali di 4 infrastrutture di ricerca ESFRI: DARIAH.it, CLARIN.it, OPERAS.it, E-RIHS.it. Nel corso degli ultimi anni, DARIAH.it ha indirizzato la sua attività verso la promozione dell'interoperabilità e l'accesso a risorse digitali legate al patrimonio culturale, sia tangibile (oggetti) che intangibile (elementi intellettuali e concettuali). Collaborando con comunità di ricerca consolidate, il gruppo si interfaccia con esperti del settore e promuove la Citizen Science al fine di plasmare lo sviluppo di un ecosistema digitale interoperabile per la ricerca nell'ambito delle Social Sciences and Humanities (SSH).

2. STATO DELL'ARTE

Nel corso del suo sviluppo e implementazione, il progetto RESTORE ha affrontato una serie di problematiche legate al trattamento dei dati, riassumibili in: 1) elevata frammentazione delle risorse digitali nei contesti di riferimento, che rischia di comprometterne il valore e ne limita la riutilizzabilità; 2) difficoltà di accesso e isolamento scientifico delle risorse di alta qualità prodotte da biblioteche, archivi e centri di ricerca; 3) eterogeneità dei formati e degli standard; 4) scarsa interoperabilità e carenza di strategie di sostenibilità a medio e lungo termine per le risorse digitali prodotte e gestite dagli attori coinvolti [8]. Inoltre, la crescente produzione di informazioni digitali accentua la sfida nell'organizzazione e nella gestione, mettendo in evidenza la necessità di criteri di selezione, strutturazione e pubblicazione dei dati per garantire qualità scientifica e interoperabilità. Il progetto, coordinato dall'OVI, ha coinvolto inizialmente istituzioni culturali del circuito GLAMs (Galleries, Libraries, Archives, Museums), a cui si sono uniti - nel corso del tempo - altri soggetti, attivi nel campo Social Sciences and Humanities (SSH) ed Heritage Science (HS) (ad esempio diagnostica e restauro del patrimonio culturale), con lo scopo di recuperare, integrare e rendere accessibili i dati, in linea con i principi FAIR[6] [16] (in breve: rintracciabilità, accesso, interoperabilità e riutilizzo). Per affrontare le sfide individuate è stato definito ed implementato un flusso di lavoro completo, capace di garantire l'integrazione e l'interoperabilità dei dati forniti da diversi soggetti, tra cui enti di ricerca, culturali e di conservazione nazionali e locali. Il flusso si articola nei seguenti passaggi: 1) acquisizione dei dati originali (nei formati in cui sono disponibili) forniti dai partner; 2) sviluppo di procedure per la normalizzazione e l'allineamento delle risorse codificate secondo gli standard dei domini di riferimento (ad es: TEI[7] per la codifica dei testi; EDM[8], MAG[9], MODS e METS[10] per la descrizione delle risorse bibliotecarie; EAD[11] e EAC[12] per la descrizione delle risorse archivistiche; ICCD[13] – in particolare la scheda OA – come sistema di catalogazione per le opere d'arte; altri standard afferenti a diverse discipline nel dominio delle scienze del patrimonio come EDF[14], HDF5[15] ecc.); 3) validazione dei dati normalizzati, attraverso la collaborazione con gli esperti di dominio; 4) mappatura e modellazione sulla base dell'ontologia scelta, CIDOC - Conceptual Reference Model[16]; 5) trasformazione dei dati in triple semantiche e caricamento nella base di dati semantica (Virtuoso Triplestore[17]); 6) esposizione di uno SPARQL endpoint per l'interrogazione della base di dati semantica e di un'interfaccia per la navigazione dei dati; 7) sviluppo di interfacce user-friendly, completamente integrate tra loro, per la visualizzazione dei dati, a cui si aggiunge l'uso di strumenti quali LodLive[18] (front-end per la visualizzazione grafica delle triple semantiche e la navigazione concettuale), EVT[19] (strumento open source per la progettazione e visualizzazione di edizioni digitali), Movio[20] (piattaforma open source multifunzionale per realizzare mostre virtuali). Inoltre, tutto il codice e la documentazione prodotti sono open source e archiviati in repository pubblicamente accessibili.

