

Atti del XV Convegno Annuale **AIUCD**

Digitale *e Public Engagement*

Pratiche e prospettive nelle Digital
Humanities

Cagliari 3-4-5 giugno 2026

a cura di

Cristina **Marras** | Andrea **Pergola** | Giampaolo **Salice**

ISBN 9791298618817



Copyright ©2026 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: Cristina Marras; Andrea Pergola; Giampaolo Salice (2026). *Digitale e Public Engagement: pratiche e prospettive nelle Digital Humanities*, Atti del XV Convegno Annuale AIUCD, Cagliari 3-5 giugno 2026, Università degli Studi di Cagliari.

Ultimo accesso agli URL in data 13 maggio 2026.

Last URL access May, 13 2026.

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 del Convegno AIUCD 2026 è disponibile online:

<https://www.aiucd2026.unica.it/companion/>

The AIUCD 2026 Conference Program is available online:

<https://www.aiucd2026.unica.it/companion/>

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 2026.

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 2026.

Gli Atti del Convegno AIUCD 2026 sono pubblicati come raccolta dei contributi forniti direttamente dagli autori e dalle autrici. I file sono stati raccolti e assemblati senza interventi redazionali significativi da parte dei curatori.

The proceedings of the AIUCD 2026 Conference are published as a collection of contributions provided directly by the authors. The files have been collected and compiled without significant editorial intervention by the editors.

Il logo di AIUCD 2026 è opera di Raffaele Argiolas

The AIUCD 2026 logo was designed by Raffaele Argiolas

La copertina è stata realizzata da Giampaolo Salice

The cover was created by Giampaolo Salice

GENERAL CHAIRS

Giampaolo Salice | Università degli Studi di Cagliari | DH UNICA

Cristina Marras | CNR-ILIESI | AIUCD

COMITATO ORGANIZZATORE / ORGANIZING COMMITTEE

Il Comitato Organizzatore è composto da Giampaolo Salice (Università degli Studi di Cagliari), Cristina Marras (CNR-ILIESI), Raffaele Argiolas, Alessandro Capra, Giommara Carboni, Andrea Pergola, Eleonora Todde (DH UNICA Università degli Studi di Cagliari), Christian D'Agata (Università di Catania), Francesca Frontini (CNR-ILC), Simone Rebora (Università di Verona), Marco Rospocher (Università di Verona), Laura Stochino (ISSASCO), Matteo Tatti e Alice Nozza (Associazione Itzokor).

Iscrizioni / Conference Registration

Eleonora **Todde** – DH UniCa – Università degli Studi di Cagliari

Atti / Conference Proceedings

Andrea **Pergola** – DH UniCa – Università degli Studi di Cagliari

Comunicazione / Communication

Raffaele **Argiolas** – DH UniCa – Università degli Studi di Cagliari

Promozione / Promotion

Giommara **Carboni** – DH UniCa – Università degli Studi di Cagliari

Tecnologo dell'Evento / Event Technologist

Alessandro **Capra** – DH UniCa

Organizzazione locale

Dipartimento di Lettere, Lingue e Beni Culturali, Università degli Studi di Cagliari, Ufficio Amministrativo:
Serena Serra, Giulia Cadoni, Claudia Serri

COMITATO SCIENTIFICO / SCIENTIFIC COMMITTEE

Il Comitato Scientifico del convegno AIUCD 2026 è composto dai Chairs di Programma e dai Chairs delle cinque aree tematiche.

Program Chairs

Giampaolo Salice | Università degli Studi di Cagliari | DH UNICA

Cristina Marras | CNR-ILIESI | AIUCD

DH e co-costruzione del sapere con le comunità: sfide, metodi, strumenti / *DH and co-construction of knowledge with communities: challenges, methods, tools*

Arianna Ciula | King's Digital Lab, London

Greta Franzini | EURAC, Bolzano

Archivi ed edizioni: descrizioni aumentate, accessibilità e sistemi informativi / *Archives and editions: augmented descriptions, accessibility, and information systems*

Federico Valacchi | Università di Macerata

Paolo Monella | Università Kore Enna | AIUCD

Testualità digitali: prospettive, sviluppi e sperimentazioni / *Digital textualities: perspectives, developments, and experimentations*

