



Atti del XIV Convegno Annuale

Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale

Verona :: 11-13 giugno 2025

A cura di:

Simone Rebora • Marco Rospocher • Stefano Bazzaco



UNIVERSITÀ
di VERONA
Dipartimento
di LINGUE
E LETTERATURE STRANIERE



ASSOCIAZIONE per
l'INFORMATICA UMANISTICA
e la CULTURA DIGITALE



ISBN 978-88-942535-9-7



Copyright ©2025 AIUCD

Associazione per l'Informatica Umanistica e la Cultura Digitale



Il presente volume e tutti i contributi sono rilasciati sotto licenza Creative Commons Attribution ShareAlike 4.0 International license (CC-BY-SA 4.0). Ogni altro diritto rimane in capo ai singoli autori.

This volume and all contributions are released under the Creative Commons Attribution Share-Alike 4.0 International license (CC-BY-SA 4.0). All other rights retained by the legal owners.

A cura di: Simone Reborà; Marco Rospocher; Stefano Bazzaco (2025). Diversity, Equity, and Inclusion: Challenges and Opportunities for Digital Humanities in the Age of Artificial Intelligence, Proceedings del XIV Convegno Annuale AIUCD, Verona 11-13 giugno 2025, Università di Verona.

Ultimo accesso agli URL in data 8 maggio 2025.

Si prega di notificare all'editore ogni omissione o errore si riscontri: segreteria [at] aiucd.org

Please notify the publisher of any omissions or errors found: segreteria [at] aiucd.org

Il programma della conferenza AIUCD 2025 è disponibile online

<https://aiucd2025.dlss.univr.it/detailed-schedule/>

The AIUCD 2025 Conference Program is available online

<https://aiucd2025.dlss.univr.it/en-gb/detailed-schedule/>

I contributi pubblicati nel presente volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante double-blind peer review, effettuata dai membri del Comitato di Programma sotto la supervisione del Comitato Scientifico di AIUCD 2025.

All the papers published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review, carried out by the members of the Programme Committee under the supervision of the Scientific Committee of AIUCD 2025.

Gli atti del convegno AIUCD 2025 sono pubblicati come raccolta di contributi in formato PDF forniti direttamente dagli autori e dalle autrici. I file sono stati raccolti e assemblati senza interventi redazionali da parte dei curatori.

The proceedings of the AIUCD 2025 conference are published as a collection of PDF contributions provided directly by the authors. The files have been collected and compiled without editorial intervention by the editors..

Il logo di AIUCD 2025 include l'immagine "Verona Dark Line Simple Minimalist Skyline With White Background" di @pabloprat/stock.adobe.com, ottenuta tramite la licenza Adobe Stock dell'Università di Verona.

The AIUCD 2025 logo includes the image "Verona Dark Line Simple Minimalist Skyline With White Background" by @pabloprat/stock.adobe.com, used under the Adobe Stock license of the University of Verona.

Il background della copertina è stato creato con tecniche di AI generativa con lo strumento "Magic Media" disponibile su Canva, usando un prompt con il tema del convegno.

The background of the cover was created using generative AI techniques with the "Magic Media" tool available on Canva, using a prompt based on the conference theme.

Comitato Organizzatore / *Organizing Committee*

General Chairs

Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Local Chair

Anna Cappellotto (Università degli Studi di Verona)

Registration Chair

Giorgia Pomarolli (Università degli Studi di Verona)

Proceedings Chair

Stefano Bazzaco (Università degli Studi di Verona)

Sponsorship Chair

Matteo Lissandrini (Università degli Studi di Verona)

Publicity Chair

Sabrina Piccinin (Università degli Studi di Verona)

Comitato Scientifico / *Scientific Committee*

Program Chairs

Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Digital Humanities e inclusione / *Inclusive DH*

Stefano Bazzaco (Università degli Studi di Verona)
Massimo Salgaro (Università degli Studi di Verona)

Archivi ed Edizioni Digitali / *Archives and Digital Editions*

Elisa Cugliana (Cologne Center for eHumanities)
Christian D'Agata (Università di Catania)

Metodi Computazionali / *Computational Methods*

Rachele Sprugnoli (Università degli Studi di Parma)
Sara Tonelli (Fondazione Bruno Kessler)

Rappresentazione di Dati e Conoscenza / *Data and Knowledge Representation*

Francesco Mambrini (Università Cattolica del Sacro Cuore)
Elena Spadini (Universität Bern)

Preservazione della Memoria e del Patrimonio Digitale / *Preservation of Memory and Digital Cultural Heritage*

Monica Berti (Universität Leipzig)
Daria Spampinato (Istituto di Scienze e Tecnologie della Cognizione-CNR)

Comitato di programma / *Program committee*

Stefano Allegrezza (Università di Macerata); Laura Antonietti (Université de Versailles Saint Quentin en Yvelines (Université Paris-Saclay); Luigi Bambaci (École pratique des hautes études, PSL); Liborio P. Barbarino (Università di Catania); Nicola Barbuti (Università degli Studi di Bari Aldo Moro (Dipartimento di Ricerca e Innovazione Umanistica); Sofia Baroncini (Leibniz Institute of European History); Andrea Bellandi (Institute for Computational Linguistics (CNR)); Mario A. Bochicchio (University of Bari, Dep.t of Computer Science); Andrea Bolioli (Independent researcher); Marco Bombieri (University of Verona); Paolo Bonora (Università di Bologna); Flavia Bruni (Università di Chieti-Pescara); Marina Buzzoni (Ca' Foscari University of Venice); Alberto Campagnolo (KU Leuven); Anna Cappellotto (Università di Verona); Emanuela Carbé (Università di Siena); Vittore Casarosa (ISTI-CNR); Raffaele Cioffi (Università di Napoli Federico II); Fabio Ciotti (Università di Roma Tor Vergata); Vincenzo Colaprice (University of Turin); Giuseppe Consolo (Università degli studi di Napoli, Federico II); Elisa Conti (Università di Catania); Salvatore Cristofaro (CNR ISTC); Giulia D'Agostino (TU Darmstadt); Elisa D'Argenio (HUN-REN Hungarian Research Centre for Linguistics); Enrico Daga (The Open University); Stefano Dall'Aglio (Università Ca' Foscari Venezia); Marilena Daquino (University of Bologna); Mauro De Bari (University of Bari Aldo Moro); Angelo M. Del Grosso (CNR-ILC); Matteo Di Franco (Università di Napoli Federico II); Giorgio Maria Di Nunzio (University of Padua); Stefano Ferilli (University of Bari); Lorenzo Ferroni (Università degli Studi di Verona); Franz Fischer (Ca' Foscari Università Ca' Foscari); Greta H. Franzini (Eurac Research); Francesca Frontini (CNR-ILC); Daniele Fusi (VeDPH, Stuttgart University); Mariangela Giglio (University of Bologna); Tiago Luis Gil (University of Brasilia); Luca Giovannini (University of Potsdam); Milena Giuffrida (Università di Catania); Edmondo Grassi (Università Telematica San Raffaele Roma); Miryam Grasso (Università di Catania); Piergiovanna Grossi (Università di Verona); Fahad Khan (CNR-ILC); Michele Lacriola (Università di Siena); Maurizio Lana (Univ. del Piemonte Orientale); Federica Lazzerini (Università degli Studi di Torino); Eleonora Litta (Università Cattolica del Sacro Cuore, Milano); Dominique Longrée (ULiège); Diego Mantoan (University of Palermo); Anna Maria Marras (University of Turin); Cristina Marras (CNR); Pietro Mazzarisi (University of Trieste); Barbara McGillivray (King's College London); Federico V. Meschini (Tuscia University); Alessio Miaschi (Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Pisa); Andrea Micheletti (University of Padua); Giulia Miglietta (Università del Salento); Paolo Monella (Università Kore di Enna); Johanna Monti (Università degli Studi di Napoli "L'Orientale"); Rossana Morriello (Università degli Studi di Firenze); Gloria Mugelli (ILC CNR); Serge Noiret (AIPH (Associazione Italiana di Public History) (European University Institute); Giuseppe Palazzolo (Università di Catania); Mafalda Papini (CNR-ILC); Enrico Pasini (UniTO/CNR-ILIESI); Giulia Pedonese (CNR (Istituto di Linguistica Computazionale "Antonio Zampolli")); Paola Peratello (Università Ca' Foscari Venezia); Federico Pianzola (University of Groningen); Chantal Pivetta (Lund University (Sweden)); Igor Pizzirusso (AIPH); Giulia Re (ILC-CNR); Giulia Renda (University of Bologna); Dario Rodighiero (University of Groningen); Roberto Rosselli Del Turco (Università di Torino); Enrica Salvatori (Università di Pisa); Emilio M. Sanfilippo (CNR); Eva Sassolini (CNR-ILC); Andrea Schimmenti (University of Bologna); Flavia Sciolette (CNR-ILC); Pietro Sichera (ILIESI-CNR); Daniele Silvi (Università di Roma 'Tor Vergata'); Giulia Speranza (University of Naples L Orientale); Francesco V. Stella (UNISI); Timothy Tambassi (Ca' Foscari University of Venice); Mirko Tavosanis (Università di Pisa); Francesca Tomasi (Università di Bologna); Simona Turbanti (University of Milano); Marco Venuti (Università di Catania); Gennaro Vessio (University of Bari Aldo Moro); Gabriele Vezzani (Università di Verona / RWTH Aachen University); Fabio Vitali (University of Bologna).