[5] https://www.h2iosc.cnr.it/
[6] FAIR: https://www.go-fair.org/fair-principles/
[7] Text Encoding Initiative - TEI: https://tei-c.org/
[8] Europeana Data Model - EDM: https://pro.europeana.eu/page/edm-documentation
[9] Administrative and Management Metadata - MAG: https://www.iccu.sbn.it/export/sites/iccu/documenti/manuale.html
[10] Metadata Object Description Schema - MODS e METS: https://www.loc.gov/standards/mods/presentations/mets-mods- morgan-ala07
[11] Encoded Archival Description - EAD: https://www.loc.gov/ead/
[12] Encoded Archival Context - EAC: https://eac.staatsbibliothek-berlin.de
[13] Istituto Centrale per il Catalogo e la Documentazione - ICCD: http://www.iccd.beniculturali.it/
[14] European Data Format - EDF: https://www.edfplus.info/
[15] Hierarchical Data Format - HDF5: https://www.hdfgroup.org/solutions/hdf5
[16] CIDOC - Conceptual Reference Model: http://www.cidoc-crm.org/
[17] Virtuoso Openlink Triplestore: https://virtuoso.openlinksw.com/
[18] LodLive: http://lodlive.it/
[19] Edition Visualization Technology – EVT: http://evt.labcd.unipi.it/
[20] Movio: https://www.gruppometa.it/it/movio

A partire dalla base di dati del progetto si vuole espandere il range di strumenti a disposizione di studiosi di varie discipline (ad es.: paleografi, filologi, lessicografi, linguisti, storici, filosofi ecc.) che si occupano di testi in italiano antico. Pertanto, DARIAH.it sta lavorando sia al potenziamento e alla FAIRificazione degli strumenti già realizzati nel contesto di RESTORE, sia allo sviluppo di strumenti di trascrizione e di lemmatizzazione (inclusa l'integrazione e la generalizzazione di strumenti esistenti) per varietà storiche di italiano. I servizi di trascrizione e lemmatizzazione sono in corso di addestramento su un dataset già trattato dal gruppo di lavoro e messo a disposizione dall'OVI e dall'Archivio di Stato di Prato per il progetto RESTORE, il già citato corpus testuale Archivio Datini, che raccoglie una selezione di lettere appartenenti al fondo omonimo, fisicamente depositato presso l'Archivio di Stato di Prato. Vista l'eterogeneità degli scriventi, il corpus comprende più varietà linguistiche e registra diverse forme grafiche e morfologiche di molti termini rilevanti per la ricostruzione del lessico quotidiano dell'epoca e del lessico tecnico legato alle attività economiche delle aziende datiniane. La lemmatizzazione approntata dai ricercatori dell'OVI include anche antroponimi, compresi eventuali soprannomi e posizioni specifiche (se l'indicazione si riferisce a una precisa personalità storica identificata), e toponimi, compresi nomi di città, paesi, distretti, località, strade, piazze, porte, chiese, monasteri, palazzi, ospedali, organizzazioni, istituzioni, ecc. Sono annotati inoltre termini relativi al campo religioso e agricolo, alle parti del corpo, alle scansioni temporali, ecc., distribuiti in 22 categorie concettuali (chiamate iperlemmi), tra cui: abbigliamento e arredamento, cibo, animali, arti e mestieri, calendario, legge ed economia politica, costruzione e architettura, medicina, monete, navigazione, parentela, cuoio e tessuti, e così via. In sintesi, il corpus è composto da: 2.511 testi; 45.259 forme; 977.034 occorrenze di cui 126.663 lemmatizzate; 6.510 lemmi e 22 iperlemmi (utilizzati per raggruppare diversi lemmi).