Simone Rebora | Università di Verona | AIUCD

Alessandro Adamou | Biblioteca Hertziana

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

Francesca Tomasi | Alma Mater Studiorum – Università di Bologna

Angelo Mario Del Grosso | CNR-ILC

Memorie, storia e patrimoni culturali digitali / *Tangible and Intangible Digital Heritage*

Michela Tardella | CNR-ILIESI

Enrica Salvatori | Università di Pisa

COMITATO DI PROGRAMMA / PROGRAM COMMITTEE

Stefano Allegrezza (Università di Macerata), Laura Antonietti (Université de Versailles Saint Quentin en Yvelines - Université Paris-Saclay), Alessio Antonini (Open University), Liborio P. Barbarino (Università di Catania), Nicola Barbuti (Università di Bari "Aldo Moro"), Stefano Bazzaco (Università di Verona), Andrea Bellandi (CNR-ILC), Giulia Benotto (CNR-ILC), Monica Berti (Leipzig University), Mario A. Bochicchio (Università di Bari "Aldo Moro"), Andrea Bolioli (Ricercatore indipendente), Marco Bombieri (Università di Verona), Flavia Bruni (Università di Chieti-Pescara), Marina Buzzoni (Università Ca' Foscari di Venezia), Alberto Campagnolo (KU Leuven), Vittore Casarosa (CNR-ILC), Raffaele Cioffi (Università di Napoli Federico II), Vincenzo Colaprice (Università di Torino), Francesca Congiu (Università degli Studi di Cagliari), Giuseppe Consolo (Università degli studi di Napoli, Federico II), Christian D'Agata (Università di Catania), Giulia D'Agostino (TU Darmstadt), Elisa D'Argenio (HUN-REN Hungarian Research Centre for Linguistics), Davide Dainese (Alma Mater Studiorum Università di Bologna), Stefano Dall'Aglio (Università Ca' Foscari Venezia), Marilena Daquino (Alma Mater Studiorum Università di Bologna), Mauro De Bari (Università di Bari "Aldo Moro"), Angelo M. Del Grosso (CNR-ILC), Matteo Di Cristofaro (UniMore), Francesca Di Donato (CNR-ILC), Giorgia Di Marcantonio (Università di Macerata), Giorgio Maria Di Nunzio (Università di Padova), Roberto Evangelista (CNR-ISPF), Stefano Ferilli (Università di Bari "Aldo Moro"), Franz Fischer (Università Ca' Foscari di Venezia), Greta H. Franzini (EURAC), Francesca Frontini (CNR-ILC), Daniele Fusi (Università Ca' Foscari di Venezia), Mariangela Giglio (Alma Mater Studiorum Università di Bologna), Tiago Luis Gil (University of Brasilia), Michela Giordano (Università degli Studi di Cagliari), Luca Giovannini (University of Potsdam), Milena Giuffrida (Università di Catania), Giovanna Granata (Università degli Studi di Cagliari), Edmondo Grassi (Università Telematica San Raffaele Roma), Miryam Grasso (Università di Catania), Piergiovanna Grossi (Università di Verona), Marina M. Guglielmi (Università degli Studi di Cagliari), Alessandro Iannella (Università degli Studi di Milano), Sabrina Iorio (Università Napoli Federico II), Michele Lacriola (Università di Siena), Alessandro Laruffa (ISSTOR), Anna Mambelli (UniMore), Francesco Mambrini (Università Cattolica del Sacro Cuore), Francesco Mamei (Università degli Studi di Cagliari), Lorenzo Mancini (CNR-ILIESI), Anna Maria Marras (Università di Torino), Cristina Marras (CNR-ILIESI), Pietro Mazzarisi (Università di Trieste), Barbara McGillivray (King's College London), Federico V. Meschini (Università della Tuscia), Alessio Miaschi (CNR-ILC), Giulia Miglietta (Università del Salento), Paolo Monella (Università Kore di Enna), Rossana Morriello (Università degli Studi di Firenze), Giulia Murgia (Università degli Studi di Cagliari), Enrico Natale (infoclio.ch), Sebastiana Nocco (CNR-ISEM), Serge Noiret (AIPH), Giuseppe Palazzolo (Università di Catania), Mafalda Papini (CNR-ILC), Enrico Pasini (Università di Torino), Giulia Pedonese (CNR-ILC), Paola Peratello (Università Ca' Foscari Venezia), Andrea Pergola (Università degli Studi di Cagliari), Ginevra Peruginelli (CNR-IGSG) University of Groningen, Federico Pianzola (University of Groningen), Fabio C. Pinna (Università degli Studi di Cagliari), Igor Pizzirusso (AIPH), Tiziana Pontillo (Università degli Studi di Cagliari), Valeria Quochi (CNR-ILC), Lisa Reggiani (CNR-ILIESI), Giulia Renda (Alma Mater Studiorum Università di Bologna), Pietro Restaneo (CNR-ILIESI), Dario Rodighiero (University of Groningen), Roberto Rosselli Del Turco (Università di Torino), Federica Rovelli (Università di Pavia), Giampaolo Salice (Università degli Studi di Cagliari), Enrica Salvatori (Università di Pisa), Emilio M. Sanfilippo (CNR-ISTC), Eva Sassolini (CNR-ILC), Manfredi Scanagatta (UniMore), Andrea Schimmenti (Alma Mater Studiorum Università di Bologna), Flavia Sciolette (CNR-ILC), Alessia Scognamiglio (CNR-ISPF), Luigi Serra (CNR-ISEM), Pietro Sichera (CNR-ILIESI), Daria Spampinato (CNR-ISTC), Giulia Speranza (Università di Napoli "L'Orientale"), Rachele Sprugnoli (Università di Parma), Francesco V. Stella (Università di Siena), Timothy Tambassi (Università Ca' Foscari di Venezia), Mirko Tavosanis (Università di Pisa), Francesca Tomasi (Alma Mater Studiorum Università di Bologna), Simona Turbanti (Università di Milano), Nicoletta Usai (Università degli Studi di Cagliari), Marco Venuti (Università di Catania), Gennaro Vessio (Università di Bari "Aldo Moro"), Gabriele Vezzani (Università di Verona), Simone Zenzaro (CNR-ILC).