Enti organizzatori / *Organisers*

AIUCD; Università di Verona: Dipartimento di Lingue e Letterature Straniere; Digital Arena for Inclusive Humanities (DAIH).

Sommario

Prefazione

Simone Rebora, Marco Rospocher, Stefano Bazzaco

I-II

Digital Humanities and Inclusion

PrevNet. A FAIR and inclusive resource for the study of proverbs in historical languages <i>Andrea Farina, Barbara McGillivray</i>	2
Potential bias in AI: cultural representation and the marginalization of African art <i>Francesca Bignotti</i>	11
Exploring data-driven narratives in Digital Humanities web-based projects: features and impact <i>Tommaso Battisti, Marilena Daquino</i>	18
Il Glossario delle Infrastrutture di Ricerca (GIR) <i>Lucia Francalanci, Alessia Scognamiglio, Irene Falini, Pietro Restaneo, Giulia Pedonese, Alessia Spadi</i>	24
Educational Impact of Storytelling and Data Visualisation in the Interpretation of Humanities Data <i>Giulia Renda, Marilena Daquino</i>	29
IncluInstIT: Un nuovo corpus per lo studio di linguaggio inclusivo su Instagram <i>Irene Caiazza, Giovanna Maria Dimitri, Liana Tronci</i>	35
Per un'analisi della rivista Umanistica Digitale in ottica DEI <i>Rossana Morriello, Lucia Sardo</i>	40
Carpe bias, quam minimum credula queries <i>Sabato Danzilli</i>	46
Evaluating bias within an epistemological framework for AI-based research in the humanities <i>Sarah Oberbichler, Cindarella Petz</i>	52
Semplificare la lettura dei manoscritti utilizzando tecnologie WEB interattive e interazioni «hover» <i>Giacomo Marchioro, Andrea Brugnoli, Francesca Carnazzi, Paolo Pellegrini, Edoardo Ferrarini</i>	60
Accessibilità e inclusione per la documentazione del restauro: gli archivi del Centro Conservazione e Restauro «La Venaria Reale» <i>Stefania De Blasi, Edi Guerzoni, Chiara Pipino</i>	66
Il patrimonio culturale digitale delle minoranze etniche: il progetto DIGICHer tra le comunità Sámi, ladine ed ebraiche in Europa <i>Matteo Cova, Eleonora De Longis</i>	72
Fulfilling GEN-der AImS: do image-generating tools discriminate? An on-field study <i>Francesco Meledandri</i>	78
From Bias Paralysis to Bias as a Category of Analysis. Introducing the Bias-Aware Framework <i>Mrinalini Luthra, Amber Zijlma</i>	86
Supporting Children with Linguistic Vulnerabilities Through Advanced, Theory-Driven Technological Solutions: The TELMI Approach for Italian children with DLD and Children with Italian as L2 <i>Arianna Compostella, Giulia Valcamonica, Mattia Gianotti, Matteo Secco, Silvia Silleresi, Fabrizio Arosio, Franca Garzotto, Maria Teresa Guasti</i>	94
Tecnologie AI per la didattica <i>Gabriele Prospero, Giulia Miglietta, Eleonora Miccoli, Mario Bochicchio</i>	104

DEA - An Innovative Technological Tool for Personalized Linguistic Training for Italian Children with Developmental Dyslexia <i>Marta Tagliani, Maria Vender, Giulia Valcamonica, Giovanni Caleffi, Franca Garzotto, Denis Delfitto</i>	110
Il ruolo delle Infrastrutture nella costruzione di un ambiente di ricerca inclusivo. Un modello di buone pratiche <i>Marta Caradonna, Nicola Giampietro, Roberta Bianca Luzietti, Monica Monachini, Valeria Quochi, Emiliano Degl'Innocenti, Alessia Spadi, Alessandra Caravale, Antonio D'Eredità, Paola Moscati, Giacomo Mancuso</i>	118
Predicting Grammatical Cases in Slovenian Varieties in Italy: A Use Case from the LORIS 1.1 Language Assistant <i>David Bordon</i>	124
Verso un futuro senza barriere: l'accessibilità dei documenti elettronici nell'European Accessibility Act <i>Stefano Allegrezza</i>	129
La gestione del nuovo sapere digitale contemporaneo. Scenari, criticità, sfide, prospettive <i>Nicola Barbuti</i>	134

Archives and Digital Editions

Learner Corpus of Creative Writing: An interdisciplinary challenge <i>Ioanna Tyrou, Katerina Florou</i>	143
Retrieval-Augmented Generation systems for enhanced access to digital archives <i>Michele Ciletti</i>	149
Preserving Clarity: The MAGIC project approach to ancient manuscripts <i>Yahya Momtaz, Stefania Conte, Guido Russo</i>	156
Digital Explorations of Historical Multilingual Practices. The Challenges of the HyperAzpilcueta Project <i>Manuela Bragagnolo, Marcus Pöckelmann, Polina Solonets, Andreas Wagner</i>	160
Digitalizzazione di un fondo archivistico per la creazione di un centro di documentazione digitale <i>Dario Baldini</i>	165
Risorgimento Digitale: Un progetto di hyperedizione per i testi risorgimentali. Le Noterelle di Abba come caso di studio <i>Vincent Mobilia</i>	171
SpaceLat: La geografia della letteratura latina tardoantica <i>Riccardo Consolini</i>	177
Verso l'edizione digitale del carteggio Canneti-Fiacchi <i>Chiara Manca, Fiammetta Sabba, Bianca Sorbara, Silvia Tripodi</i>	184
L'edizione critica digitale della 'Scienza nuova' di Giambattista Vico in Scholarly Digital Edition <i>Alessia Scognamiglio, Roberto Evangelista, Manuela Sanna, Salvatore Prinzi, Stefano Veneroni, Chiara Aiola, Luca de Santis</i>	190
L'edizione digitale del papiro P.Tor.Choach. 12 in collaborazione con il Museo Egizio di Torino <i>Chiara Senatore</i>	197
Dal palcoscenico al digitale: modelli di data visualization per la valorizzazione dell'Archivio Teatro delle Marionette di Gianni e Cosetta Colla <i>Elena Radaelli</i>	202
Metodologie computazionali per l'organizzazione di archivi nati digitalmente <i>Mariangela Giglio</i>	208
Archivi digitali per la conservazione e valorizzazione del patrimonio culturale: il caso del Santuario della Madonna di Carufo <i>Caterina Ciccotti</i>	215