3. TRASCRITTORE

Nell'ambito del progetto è stato selezionato un dataset di circa 300 lettere appartenenti al corpus Archivio Datini, ciascuna associata alla rispettiva trascrizione e riproduzione digitale. L'associazione è stata resa possibile grazie al lavoro di integrazione e allineamento di dati, metadati e immagini effettuato durante la costruzione della base di dati semantica. Questa raccolta è stata scelta come caso di studio per addestrare uno strumento finalizzato al riconoscimento e alla trascrizione automatica del testo. Le lettere, redatte in scrittura mercantesca da diversi mittenti, costituiscono parte del carteggio commerciale e privato di Francesco Datini, così come precedentemente descritto. Il progetto di ricerca si propone di implementare un sistema di HTR (Handwritten Text Recognition), mirato alla trascrizione automatica della scrittura mercantesca. L'HTR utilizza modelli di apprendimento automatico, come reti neurali artificiali, per estrarre e interpretare caratteri scritti a mano in immagini, trasformandoli in testo digitale [1, 2, 5, 12]. Negli ultimi anni, l'HTR si è affermato come modello predominante nello sviluppo di strumenti, tra i quali, Loghi[21], eScriptorium[22] e Transkribus[23], dedicati alla trascrizione automatica di testi antichi. Nell'ambito di questo lavoro, il gruppo di ricerca DARIAH.it attivo presso l'OVI ha avviato una valutazione degli strumenti esistenti e del loro utilizzo da parte della comunità di ricerca di riferimento che ha portato all'individuazione del software eScriptorium - gratuito e open source, basato sull'OCR engine Kraken, anch'esso gratuito e open source - come uno dei modelli di riferimento per il progetto pilota in corso di sviluppo.

Il procedimento per la trascrizione automatica di testi comunemente comprende diverse fasi: preelaborazione dell'immagine, segmentazione, OCR/HTR, e post-elaborazione. La fase di segmentazione mira a individuare le linee di testo, preparandole per il successivo processo di trascrizione. Per questa fase - con l'obiettivo di addestrare un modello di segmentazione sperimentando l'utilizzo di eScriptorium, di cui è stata approntata una istanza locale - è stato selezionato un gruppo di lettere che presentano similitudini nel layout della pagina. Una delle sfide principali del progetto in corso di sviluppo risiede nella corretta gestione della variabilità dello stile di scrittura degli autori e nella tipologia corsiva della mercantesca; pertanto, il corpus di addestramento del carteggio Datini è stato selezionato anche tenendo conto del numero elevato di mani e mittenti.

Il progetto pilota fornirà un sistema ottimizzato per il riconoscimento del segno grafico e la successiva trascrizione assistita della scrittura, con particolare riferimento al carteggio considerato. Il sistema darà inoltre la possibilità di annotare la trascrizione in base a diverse tipologie di criteri (paleografico, filologico, storico, ecc.).

4. LEMMATIZZATORE

Questa parte del contributo è dedicata alla descrizione di alcuni esperimenti relativi all'annotazione linguistica automatica (in particolare *Part-of-Speech tagging* e lemmatizzazione) di varietà storiche di italiano. Proprio per la complessità dello

[21] https://github.com/rvankoert/loghi
[22] https://gitlab.com/scripta/escriptorium
[23] https://readcoop.eu/transkribus/

studio dell'italiano antico la scelta dei testi su cui operare è di fondamentale importanza quale requisito preliminare per procedere alla lemmatizzazione. I materiali scelti vengono organizzati in:

- Corpus di addestramento: uno o più testi, intesi come collezione di frasi compiute, lemmatizzati in modo esaustivo. Si considerano parte di questo corpus sia i testi usati per l'addestramento che quelli usati per la valutazione.
- Corpus di lingua (o lessico): insieme di testi che comprenda il corpus di addestramento, ma con numero totale di occorrenze molto maggiore. Il corpus di lingua non deve essere lemmatizzato esaustivamente ma è preferibile (anche se non indispensabile) che sia lemmatizzato esaustivamente per forme. In particolare, per l'esperimento che si sta descrivendo si prende come riferimento il TLIO (Tesoro della Lingua Italiana delle Origini)[24] dove tutte le forme sono lemmatizzate almeno una volta, anche se non in tutte le occorrenze.