Enti organizzatori / Organisers

Il XV convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) è organizzata dal DH UNICA - Centro Interdipartimentale per l'Umanistica Digitale e dal Dipartimento di Lettere, Lingue e Beni Culturali dell'Università di Cagliari in collaborazione con AIUCD.

Sommario

PREMESSA

I-IV

Cristina Marras, Andrea Pergola, Giampaolo Salice

[1/A] INTELLIGENZA ARTIFICIALE GENERATIVA E RIFLESSIONE TEORICA

- Una Agenda sull'AI generativa per Digital Humanities 6
Fabio Ciotti
- L'errore dell'IA come euristica per una deliberazione assistita 14
Valerio De Luce; Tiberio Uricchio
- Verso una memoria artificiale? Il progetto Johannes tra questioni tecnologiche, archeologiche e morali 21
Stefano Bertoldi
- AI confini della realtà: il pensiero umano nell'epoca della sua riproducibilità generativa 31
Daniele Silvi

[1/B] CORPORA DIGITALI: REQUISITI E SOLUZIONI GESTIONALI

- Verba volant, emails manent*: una proposta di metodologia per l'acquisizione e la preservazione degli archivi di posta elettronica di personaggi illustri 38
Alberto Abis, Stefano Allegrezza, Fabrizio Stupino
- Sfide alla sostenibilità dei progetti di digitalizzazione: i casi studio degli archivi manicomiali calabresi 45
Emanuela Nicole Donato, Maria Natalia Federico, Grazia Serratore
- A technical solution for a digital cultural heritage corpus: Notes on collaboration, requirements and sustainability 53
Arianna Ciula, Miguel Vieira, Geoffroy Noel, Neil Jakeman, Zihao Lu, Tiffany Ong, Simona Stoyanova, Jonathan Prag, Alessia Coccato, Ryan Heuser