Il progetto ArPeR. Per un Archivio dei periodici romaneschi <i>Martina Ludovisi</i>	222
Un modello integrato per il Roman d'Alexandre del codice Correr 1493: annotazione linguistica e edizione critica digitale <i>Giacomo Costa, Simone Zenzaro</i>	228
Descrivere la catastrofe: documentare la diegesi per la catalogazione di opere distopiche e post-apocalittiche <i>Luca Paolo Bruno, Valeria Stabile, Juan Scassa, Carmelo Caruso, Ludovica Pannitto</i>	232
Un'applicazione pratica per l'edizione digitale di testi agiografici e calendariali <i>Luca Avellis</i>	238
Embracing flexibility: new EVT features for critical editing, accessibility and inclusivity <i>Roberto Rosselli Del Turco, Davide Cucurnia, Marina Buzzoni</i>	244
Riscoperte poliane: edizione digitale di un manoscritto inedito de Il Milione <i>Giulia Fabbris, Samuela Simion, Fabio Soncin</i>	251
«Proximior perfectioni»: criticità e future prospettive del progetto Dante Limina <i>Elisabetta Tonello</i>	257
TEI Encoding as a Unified Structure for Multilingual Digital Editions: The LeggoManzoni Case Study <i>Mariia Levchenko, Beatrice Nava, Ersilia Russo</i>	264
Verso l'implementazione di un sistema di riconoscimento di allusioni al lessico dantesco nelle testimonianze del Lager: il caso d'uso in Voci dall'Inferno <i>Carla Congiu, Angelo Mario Del Grosso, Marina Riccucci</i>	270
Il progetto RETI (REndering Texts and Images): metodologia e primi risultati <i>Chiara Barbero, Matteo Di Franco, Federica Lazzerini, Annamaria Persia</i>	276
Edizione digitale ed autorialità plurima: quali sfide? <i>Stefania Tesser</i>	283
Il corpus di prosa letteraria del progetto RIND (1830-1930). Assunti teorici e vincoli pratici <i>Stefano Ondelli, Pietro Mazzarisi</i>	289
ZoneRW: verso un'integrazione con Kraken ed eScriptorium per il riconoscimento e la gestione avanzata delle regioni di interesse <i>Pietro Sichera, Angelo Mario Del Grosso, Laura Mazzaguo, Daria Spampinato</i>	297
L'edizione digitale dei primi 16 Taccuini di Paolo Orsi <i>Giuseppina Monterosso, Andrea Bolioli, Elisa Bonacini, Gianmario Cattaneo, Dario Gonella, Anna Maria Marras, Salvatore Spina, Paola Venuti</i>	303
Artificial intelligence vs human handwriting: annotating damaged manuscripts <i>Dumitru Scutelnic, Laura Gazzani, Paolo Pellegrini, Claudia Daffara</i>	308
Modellazione, interoperabilità e riuso in DiScEPT <i>Tiziana Mancinelli, Hansmichael Hohenegger, Federico Boschetti, Angelo Mario Del Grosso, Eleonora De Longis, Gloria Mugelli</i>	314
Navigating the Digital Transition: Lessons from a Hybrid Critical Edition Project <i>Elisa Bastianello, Reto Baumgartner</i>	319
Human-LLM Synergy in Higher Education Publishing: Two ChatGPT Use Cases within Editorial Pipelines <i>Gianluca Pavani</i>	327
Taming the Hydra: A Model for Textual Dynamics and Constellations of Goethe's Venetian Epigrams <i>Daniele Fusi, Matteo Zupancic, Franz Fischer, Claus Zittel</i>	334

Computational Methods

The Influence of AI Tools on University Students' Writing Style: A Stylometric Analysis of Narrative Texts <i>Dimitris Bilianos, Katerina Florou</i>	343
Usare i Large Language Model per l'analisi del testo narrativo: strategie di prompt engineering per il riconoscimento del discorso indiretto libero nella narrativa italiana 1830-1930 <i>Aurora Argenzio, Fabio Ciotti, Anna Chiara Corradino</i>	349
Historical GIS e metodologie digitali per una storia della copertura boschiva <i>Vincenzo Colaprice</i>	357
Experiments on the Use of LLMs for the Translation of the Babylonian Talmud <i>Mafalda Papini, Davide Albanesi, David Dattilo, Emiliano Giovannetti, Simone Marchi</i>	363
Metodi di allineamento testuale bilingue per un'edizione genetica digitale dei Mémoires di Carlo Goldoni <i>Matteo Zibardi</i>	368
Eastern Law in Western Words: Analyzing Roman Legal Terminology in Medieval Charters <i>Tamás Kovács, Angelos Nikolaou, Johannes Laroche, Georg Vogeler</i>	375
Il corpus del Digesto: approcci e metodi computazionali per la creazione di risorse linguistiche <i>Alessandra Cinini, Paola Marongiu, Eva Sassolini</i>	379
Preliminary Results for the Explanation of Neural Network-based Handwriting Identification in Historical Manuscripts <i>Riccardo De Cesaris, Valerio Caravani, Arianna Pastorini, Serena Ammirati, Paolo Meriardo</i>	386
From Documents to Data: Digital Technologies in the Study of Notarial Charters <i>Franziska Decker, Sandy Aoun, Giuseppe Consolo</i>	392
Verso la svolta computazionale della critica dantesca <i>Fara Autiero, Vittorio Celotto, Gennaro Ferrante, Chiara Fusco, Sandra Gorla, Giuseppe Andrea Liberti, Mariangela Palomba, Serena Picarelli, Stefano Angelo Rizzo, Silvia Tripodi</i>	397
Phylo-1-preview. Un modello T5-Base per l'emendazione dei testi antichi <i>Giuseppe Ferrara</i>	404
«Glottolab: A Linguistic Adventure»: Lo sviluppo di un'attività gamificata incentrata sulla linguistica <i>Cecilia Cattaneo, Claudia Roberta Combei, Chiara Zanchi</i>	411
Concordanze e NLP: idee, metodi e regole per l'applicazione alla lingua italiana <i>Pietro Sichera, Christian D'Agata, Giuseppe Palazzolo</i>	419
Reverse Engineering Critical Apparatuses for HTR Ground Truth Creation: The Case of Kennicott's Collation of the Hebrew Bible <i>Luigi Bambaci, Nachum Dershowitz, Daniel Stökl Ben Ezra</i>	426

Data and Knowledge Representation

Prototyping an Atlas of Early Modern English Drama: An Experiment on DraCor Data <i>Luca Giovannini, Andreas Wagner</i>	435
ATLAS: A data model for describing FAIR Digital Humanities research outcomes <i>Chiara Martignano, Giorgia Rubin, Sebastiano Giacomini, Alessia Bardi, Marina Buzzoni, Marilena Daquino, Riccardo Del Gratta, Angelo Mario Del Grosso, Franz Fischer, Roberto Rosselli Del Turco, Francesca Tomasi</i>	440