Alla costruzione dei corpora si è affiancata un'analisi esplorativa sullo stato dell'arte delle risorse e degli strumenti dedicati al trattamento automatico dell'italiano antico [6, 7, 13, 14, 15], da cui sono emerse due possibili tipologie di approcci: da una parte, l'uso di strumenti specifici per l'italiano antico, dall'altra lo sviluppo o il riadattamento di strumenti di annotazione addestrati sull'italiano contemporaneo.

La lemmatizzazione procede per "frasi", dove per "frase" si intende la parte di testo tra l'inizio o un segno di interpunzione forte e la fine o il segno di interpunzione forte successivo, ovvero quello che si definisce un periodo. Una volta individuate le frasi la procedura si può riassumere nei seguenti punti: 1) per ogni frase si considerano tutte le occorrenze; 2) tramite il lessico si associano ad ogni occorrenza tutti i lemmi a cui corrispondono; 3) si costruiscono le catene di lemmi della frase. Una catena di lemmi consiste in una sequenza di due o più lemmi (lo standard si attesta su catene di cinque lemmi ma non è dogmatico), intese come terne costituite da forma standard, categoria grammaticale e disambiguatore. Per ogni occorrenza viene costruito un insieme di catene di lemmi in tutte le combinazioni disponibili nel lessico già indentificate al punto 2; 4) ad ogni catena di lemmi si associa una probabilità intrinseca calcolata sulla base della probabilità forma/lemma e della probabilità lemma/struttura; 5) una volta note le probabilità intrinseche di tutte le catene di lemmi della frase si stima per l'intera frase la concatenazione di lemmi di massima probabilità. Per determinare la sequenza di lemmi di massima probabilità potrebbe sembrare necessario calcolare la probabilità di tutte le combinazioni di catene di lemmi possibili, con relativo costo computazionale piuttosto alto. In realtà si vede che partendo da un lato della frase (l'inizio o la fine è equivalente) ed aggiungendo di volta in volta un "anello" alla catena, stimando le probabilità tramite l'algoritmo di Viterbi[25], si riesce a mantenere la complessità algoritmica della procedura sotto controllo.

L'obiettivo dello studio è la realizzazione di un servizio per la lemmatizzazione semiautomatica di testi scritti in italiano antico basato sul machine learning. La realizzazione di un tale servizio è strettamente collegata alla costruzione di uno o più corpora annotati rappresentativi delle varietà storiche, da poter usare in fase di addestramento e di valutazione.

Uno di questi dataset è costituito dal sotto corpus creato a partire dalle lettere appartenenti all'Archivio Datini, precedentemente descritto. Gli esperimenti condotti su questi testi hanno evidenziato sia le problematiche connesse alle peculiarità specifiche di testi antichi, come l'alta variabilità a tutti i livelli di analisi linguistica (grafico, morfologico, sintattico e lessicale), sia i problemi derivanti dall'adattamento di strumenti preesistenti a varietà linguistiche differenti. Una parte dei test è stata effettuata sul sistema di lemmatizzazione semiautomatica realizzato nel 2010 da Domenico Iorio-Fili [9, 10] e inserito all'interno della versione 4.0 di GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), il software che gestisce la versione elettronica dei corpora testuali dell'OVI. Si tratta di uno strumento propriamente costruito per il trattamento dell'italiano antico e in particolare per il caso specifico del Corpus OVI, caratterizzato da dimensioni relativamente ridotte, ma da un'elevata complessità e variabilità del materiale linguistico. Altre prove sono state invece condotte con strumenti di NLP (Natural Language Processing) addestrati sull'italiano contemporaneo e sviluppati in ambiente Python.