[1/C] CO-PRODUZIONE DI ARCHIVI E COLLEZIONI DI COMUNITÀ

- L'Archivio Digitale di Vigne Nuove. Costruire un archivio di comunità, tra teoria e pratica 62
Manfredi Scanagatta
- Participatory Digital Archives as a Method for Co-constructing Cultural Memory: The Bunjevke Women's Platform 68
Sandra Iršević PhD
- Il ritorno dei partigiani sardi. Una collezione digitale co-prodotta 73
Walter Falgio, Laura Stochino, Matteo Quarantiello
- Dalla materialità delle fonti al dato condiviso: metodi ibridi per un archivio digitale partecipativo 78
Tiziana Pasciuto

[2/A] LARGE LANGUAGE MODELS: APPLICAZIONI UMANISTICHE

Scholarly Opinion Mining using LLMs and Knowledge Graphs: the case of the Van den vos Reynaerde	87
Andrea Schimmenti, Joris van Zundert, Fabio Vitali, Marieke van Erp	
Un LLM per le Scienze Umane e Sociali? ReSearch_SSH come esperimento europeo	95
Adam Faci, Alessio Miaschi, Anne Combe, Pascal Cuxac ⁴ , Francesca Frontini, Nicolas Larrousse, Stéphane Pouyllau	
L'AI a servizio della Musicologia: costruzione di thesauri musicologici con il supporto dei Large Language Models	102
Manuela Grillo, Paolo Bonora	
Per una formazione del filologo computazionale del futuro: Domain Specific Language e Large Language Model	107
Christian D'Agata, Angelo Mario Del Grosso, Federico Boschetti	

[2/B] OPEN SCIENCE, SOSTENIBILITÀ E PRATICHE APERTE

Are Digital Humanities really committed to open? An exploratory study on the availability of methodological workflows and open peer review practices	116
Silvio Peroni	
Monitorare la transizione: l'Osservatorio sulle pratiche di Open Science nelle SSH come strumento di Public Engagement	122
Pamela Barletta, Marta Caradonna, Nicola Giampietro, Alice Orrù, Federico Silvestri	
Nel 25° anniversario di Wikipedia: l'ecosistema Wikimedia e le Digital Humanities tra pratiche collaborative e sostenibilità della conoscenza	130
Piergiovanna Grossi	
Una pausa e tre domande. La crisi della memoria nell'era dell'accesso	137
Enrica Salvatori, Federico Valacchi	

[2/C] MODELLI DIGITALI PER GESTIRE E CONDIVIDERE FONTI MULTIDISCIPLINARI

Digital Repatriation in Practice: Sharing Ethnographic Photography with Communities of Origin at the Náprstek Museum	142
Daniela Šnapková	
Una storia da ascoltare. Descrivere, mappare, connettere: un modello per la valorizzazione degli archivi musicali	147
Martina Gremignai, Luca Andrea Ludovico	
LETSDigit: Archivio digitale della letteratura a Trieste	154
Alina Jill Simeone, Cristina Fenu, Marina Buzzoni, Roberto Rosselli Del Turco	
La mostra <i>Coerenza In Coerenza</i> (1984) oltre la fotogrammetria: ricostruzione digitale tramite reti neurali	161
Filippo Yahia Masri, Maria Carmelita	

[3/A] AI E NLP PER TESTI E RECUPERO DELL'INFORMAZIONE

Creating a new gender-fair Italian morphological classifier based on UmBERTo fine-tuning	168
Irene Caiazzo, Giovanna Maria Dimitri	
Assessing LLM Capabilities in Cultural Heritage: Recent Trends and Gaps in Benchmarks	175

Remo Grillo, Gianmarco Spinaci, Lukas Klic, Giovanni Colavizza

L'intelligenza artificiale per il potenziamento dell'information retrieval nei discovery tool delle biblioteche 183

Ilaria Belvedere, Simona Turbanti

Sinergie ludonarrative: l'evoluzione del racconto videoludico tra divulgazione, tecnica ed intelligenza artificiale 191

Giulia Angelini, Tommaso Mazzoli

[3/B] DH PER STORIA, DIRITTO E ARTE

Archivi ibridi di persona e preservazione forense dei materiali nativi digitali 198

Mariangela Giglio, Adele Gorini

Digital Humanities e studi neogreci: il progetto della Biblioteca Digitale dell'Osservatorio "Mario Vitti" 204