Cantautorato e Digital Humanities. Per una valorizzazione dell'opera di Fabrizio De André, Lucio Dalla, Gianmaria Testa <i>Marcello Ranieri</i>	448
OWL Ontology on the European Integration Process between 1949 and 1979 <i>Lorenzo Galvagno</i>	453
LiITA, una Knowledge Base di risorse interconnesse per l'italiano <i>Eleonora Litta, Marco Passarotti, Paolo Brasolin, Valerio Basile, Cristina Bosco, Andrea Di Fabio</i>	460
Dai limina a LiMINA: un database per i marginalia alla Commedia <i>Serena Malatesta, Beatrice Mosca</i>	466
Dai Materiali Didattici alle Piattaforme FAIR: Costruire un'Infrastruttura di Training in H2IOSC <i>Giulia Pedonese, Francesca Frontini, Roberta Ottaviani, Federico Boschetti, Alessia Spadi, Lucia Francalanci, Alessia Scognamiglio, Pietro Restaneo, Antonina Chaban, Jana Striova, Laura Benassi</i>	473
IlluminAI: un sistema di navigazione interattivo per i manoscritti miniati rinascimentali <i>Valeria Minisini, Giorgio Gosti, Bruno Fanini</i>	478
Making Germanic Cultural Heritage accessible to students: a proposal for a case study <i>Chiara De Bastiani, Giulia Fabbris</i>	485
/DH.arc Vocabularies: Making semantic artefacts more visible and accessible using SKOS <i>Laurent Fintoni</i>	492
Modeling an Ontology for Heritage Science: Challenges and Key Strategies <i>Erica Scarpa, Riccardo Valente, Irene Rossi</i>	499
Linked Open Data and IIIF for connecting manuscripts images with their transcriptions: a case study from the Veneranda Biblioteca Ambrosiana <i>Lorenza Talarico</i>	505
Describing Monastic Iconography Using Semantic Data: A Preliminary Investigation <i>Sofia Baroncini, Francesco Mele</i>	511
A Linguistic Knowledge Graph of Word Borrowings from Portuguese <i>Anas Fahad Khan, Ana Salgado</i>	519
E.T and Visual Culture Ontology (ETVCO): Perspectives on Extraterrestrial Influence in Visual Heritage <i>Kaosaier Wusiman, Simone Casazza</i>	524
Automating XML-TEI Encoding of Unpublished Correspondence: A Comparative Analysis of two LLM Approaches <i>Marco De Cristofaro, Daniel Zilio</i>	531
Modelli e tecnologie integrate e innovative per una cittadinanza digitale equa e sostenibile <i>Cristina Marras, Vittoria Fabiani, Enrico Pasini, Lisa Reggiani, Pietro Sichera, Paolo Ongaro, Martina Rossi</i>	538
A Case Study in Cultural Heritage: A System Linking Three Open Data Tools – Digital Philology for Dummies (DPhD), Edition Visualization Technology (EVT), and a Relational Database <i>Renato Caenaro, Chantal Pivetta</i>	544
Modeling Intermediality and Interpretations in Contemporary Combinatory Literature: Revealing Il Giuoco dell'Oca by Edoardo Sanguineti <i>Enrica Bruno, Maria Francesca Bocchi, Francesca Tomasi</i>	551
From Metadata to Storytelling: A Framework For 3D Cultural Heritage Visualization on RDF Data <i>Sebastian Barzaghi, Simona Colitti, Arianna Moretti, Giulia Renda</i>	558
Between Text and Icon: Towards A Representational Model for Ekphrastic Relations <i>Maria Francesca Bocchi, Carlo Teo Pedretti, Fabio Vitali</i>	566

Preservation of Memory and Digital Cultural Heritage

Life and Death of DH Projects: A Preliminary Investigation of Their Lifecycles in Italy <i>Erica Andreose, Giorgia Crosilla, Remo Grillo, Gianmarco Spinaci</i>	575
Research on Street Art in the Digital Space <i>Aleksandra Tselikova</i>	581
Motion Visualisation of Dancers' Performances <i>Giacomo Alliata, Loïc Serafin, Samy Mannane, Sarah Kenderdine</i>	587
Entità in relazione: policies, soluzioni tecnologiche e modelli lessicali per un (eco)sistema informativo integrato <i>Herbert Natta, Michela Tardella, Eleonora Lattanzi, Gianluca Rossi, Roberta Maggi</i>	593
Preserving and enhancing cultural heritage: the Digest project <i>Alessandra Cinini, Paola Marongiu, Eva Sassolini, Monica Monachini</i>	600
The Staccioli Digital Archive: Using Knowledge Graphs to power digital art history catalogues and art exhibitions <i>Klaus Werner, Pietro Liuzzo, Alessandro Adamou</i>	608
Fantàsimè: Interactive Drama per la valorizzazione del Patrimonio Culturale <i>Maria Chiara Provenzano, Eleonora Miccoli, Mario A. Bochicchio</i>	613
Soluzioni phygital e mediazione culturale: riflessioni digiteconomiche nell'era dell'IA <i>Nicola Barbuti, Mauro De Bari</i>	619
MeMo: Una mappa letteraria digitale per la memoria del Mezzogiorno <i>Laura Giurdanella, Giuseppe Palazzolo, Bernardo De Luca, Fara Autiero, Marco Gatto, Sabatino Peluso, Concetta Maria Pagliuca, Andrea Schembari</i>	627
The relationship between art and sound: An experiment on the engagement of the cultural tourist <i>Sara Benetti, Nicola Orio</i>	633
Analisi RTI delle iscrizioni runiche del Leone del Pireo (Arsenale di Venezia) <i>Paola Peratello, Elisa Corrà</i>	639
Torino anni Ottanta. Digitalizzazione del patrimonio documentario e ricostruzione virtuale delle mostre negli spazi pubblici e privati <i>Filippo Yahia Masri</i>	646
Linguistica dei corpora e informatica umanistica per la valorizzazione plurilingue del patrimonio culturale: implementazione del progetto UniVOCItà <i>Rita Gramellini, Valeria Zotti</i>	651
«Il mio sommario dunque è tutto qui?» Per Franco Fortini <i>Emmanuela Carbé, Mariangela Giglio, Pietro Orlandi, Jacopo Maria Romano, Giulio Quaresima</i>	658

Concordanze e NLP: idee, metodi e regole per l'applicazione alla lingua italiana

Pietro Sichera¹, Christian D'Agata², Giuseppe Palazzolo³

¹CNR ILIESI, Italia pietro.sichera@cnr.it

²Università degli Studi di Catania, christian.dagata@unict.it

³Università degli Studi di Catania, giuseppe.palazzolo@unict.it

ABSTRACT (ITALIANO)

La lemmatizzazione automatica dei testi rappresenta uno strumento fondamentale nell'elaborazione del linguaggio naturale (NLP), in quanto consente di associare le occorrenze testuali ai loro lemmi, ovvero alle voci di vocabolario da cui derivano. Questo processo, essenziale per l'analisi linguistica e semantica, presenta tuttavia sfide significative quando applicato alla lingua italiana. In particolare, le peculiarità morfologiche della lingua, come la presenza di enclitiche e omografi, rendono necessaria l'adozione di approcci che coniughino tecnologie avanzate e interventi di disambiguazione personalizzati.