5. CONCLUSIONI

Visti i riscontri positivi da parte della comunità scientifica circa le possibilità offerte da RESTORE per la ricostruzione del lessico e della fitta rete di persone e luoghi gravitanti attorno alla figura di Francesco di Marco Datini nel Mediterraneo del XIV sec., l'Archivio Datini si configura come l'oggetto di studio ideale per testare l'efficacia, anche in termini di riutilizzabilità, di due strumenti, trascrittore e lemmatizzatore, indubitabilmente utili alla ricerca scientifica in molteplici discipline che hanno come punto di partenza l'interpretazione di testi in italiano antico. Lo sviluppo di queste tecnologie da parte di DARIAH.it si inserisce nella progettazione di uno strumento pilota dedicato alla filologia digitale (Digital

[24] http://tlio.ovi.cnr.it/TLIO/

[25] L'algoritmo Viterbi è un algoritmo ideato da Andrew Viterbi e generalmente utilizzato per trovare la migliore sequenza di stati (detta Viterbi path) in una sequenza di eventi osservati in un processo markoviano. Da Wikipedia: https://it.wikipedia.org/wiki/Algoritmo_di_Viterbi

Philology Hub) (vd. Fig. 1) previsto in seno a H2IOSC, per l'ideazione della quale si stanno seguendo le linee guida esposte in Leonardi (2021) [11] a lungo discusse con il gruppo DARIAH.it attivo presso l'OVI. La progettazione di questo pilot per la ricerca filologica rientra inoltre tra gli obiettivi della collaborazione fra lo Spoke 3 del progetto CHANGES[26] («Digital library, archives and philology») ed H2IOSC, con particolare riferimento all'attività di sviluppo del Digital Philology Hub, coordinata dall'OVI per DARIAH.it; nel medesimo contesto di collaborazione si colloca l'istituzione del corso di dottorato FROID[27] («Filologia Romanza e Italiana Digitale») presso la Scuola Normale Superiore[28], che vede la compartecipazione di DARIAH.it e dell'OVI attraverso il finanziamento di una borsa di dottorato legata allo sviluppo dell'Hub: primi esempi della fruttuosa sinergia tra i progetti PNRR IR H2IOSC e PE CHANGES.

Figura 1. Il Digital Philology Hub: le fasi del lavoro filologico descritte in L. Leonardi, "Filologia digitale del Medioevo italiano", pubblicato in Italianistica digitale = «Griseldaonline», 20, 2 (2021), pp. 77-89.

6. RINGRAZIAMENTI

Progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" Componente 2 "Dalla ricerca all'impresa" Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" Azione 3.1.1 "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti" - Codice progetto IR0000029 - CUP B63C22000730005. Soggetto attuatore CNR.

[26] CHANGES: https://sites.google.com/uniroma1.it/changes/home
[27] FROID: https://www.sns.it/it/disciplinacorso-di-laurea/corso-phd/filologia-romanza-e-italiana-digitale-froid
[28] SNS: https://www.sns.it/it

BIBLIOGRAFIA

[1] Cascianelli, Silvia, Marcella Cornia, Lorenzo Baraldi, Maria Ludovica Piazzi, Rosiana Schiuma, and Rita Cucchiara. "Learning to Read L'Infinito: Handwritten Text Recognition with Synthetic Training Data." In *Computer Analysis of Images and Patterns. CAIP 2021*, edited by Nicolas Tsapatsoulis, Andreas Panayides, Theo Theocharides, Andreas Lanitis, Constantinos Pattichis, and Mario Vento, 13053:340–350. Lecture Notes in Computer Science. Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-89131- 2_31

[2] Clérice, Thibault, Malamatenia Vlachou-Efstathiou, and Alix Chagué. "CREMMA Medii Aevi: Literary Manuscript Text Recognition in Latin." *Journal of Open Humanities Data* 9 (2023): 1–19. https://doi.org/10.5334/johd.97

[3] Coradeschi, Francesco, Leonardo Canova, Emiliano Degl'Innocent, Carmen Di Meo, Maurizio Sanedi, Alessia Spadi, and Federica Spinelli. "The RESTORE Project: A Final Review." In *Proceedings of the 19th Conference on Information and Research Science Connecting to Digital and Library Science*, edited by Alessia Bardi, Alex Falcon, Stefano Ferilli, Stefano Marchesin, and Domenico Redavid, 167–179. Bari, 2023.