Georgios Katsantonis

artresearch.net: A Research Infrastructure for Exploring Interpretive Plurality in Art-Historical Photo Archives 211

Marilena Daquino, Francesca Mambelli, Alessandro Adamou, Pietro Maria Liuzzo, Stefanie Schneider, Spyros Koulouris, Artem Kozlov, Rafael Brundo Uriarte

Legislative Changes Concerning the Reproduction of Images of Italian Cultural Heritage Assets between 2023 and 2025 217

Marco Ciurcina, Piergiovanna Grossi

[3/C] ANALISI COMPUTAZIONALE DEL TESTO LETTERARIO

Flaiano postcoloniale? Un approccio computazionale alla *vexata quaestio* di *Tempo di uccidere* 226

Silvia Lilli, Daniel Raffini

Verso una rappresentazione del macrotesto: filologia d'autore e analisi quantitativa nei racconti di Fausta Cialente 234

Emmanuela Carbè

Analisi fluidodinamica dei campi semantici in testi letterari: il Numero di Reynolds come modello di turbolenza testuale 241

Pietro Sichera, Carla Perrone, Antonio Sichera

Dentro l'Officina del Testo: Modellare Strutture Letterarie Complesse 248

Enrica Bruno

[4/A] STUDI LETTERARI DIGITALI E DISTANT READING

Tracing Intertextuality at Scale. A Network Study of Poetic Influence 255

Gabriele Vezzani

The Landscape of Italianistica: Italian Literary Studies and Digital Humanities in Semantic Space 261

Maria Levchenko

L'Edicola della Storia di Trieste: NLP e *distant reading* Per il confronto tra periodici storici italiani e sloveni nella Trieste del 1902 267

Cristina Fenu, Giovanni Pinna

Literary Tropes as Prescriptive Devices: A Case Study from BookTok 274

[4/B] MUSEI E ACCESSO DIGITALE

- Rigenerare un patrimonio desueto con il phigital: la Casa Museo Bartolo Longo 282
Mauro De Bari
- Musei senza confini: il digitale come infrastruttura di memoria partecipata nei musei delle aree interne. Il GeoMuseo Monte Arci Stefano Incani di Masullas e il Museo Paleontologico PARC di Genoni 290
Paola Palmas
- Pubblici remoti: la Realtà Virtuale come infrastruttura culturale di accesso 296
Elisa Smeraldo

[4/C] FORMAZIONE, DIDATTICA E LABORATORI DH

- Gli Hunger Games del LabCD. Un diario di sopravvivenza 304
Enrica Salvatori, Vittore Casarosa, Francesco Ricciardi
- It's Buildin' Time: A Project-based Approach for Teaching Data Management to Cultural Heritage Graduate Students 309
Sebastian Barzagli, Alba d'Elia, Esther Montserrat Giordano, Giulia Guidarelli, Gianluca Petrosillo, Elisabetta Sabattini, Fiammetta Sabba, Lucia Sardo
- Integrazione di tecnologie per la traduzione e CLIL nella creazione di risorse didattiche per il sardo 319
Gianfranco Fronteddu, Igor Deiana

[5/A] LINKED OPEN DATA, MODELLI SEMANTICI, INFRASTRUTTURE

- Reti sociali e linked data: persone e co-occorrenze documentarie nel Portale delle fonti per la storia della Repubblica italiana 329
Herbert Natta, Gianluca Rossi, Roberta Maggi
- Web Portals on Linked Open Data: Testing Sampo-UI in Projects on Early Modern Bibliographic Data and Chinese Calligraphy and Arts 336
Arianna Moretti, Katarina Lučić, Iiro Lassi Ilmari Tiihonen, Jonas Paul Fischer
- Risorse Terminologiche e Linked Open Data: il *Glossario delle Infrastrutture di Ricerca (GIR)* come Caso di Studio 345
Lucia Francalanci, Alessia Scognamiglio, Giulia Pedonese, Michele Mallia, Irene Falini, Fahad Khan
- Orchestrare servizi e scenari avanzati: fruizione e necessità delle comunità di ricerca nel marketplace di H2IOSC 352
Pietro Sichera, Cristina Marras, Enrico Pasini, Vittoria Fabiani, Paolo Ongaro, Michele Scapicchi, Chiara Di Pietro