In questo lavoro si analizza l'uso combinato di spaCy e UDPipe, per la lemmatizzazione automatica di testi italiani, evidenziandone i punti di forza e le limitazioni. Tali strumenti sono integrati con il software LiotroConcord_v2, sviluppato su piattaforma 4D, implementando regole personalizzate per la gestione di casi complessi. Questo studio si propone di discutere l'importanza di un approccio ibrido alla lemmatizzazione, che integri metodi automatici e correttivi manuali, al fine di migliorare la qualità dell'analisi linguistica nel contesto delle edizioni digitali e della ricerca filologica. In particolare, vengono approfonditi i contributi metodologici e tecnologici del sistema sviluppato, con uno sguardo alle potenziali applicazioni future e alla possibilità di estendere tali soluzioni a ulteriori ambiti di ricerca. Nella seconda parte viene presentato un caso di studio specifico relativo alla codifica e alla lemmatizzazione de I Viceré.

Parole chiave: lemmatization; concordance; NLP; model; digital scientific editions.

ABSTRACT (ENGLISH)

Concordances and NLP: ideas, methods and rules for application to the Italian language

Automatic text lemmatization represents a fundamental tool in natural language processing (NLP), as it allows textual occurrences to be associated with their lemmas, i.e., the vocabulary items from which they are derived. However, this process, which is essential for linguistic and semantic analysis, presents significant challenges when applied to the Italian language. In particular, the morphological peculiarities of the language, such as the presence of enclitics and homographs, make it necessary to adopt approaches that combine advanced technologies and customized disambiguation interventions. This paper analyzes the combined use of spaCy and UDPipe, for automatic lemmatization of Italian texts, highlighting their strengths and limitations. These tools are integrated with LiotroConcord_v2 software, developed on a 4D platform, implementing custom rules for handling complex cases. This study aims to discuss the importance of a hybrid approach to lemmatization, integrating automatic methods and manual corrections, in order to improve the quality of linguistic analysis in the context of digital editions and philological research. In particular, the methodological and technological contributions of the developed system are discussed in depth, with a look at potential future applications and the possibility of extending such solutions to further areas of research. In the second part, a specific case study related to the encoding and lemmatization of I Viceré is presented.

Keywords: lemmatization; concordance; NLP; model; digital scientific editions

1. INTRODUZIONE

La lemmatizzazione automatica dei testi rappresenta uno dei pilastri fondamentali nell'elaborazione del linguaggio naturale (Natural Language Processing, NLP), in quanto consente di collegare ogni occorrenza testuale al suo lemma, ovvero alla forma di base da cui deriva. Questo processo risulta particolarmente importante nel contesto dell'umanistica digitale, dove l'analisi dei testi richiede una comprensione accurata delle strutture linguistiche, al fine di creare strumenti affidabili per la ricerca filologica, la costruzione di edizioni digitali e l'interrogazione semantica dei corpora.

La lingua italiana, tuttavia, presenta una serie di peculiarità che rendono la lemmatizzazione un compito particolarmente complesso. Tra queste si annoverano la presenza di enclitiche, che si legano a forme verbali generando token composti, e gli omografi, parole con la stessa forma grafica ma significati e categorie grammaticali differenti. Questi fenomeni richiedono soluzioni avanzate che vadano oltre la

semplice associazione tra forma e lemma, includendo anche la disambiguazione contestuale e la segmentazione delle forme complesse.

La difficoltà di questi problemi è ulteriormente amplificata quando le variazioni linguistiche, le forme arcaiche e la mancanza di standardizzazione rendono inapplicabili molti strumenti di NLP progettati per il linguaggio contemporaneo. Pertanto, l'adozione di sistemi combinati, che uniscano le capacità di diverse tecnologie NLP, può rappresentare una soluzione efficace per affrontare tali complessità.

Questo articolo si concentra sull'analisi e l'integrazione di due tra i più avanzati strumenti di lemmatizzazione automatica: spaCy¹ e UDPipe.² Entrambi i sistemi sono stati scelti per la loro capacità di gestire testi italiani con un alto grado di accuratezza e per la possibilità di personalizzare le pipeline di analisi linguistica. Tuttavia, per affrontare le limitazioni intrinseche di ciascun sistema, è stato sviluppato un approccio ibrido che combina i risultati di spaCy e UDPipe con interventi di disambiguazione e segmentazione realizzati tramite il software LiotroConcord_v2, una piattaforma progettata per integrare l'automazione e il controllo manuale nell'elaborazione dei testi.

L'obiettivo del presente lavoro è duplice: da un lato, evidenziare le sfide specifiche della lemmatizzazione automatica per la lingua italiana e, dall'altro, presentare un metodo pratico che sfrutti l'integrazione di tecnologie avanzate per migliorare l'affidabilità dell'analisi linguistica. In particolare, si pone l'accento sull'importanza di approcci personalizzabili per il trattamento delle enclitiche, degli omografi e delle forme linguistiche non standard, contribuendo a una migliore comprensione e accessibilità dei testi nel contesto dell'umanistica digitale.

2. STATO DELL'ARTE

Storicamente, la lemmatizzazione è stata affrontata tramite due approcci principali: basati su regole e basati su modelli statistici. Mentre i sistemi tradizionali si affidavano a dizionari e regole morfologiche, le tecnologie contemporanee sfruttano modelli di apprendimento supervisionato e reti neurali per elaborare testi con maggiore accuratezza e flessibilità.

La lingua italiana presenta peculiarità che mettono alla prova le capacità dei lemmatizzatori. Tra queste, si evidenziano le enclitiche e le forme composte, come nel caso di parole quali "trovandotelo", che contengono forme verbali con pronomi oggetto e indiretti annessi, richiedendo una segmentazione corretta per individuare i lemmi "trovare", "te" e "lo". Un ulteriore ostacolo è costituito dagli omografi, ovvero termini come "calcio", che possono avere significati e categorie grammaticali diverse a seconda del contesto, come sostantivo o verbo. Le variazioni diacroniche e arcaiche presenti nei testi letterari e storici, ad esempio "vo" invece di "vado", complicano ulteriormente l'analisi automatica. Sebbene i modelli moderni come quelli di spaCy e UDPipe riescano a gestire alcune di queste problematiche attraverso reti neurali e addestramenti su corpora diversificati, non sono privi di limitazioni. Ad esempio, i modelli pre-addestrati per l'italiano spesso non distinguono correttamente tra parole contenenti enclitiche o forme non standard, richiedendo interventi manuali o l'uso di regole personalizzate.

Uno dei contributi metodologici fondamentali nel trattamento dei testi italiani è il metodo concordanziale di Giuseppe Savoca (Savoca G., 2000), che ha gettato le basi per un'analisi testuale sistematica nell'umanistica digitale. Come descritto da Di Silvestro & al. (2022), questo approccio si concentra sull'analisi lessicografica delle parole, costruendo concordanze dove ogni occorrenza è messa in relazione con il suo lemma attraverso il contesto d'uso. Questo metodo attribuisce particolare importanza alla parola come unità fondamentale dell'analisi, tralasciando informazioni grammaticali come tempo e numero nei verbi. Tale impostazione nasce dall'esigenza di mantenere i dati gestibili nel contesto di un processo quasi completamente manuale, evitando una granularità eccessiva nella catalogazione. Sebbene il metodo non sia automatizzato, ha influenzato lo sviluppo di strumenti di umanistica digitale, evidenziando l'importanza di approcci manuali per affrontare casi complessi.