[4] Degl'Innocenti, Emiliano, Monica Monachini, Alberto Bucciero, Enrico Pasini, Bruno Fasini, and Francesca Frontini. "H2IOSC: Humanities and Heritage Open Science Cloud." In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, edited by Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 63-64, 2023. https://doi.org/10.6092/unibo/amsacta/7721

[5] Dhiaf, Marwa, Ahmed Cheikh Rouhou, Yousri Kessentini, and Sinda Ben Salem. "MSdocTr-Lite: A Lite Transformer for Full Page Multi-Script Handwriting Recognition." *Pattern Recognition Letters* 169 (2023): 28–34. https://doi.org/10.1016/j.patrec.2023.03.020

[6] Favaro, Manuel, Marco Biffi, and Simonetta Montemagni. "Risorse linguistiche di varietà storiche di italiano: il progetto TrAVaSI." In *Proceedings of the Seventh Italian Conference on Computational Linguistics. CLiC-It 2020 (Bologna, Italy, March 1-3, 2021)*, edited by Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, 178–86. Torino: Accademia University Press, 2020.

[7] Favaro, Manuel, Marco Biffi, and Simonetta Montemagni. "Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione." In *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data, JADT 2022*, edited by Michelangelo Misuraca, Germana Scepi, and Maria Spano, 1:392–399. Napoli: Vadistat press, 2022.

[8] Hilbert, Martin. "How Much Information Is There in the 'Information Society'?" *Significance* 9, no. 4 (2012): 8–12.

[9] Iorio-Fili, Domenico. "Il Lemmatizzatore Semiautomatico Di GATTO4." In Dizionari e Ricerca Filologica, Atti Della Giornata Di Studi in Memoria Di Valentina Pollidori. Bollettino Dell'Opera Del Vocabolario Italiano, Supplemento III:41–56, 2010.

[10] Iorio-Fili, Domenico. "Un Nuovo Strumento Di Lemmatizzazione Automatica per Corpora Testuali Di Ridotte Dimensioni. Applicazione All'italiano Antico." *Bollettino Dell'Opera Del Vocabolario Italiano* XV (2010): 367–391.

[11] Leonardi, Lino. "Filologia Digitale Del Medioevo Italiano." *Italianistica Digitale, Griseldaonline* XX, no. 2 (2021): 77–89.

[12] Lombardi, Francesco, and Simone Marinai. "Deep Learning for Historical Document Analysis and Recognition - A Survey." *Journal of Imaging* 6, no. 10 (2020): 110. https://doi.org/10.3390/jimaging6100110

[13] Montemagni, Simonetta. "Trattamento automatico del linguaggio e Digital Humanities: metodi e strumenti, sfide." In *Digital Humanities. Metodi, strumenti, saperi*, edited by Fabio Ciotti, 160–177. Roma: Carocci, 2023.

[14] Pennacchiotti, Marco, and Fabio M. Zanzotto. "Natural Language Processing Across Time: An Empirical Investigation on Italian." In *Proceedings of GoTAL - 6th International Conference on Natural Language Processing, LNAI 5221*, edited by Bengt Nordström and Aarne Ranta, 5221:371–382. Lecture Notes in Computer Science. Gothenburg: Springer, 2008.

[15] Piotrowski, Michael. "Natural Language Processing for Historical Texts." *Synthesis Lectures on Human Language Technologies* 5, no. 2 (2012): 1–157. https://doi.org/10.2200/S00436ED1V01Y201207HLT017

[16] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.