[5/B] PATRIMONIO CULTURALE: SISTEMI INFORMATIVI INTEGRATI

- Architettura e *Digital Humanities*. Conoscenza e valorizzazione del Patrimonio attraverso la costruzione di sistemi informativi integrati 361
Donatella Rita Fiorino, Caterina Giannattasio, Elisa Pilia, Valentina Pintus, Andrea Pirinu, Marcello Schirru
- Tra testo e architettura digitali per lo studio delle trasformazioni storico-artistiche dei monasteri benedettini 368

Gianmario Guidarelli, Paolo Borin, Sonia Cavicchioli, Chantal Pivetta, Renato Canearo,	
Dalla pergamena al virtuale. Manoscritti miniati nello spazio tridimensionale	375
Valeria Minisini, Bruno Fanini, Giorgio Gosti	
Modeling and visualizing palimpsests with IIIF Presentation API 3.0	382
Panagiotis Leontaridis, Giacomo Marchioro, Glen Robson	

[5/C] CROWDSOURCING E PUBLIC ENGAGEMENT

<i>Tag ANIMALx</i> : Crowdsourcing the Non-Human in Cultural Heritage Collections	389
Daniela Teixeira Gomes, Carla Vieira, Joana Vieira Paulino	
From perception to memory: Public Engagement and Neurohumanistic Approaches to the Collective Reconstruction of Archaeological Heritage	396
Grazia Solenne, Vincenza Ferrara, Elisa Corrò	
Comunicare la ricerca archeologica nell'ecosistema digitale: pratiche, metodi e sperimentazioni su Instagram	404
Marco Demuru	
Public engagement nell'era dei synthetic media: per una co-interpretazione critica tra Digital Humanities e Media Education	411
Gabriele Prosperi, Mario A. Bochicchio	

[6/A] MODELLAZIONE TESTUALE E LINGUISTICA COMPUTAZIONALE

Verso un dizionario narrativo computazionale degli usi linguistici in aree ad alta vulnerabilità sociale	420
Michela Bandini, Silvia Piccini, Andrea Bellandi, Emiliano Giovannetti	
Verso la creazione di una Treebank per il sardo	426
Nicoletta Puddu, Manuela Sanguinetti, Luigi Talamo	
A Learning-by-Doing Approach to Spoken Data Collection: The Case of the LPSP Corpus	432
Claudia Roberta Combei, Gaia Eleonora Di Raimondo, Sofia Maestri, Benedetta Romanazzi, Elena Scotti	
Digital Humanities e Letteratura Elettronica in Italia: marginalità e convergenze	440
Giuseppe Arena	

[6/B] CO-CREAZIONE, RICERCA PARTECIPATIVA, COMUNITÀ

Metodi e strumenti di sperimentazione tra co-creazione, Digital Humanities, e ricerca demo-etno-antropologica: la digitalizzazione come pratica relazionale ed etica.	448
Eleonora De Longis, Matteo Cova, Giovanni Bertelli	
Integrare strumenti digitali e conoscenze della comunità per una progettazione urbana resiliente al clima: l'iniziativa Dundrum by Design	455
Chiara Cocco, Mattia Leone, Antonio Savino, Miruna Popa, Sara Tedesco	
HCI e AI per il Public Engagement: casi di studio per la valorizzazione del patrimonio culturale	462
Samuel Aldo Iacolina, Manuela Angioni, Valentina Marotto, Francesca Mura, Piergiorgio Palla	

The Automation of Research in Digital Humanities: Artificial Researchers, Digital Archives, and the Democratization of Scholarly Access	468
Roberto Di Quirico	

[6/C] AI E AUTOMAZIONE PER ARCHIVI E PATRIMONIO

Dal problema storico alle IA: Il Database on the Slave Trade between the Mediterranean and the Atlantic (15 th -16 th Centuries)	477
Salvatore Spina, Carlo Taviani, Jörg Hörnschemeyer	
Peirce Interprets Peirce: Digitization, Automation, and Interpretation in Charles Peirce's Manuscripts	485
Alessandro Adamou, Sebastian Feil, Davide Picca, Carlo Teo Pedretti, Dario Rodighiero,	
Lorenzo Zangari	
A New Semantic Model for Mythologiae: from Data to Interpretation	494
Bianca La Manna	