Nel presente lavoro, il metodo concordanziale trova un'applicazione moderna nell'uso del software LiotroConcord_v2, che combina strumenti di lemmatizzazione automatica (spaCy e UDPipe) con regole

1 spaCy è una libreria Python open-source, progettata per essere altamente scalabile e performante, utilizzando modelli pre-addestrati basati su reti neurali per l'analisi morfosintattica e la lemmatizzazione <https://spacy.io/> (cons. 26/01/2025)

2 UDPipe, sviluppato dall'Institute of Formal and Applied Linguistics, sfrutta i dati Universal Dependencies (UD) per addestrare modelli di analisi sintattica e morfologica su corpora annotati. La combinazione di analisi sintattica e morfologica rende UDPipe particolarmente adatto per lingue ricche di morfologia come l'italiano. <https://ufal.mff.cuni.cz/udpipe> (cons. 26/01/2025)

personalizzate implementate manualmente. Questo approccio consente di colmare le lacune dei modelli NLP, estendendo l'efficacia della lemmatizzazione anche ai testi storici e letterari. Nonostante i progressi significativi, gli strumenti attuali presentano alcune limitazioni. La mancanza di personalizzazione rimane un problema centrale, poiché i modelli pre-addestrati spesso non tengono conto delle specificità di corpora storici o letterari, come la ridotta efficacia nella risoluzione delle ambiguità semantiche, legate a fenomeni come l'omografia e la polisemia: la dipendenza dai corpora annotati limita l'applicabilità dei modelli a domini linguistici poco rappresentati, evidenziando la necessità di soluzioni più flessibili e adattabili (Ciula & al., 2023).

3. NLP E LEMMATIZZAZIONE AUTOMATICA

Il processo di lemmatizzazione automatica per la lingua italiana può presentare diverse problematiche. I sistemi NLP offrono certamente un valido supporto al processo, ma alcuni passi devono essere trattati con un automatismo limitato, come la gestione delle enclitiche, degli omografi e delle forme linguistiche non standard. Per affrontare tali complessità, il flusso di lavoro integrato sviluppato nel software LiotroConcord_v2 utilizza un approccio ibrido basato su spaCy e UDPipe, combinato con regole personalizzate e interventi manuali. Questo sistema si basa su un'analisi comparativa degli output generati dai due strumenti di NLP, con ulteriori correzioni specifiche implementate all'interno della piattaforma 4D. Il processo inizia con la preparazione del testo, che può essere fornito in formato TXT o XML-TEI. Nel caso di file XML, LiotroConcord_v2 include un modulo dedicato all'estrazione del contenuto dal tag <body>, escludendo automaticamente le sezioni ritenute non rilevanti, grazie a un sistema di esclusione basato su tag e attributi. Una volta estratto, il testo viene normalizzato attraverso una serie di trasformazioni che includono l'uniformazione dei caratteri di newline, la rimozione degli spazi ridondanti o superflui e la sostituzione di eventuali apici o virgolette non standard con caratteri uniformi. La normalizzazione garantisce un input coerente e privo di elementi che potrebbero interferire con l'analisi linguistica. La tokenizzazione viene realizzata all'interno della piattaforma 4D mediante espressioni regolari altamente personalizzate. Questo approccio permette di identificare con precisione specificità linguistiche dell'italiano, come parole contenenti enclitiche o apostrofi sia iniziali che finali, che spesso non vengono trattate adeguatamente dai sistemi NLP standard. Un esempio è dato dalla parola 'Ntoni, presente nel testo I Malavoglia. Entrambi i sistemi, spaCy e UDPipe, tendono a ignorare l'apostrofo iniziale, che segnala l'afèresi, riconoscendo il token semplicemente come Ntoni. Questa interpretazione non è corretta, poiché l'apostrofo è parte integrante del nome e ne modifica la struttura morfologica. Dopo la tokenizzazione, il testo viene analizzato separatamente sia da spaCy³ sia da UDPipe.⁴ Ciascun sistema elabora i token, assegnando loro un lemma e una categoria grammaticale (POS), e gli output generati vengono salvati in file distinti per consentire un confronto diretto.

```
if __name__ == "__main__":
    import argparse

    parser = argparse.ArgumentParser(description="Analizza un file di testo con SpaCy e UDPipe.")
    parser.add_argument("file_path", help="Percorso al file di testo da analizzare.")
    parser.add_argument("udpipe_model_path", help="Percorso al file modello UDPipe.")
    parser.add_argument("language_code", help="Codice lingua (es. 'it' o 'en').")
    args = parser.parse_args()

    analyze_text(args.file_path, args.udpipe_model_path, args.language_code)
```

Figura 1. Entry point dello script Python: esegue l'analisi di un file di testo utilizzando SpaCy e UDPipe, leggendo i parametri dalla riga di comando

Il software confronta quindi i risultati dei due sistemi. Quando entrambi concordano sul lemma e sulla categoria grammaticale, LiotroConcord_v2 accetta automaticamente l'output come valido. Tuttavia, nei casi in cui spaCy e UDPipe producono risultati discordanti, oppure quando il lemma non è presente nel vocabolario di riferimento interno a LiotroConcord_v2, il software, come vedremo nel capitolo successivo, interviene applicando regole personalizzate per risolvere le ambiguità o adattare i risultati al modello di riferimento.

³ Il modello utilizzato è it_core_news_sm

⁴ Il modello utilizzato è italian-isdt-ud-2.5

4. ERRORI DEI SISTEMI E METODI DI CORREZIONE

Un esempio di adattamento è la forma "del" che in questo modello viene ricondotta unicamente al lemma "di", senza esplicitare la combinazione articolata. Questo approccio differisce da quello adottato da spaCy e UDPipe: per spaCy, "del" è lemmatizzato come "di il ADP", mentre UDPipe separa ulteriormente il token in "di ADP"⁵ e "il DET".⁶ Per rispettare il modello di riferimento, LiotroConcord_v2 traduce automaticamente l'analisi di spaCy e UDPipe nella rappresentazione consolidata del lemma "di".

Il software effettua anche controlli specifici su differenze semantiche sottili ma sostanziali: se un verbo è preceduto da un articolo, il sistema interpreta la combinazione come un infinito sostantivato.

Come dicevamo, il sistema considera non plausibili i risultati che includono lemmi assenti dal vocabolario di riferimento, come nel caso delle enclitiche o, più in generale, delle forme o dei lemmi inesistenti.

Nel caso del token "tramutandotela", entrambi i sistemi restituiscono un lemma errato:

"tramutandotelare". Questo errore dimostra la difficoltà di segmentare correttamente i componenti morfologici, in quanto né spaCy né UDPipe riescono a riconoscere i singoli lemmi "tramutare", "te" e "la". Al contrario, LiotroConcord_v2, grazie all'implementazione di una regola personalizzata che riconosce le enclitiche come le particelle pronominali o avverbiali, scompone correttamente il token nei suoi elementi, associando ciascuno di essi al lemma e alla categoria grammaticale corrette.

```
ARRAY TEXT($sarrEnclitiche;17)
SarrEnclitiche{1}:="glie"
SarrEnclitiche{2}:="me"
SarrEnclitiche{3}:="mi"
SarrEnclitiche{4}:="te"
SarrEnclitiche{5}:="ti"
SarrEnclitiche{6}:="le"
SarrEnclitiche{7}:="lo"
SarrEnclitiche{8}:="li"
SarrEnclitiche{9}:="la"
SarrEnclitiche{10}:="ci"
SarrEnclitiche{11}:="ce"
SarrEnclitiche{12}:="vi"
SarrEnclitiche{13}:="ve"
SarrEnclitiche{14}:="si"
SarrEnclitiche{15}:="ne"
SarrEnclitiche{16}:="se"
SarrEnclitiche{17}:="gli"

: [([Lemmi]Categoria="sm") | ([Lemmi]Categoria="sf")\
  | ([Lemmi]Categoria="ag") | ([Lemmi]Categoria="di")\
  | ([Lemmi]Categoria="ie") | ([Lemmi]Categoria="in")\
  | ([Lemmi]Categoria="po") | ([Lemmi]Categoria="re"))]
  ScategoriaGiusta:="ar"
  Slemmagiusto:=""
  If ($sarrSplitted2{Si-1}#"1")
    Slemmagiusto:=Lowercase($sarrSplitted2{Si-1})
  Else
    If ([Lemmi]Categoria="sm")
      Slemmagiusto:="lo"
    End if
```

Figura 2. Frammenti di codice 4D per la gestione personalizzata delle enclitiche e degli articoli

Un caso di discordanza tra spaCy e UDPipe è rappresentato dal termine "bramosia" (in "... la vaga bramosia dell'ignoto..." da I Malavoglia). UDPipe lo analizza erroneamente come aggettivo (bramosio), mentre spaCy lo classifica correttamente come sostantivo.

Gli esempi di output permettono di comprendere l'efficacia del sistema integrato e le difficoltà che spaCy e UDPipe incontrano nell'analisi linguistica italiana. Tuttavia, grazie al confronto con il dizionario di riferimento integrato in LiotroConcord_v2, il software conferma l'interpretazione corretta, avendo già catalogato il lemma "bramosia" e le relative forme "bramosia" e "bramosie", eliminando così ogni ambiguità.

5. UN CASO DI STUDIO: I VICERÉ DALL'OCR ALL'OUTPUT DELLA LEMMATIZZAZIONE IN XML-TEI

Un caso di studio utile a comprendere il workflow completo (cfr. Figura 3) del sistema di lemmatizzazione è quello dell'edizione scientifica digitale commentata de *I viceré* di Federico De Roberto, promossa all'interno del progetto [PROJECT].

⁵ "ADP" sta per *Adposition*; in italiano corrisponde alle preposizioni "di", "a", "da" ...

⁶ "DET" sta per *Determiner*; in italiano corrisponde a varie categorie grammaticali: articoli determinativi (es. "il", "la", "i", "le"), articoli indeterminativi (es. "un", "una"), pronomi dimostrativi usati come determinanti (es. "questo", "quello"), determinanti indefiniti (es. "alcuni", "ogni").

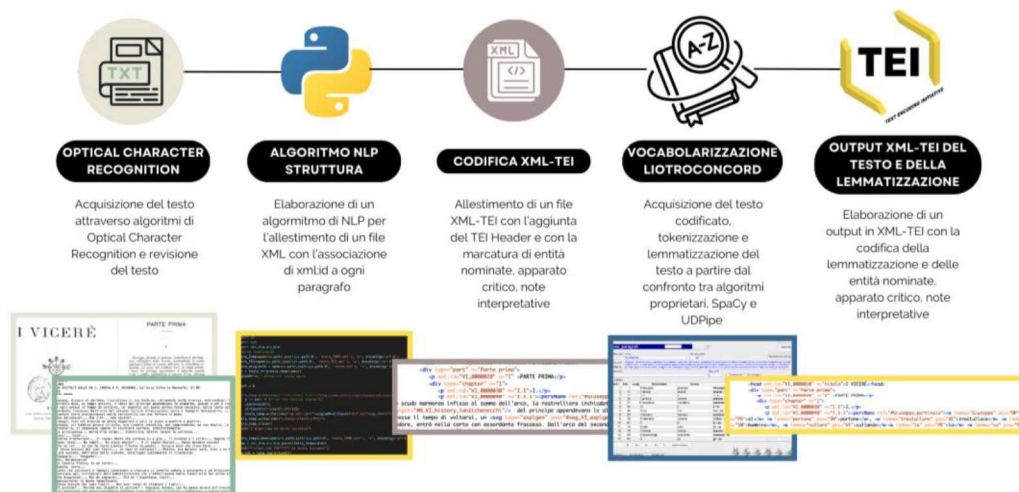


Figura 3. Workflow per la codifica e vocabolarizzazione dei Viceré

A partire dall'acquisizione delle immagini ad alta definizione dell'edizione Galli del 1894 e del conseguente riconoscimento automatico dei caratteri (OCR) è stato prodotto un documento in formato solo testo a cui è stato applicato un algoritmo di *Natural Language Processing* in Python. In particolare, sono state utilizzate le *regular expressions* e il metodo *readline* per costruire la struttura del documento XML associando un `@xml:id` univoco (fondamentale anche per associare le righe nel sistema di lemmatizzazione). Infine, attraverso l'API del *Document Object Model* è stato *parsato* il testo producendo il documento xml, poi marcato manualmente in linea con il modello di codifica promosso dalla *Text Encoding Initiative* (TEI). In particolare, sono state marcate le entità nominate (`<persName>`, `<placeName>`, `<orgName>`, `<eventName>`), i riferimenti a personaggi, luoghi, organizzazioni ed eventi (`<rs>`), alcune parole in disuso o di difficile comprensione (`<distinct>`), il discorso diretto (`<q>`), alcuni segmenti testuali a cui associare delle note di esegesi (`<seg>`) e infine alcune note di carattere storico-bibliografico o interpretativo (con dei *pointer* `<ptr>` e delle `<note>`). Si veda come esempio la codifica del primo paragrafo del primo capitolo:

```
<p xml:id="VI_0000040" n="I.1.1"><persName ref="#Giuseppe_portinaio">Giuseppe</persName>,
dinanzi al portone, trastullava il suo bambino, cullandolo sulle braccia, mostrandogli lo
scudo marmoreo infisso al sommo dell'arco, la rastrelliera inchiodata sul muro del
vestibolo dove, ai tempi antichi, i <distinct ana="lanzo">lanzi</distinct><ptr
type="history" target="#N_VI_history_lanzicheneccchi"/> del principe appendevano le
alabarde, quando s'udi e crebbe rapidamente il rumore d'una carrozza arrivante a tutta
carriera; e prima ancora che egli avesse il tempo di voltarsi, un <seg type="explgen"
ana="#seg_VI_explgen_legnetto">legnetto</seg> sul quale pareva fosse nevicato, dalla tanta
polvere, e il cui cavallo era tutto spumante di sudore, entrò nella corte con assordante
fracasso. Dall'arco del secondo cortile affacciaronsi servi e famigli: <persName
ref="#BaldassarreCrimi">Baldassarre</persName>, il maestro di casa, schiuse la vetrata
della loggia del secondo piano, intanto che <persName ref="#SalvatoreCerra">Salvatore
Cerra</persName> precipitavasi dalla carrozzella con una lettera in mano.</p>
```

Il documento così marcato è stato acquisito in *input* da LiotroConcord che, dopo la *tokenizzazione*, ha associato il lemma e il POS corrispondente a ogni forma in base al confronto tra il modello utilizzato e quelli promossi da SpaCy e UDPipe, come detto nei paragrafi precedenti. La *Graphical User Interface* (GUI) di LiotroConcord (cfr. Figura 4) permette dunque al concordatore di intervenire correggendo gli omografi, le associazioni o *tokenizzazioni* errate, i neologismi, ecc. Un importante aspetto riguarda anche la decisione del concordatore su alcuni aspetti che necessitano di una decisione interpretativa (come la scelta tra aggettivo e participio nei participi passati, tra articolo e numerale negli indeterminativi, ecc.).

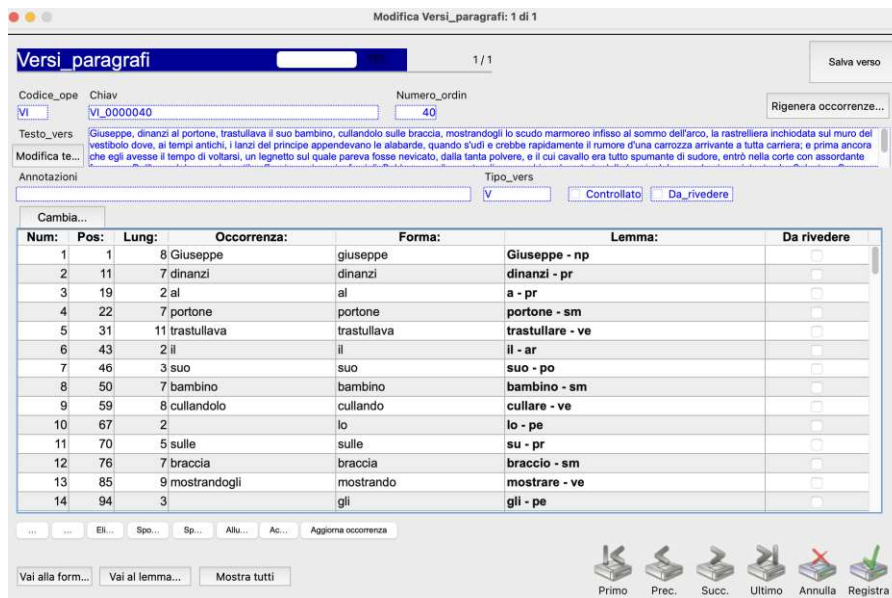


Figura 4. Schermata di LiotroConcord con la lemmatizzazione del primo paragrafo de I Viceré

Infine, la lemmatizzazione viene integrata nella codifica attraverso il tag <w> e gli attributi @lemma e @pos, così da rendere i dati aperti e interoperabili. Si veda dunque l'output arricchito del primo paragrafo:

```
<p xml:id="VI_0000040" n="I.1.1"><persName ref="#Giuseppe_portinaio"><w lemma="Giuseppe"
pos="NP">Giuseppe</w></persName>, <w lemma="dinanzi" pos="PR">dinanzi</w> <w lemma="a"
pos="PR">al</w> <w lemma="portone" pos="SM">portone</w>, <w lemma="trastullare"
pos="VE">trastullava</w> <w lemma="il" pos="AR">il</w> <w lemma="suo" pos="PO">suo</w> <w
lemma="bambino" pos="SM">bambino</w>, <w lemma="cullare" pos="VE">cullando</w><w
lemma="lo" pos="PE">lo</w> <w lemma="su" pos="PR">sulle</w> <w lemma="braccio"
pos="sm">braccia</w>, <w lemma="mostrare" pos="VE">mostrando</w><w lemma="gli"
pos="PE">gli</w> <w lemma="lo" pos="AR">lo</w> <w lemma="scudo" pos="SM">scudo</w> <w
lemma="marmoreo" pos="AG">marmoreo</w></p>
```

La possibilità di far dialogare realmente i dati della codifica (come quello sul discorso diretto o sulle entità nominate) e la lemmatizzazione diventa potenzialmente non soltanto un arricchimento della codifica, ma una vera e propria occasione per effettuare interrogazioni complesse con lo sviluppo di concordanze selettive (per capitoli, parti, personaggi) aprendo a un nuovo modo di intendere la vocabolarizzazione, la codifica, *distant* e *close reading* e l'applicazione ermeneutica di tali teorie e modelli.

6. CONCLUSIONI

Questo lavoro ha dimostrato come un approccio ibrido, che integra spaCy, UDPipe e il software personalizzato LiotroConcord_v2, possa affrontare efficacemente le complessità linguistiche della lingua italiana, garantendo una lemmatizzazione accurata grazie all'uso di tecnologie avanzate, combinato con regole manuali mirate.

I prossimi sviluppi prevedono l'addestramento di modelli NLP basati sulle concordanze già gestite da LiotroConcord_v2. Questo processo potrebbe migliorare ulteriormente l'accuratezza degli strumenti esistenti, affinando la capacità di spaCy e UDPipe di gestire le specificità della lingua italiana. I nuovi modelli, addestrati su un corpus così ricco e curato, potrebbero ridurre gli errori di lemmatizzazione e disambiguazione, offrendo risultati più precisi.

BIBLIOGRAFIA

Alishahi, A., Chrupała, G., & Linzen, T. (2019). Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, 25(4), 543–557. <https://doi.org/10.1017/S135132491900024X>.

Blaise, A. (2022). Do Large Language Models Understand Us? *Daedalus*, 151(2), 183–197. https://doi.org/10.1162/daed_a_01909.

- Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6, 213–234. <https://doi.org/10.1146/annurev-linguistics-011619-030303>.
- Borges, J. L. (1998). The library of Babel. In *Collected Fictions*.
- Ciotti, F. (2023). "La codifica del testo, XML e la TEI", in *Digital Humanities. Metodi, strumenti, saperi*. Carocci.
- Ciula, A., Eide, Ø., Marras, C., & Sahle, P. (2023). Modelling between digital and humanities: Thinking in practice. Open Book Publishers. <https://doi.org/10.11647/OBP.0369>.
- Clarke, A. C. (1967). *The nine billion names of God*. Harcourt.
- Di Silvestro, A., D'Agata, C., Palazzolo, G., & Sichera, P. (2022). Conservazione e fruizione di banche dati letterarie: L'archivio della poesia italiana dell'Otto/Novecento di Giuseppe Savoca. In F. Ciraci, G. Miglietta, & C. Gatto (Eds.), *Quaderni di Umanistica Digitale. AIUCD 2022 - Proceedings* (pp. 1–12). AIUCD. <https://doi.org/10.6092/unibo/amsacta/6848>.
- Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models and other generative models. *Philosophical Technology*, 36, 15. <https://doi.org/10.1007/s13347-023-00621-y>.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.
- Hassija, V., Chamola, V., Mahapatra, A., et al. (2024). Interpreting Black-Box Models: A review on explainable artificial intelligence. *Cognitive Computation*, 16, 45–74. <https://doi.org/10.1007/s12559-023-10179-8>.
- Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4129–4138). New York: Association for Computational Linguistics.
- Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., et al. (2023). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *bioRxiv*. <https://doi.org/10.1101/2022.06.08.495348>.
- Lodder, J. (2009). Binary Arithmetic: From Leibniz to von Neumann. <https://doi.org/10.5948/UPO9780883859742.023>.
- Mickus, T., Paperno, D., & Constant, M. (2022). How to dissect a muppet: The structure of transformer embedding spaces. *Transactions of the Association for Computational Linguistics*, 10, 981–996. https://doi.org/10.1162/tacl_a_00501.
- Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., et al. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(1), 1–40. <https://doi.org/10.1145/3605943>.
- Palazzolo, G. (2023). Lessicografia letteraria e metodo concordanziale: Storia, modelli e sfide per le Digital Humanities. *Linguistica e Letteratura*, 48(1–2), 23–45. <https://digital.casalini.it/10.19272/202301602006>.
- Primiero, G. (2020). *On the Foundations of Computing*. Oxford: Oxford University Press.
- Savoca, G. (2000). *Lessicografia letteraria e metodo concordanziale*. Firenze: Olschki.
- Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Somerset, NJ: Association for Computational Linguistics.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>.