[7/A] ONTOLOGIE, FAIR DATA E MODELLAZIONE DELLA CONOSCENZA

The ATLAS Guidelines for producing FAIR research products in the Digital Humanities	502
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	
RATIO: strumenti per la creazione e la gestione di risorse e collezioni FAIR nell'ambito delle scienze umane	509
Federico Silverstri, Lorenzo Mancini, Laura Baggiani, Marco Bagiacchi	
Making Uncertainty Explicit: The Preservation of Interpretive Complexity across Digital Humanities Project Workflows	516
Tommaso Battisti, Valentina Pasqual, Giulia Renda, Marilena Daquino	
Iconclass as a Semantic Framework for the Analysis of the Sacred Space	524
Polina Voronova	

[7/B] GIS E SPAZIALIZZAZIONI DIGITALI

Gli <i>Historical GIS</i> e la storia forestale: potenzialità e applicazioni	531
Vincenzo Colaprice	
Le nazioni genovesi nell'Atlante digitale di Storia Marittima del Regno di Sardegna (ASMSA). Lavori in corso	539
Giampaolo Salice, Giommara Carboni	
Integrating Romanian Toponymy in a GIS Background: dictionary forest names vs. OpenStreetMap forest data	547
Roxana Patras, Ana Odochiciuc, Mihai-Bogdan Atanasiu, Constantin Răchită	
Digital Humanities per narrare i territori: un WebGIS sui porti minori della Sardegna dal medioevo alla contemporaneità. Una proposta di implementazione collaborativa	554
Luigi Serra, Sebastiana Nocco	

[7/C] MODELLI, METODI E PROSPETTIVE DELL'EDIZIONE DIGITALE

Roman and Justinianic Legal Terminology in Thirteenth-Century Western European Diplomatic Sources	562
---	-----

Tamás Kovács, Angelos Nicolaou, Johannes Laroche, Georg Vogeler	
In Search of Lost Time: an Innovative Model for Representing Autograph Texts	567
Giulia Baldelli, Daniele Fusi, Matteo Zupancic, Franz Fischer, Claus Zittel	
<i>Naples Dante Project</i> . Un ambiente integrato per l'analisi bibliografica, filologica e iconografica	573
Gennaro Ferrante, Andrea La Veglia, Daniele Fusi, Stefano Angelo Rizzo, Angelo Eugenio Mecca, Daniele Duranti, Anna Sviridova	
La curatela biblioteconomica delle raccolte digitali MAB: linee gestionali, requisiti, workflow	580
Nicola Barbuti	

[8/A] GRAFI, LINKED DATA E MAPPING

DOG: an open-source ecosystem for mapping, narrating, and exploring humanities data	588
Alessandro Laruffa, Alessandro Capra	
Beyond Spatial Data: Exploring Gender and Property in the Catasto Generale Toscano through the Florentia Illustrata Semantic Platform	596
Erica Andreose , Remo Grillo , Gianmarco Spinaci	
Per la storia del mercato dell'arte tra otto e novecento: raccolta, modellazione e valorizzazione dei dati sugli antiquari italiani	602
Valentina Rossetti, Marilena Daquino, Francesca Mambelli, Ludovica Pannitto, Francesca Tomasi	
Un grafo per la ricerca storico-filologica in archivio: il caso del fondo Giuseppe Albini	609
Lucia Giagnolini, Mohamed Iheb Ouerghi, Valentina Pasqual, Cecilia Tamagnini, Francesca Tomasi	

[8/B] GEOGRAFIE LETTERARIE E MULTIMEDIALI

'Spazializzare' la narrativa ecocritica italiana: la piattaforma dell'Atlante digitale della letteratura ecologica italiana	617
Valentina Pasqual	
MeMo: mappe digitali, memoria letteraria e public engagement nel Mezzogiorno	624
Laura Giurdanella; Giuseppe Palazzolo	
Una proposta di approccio computazionale alle geografie del cinema italiano	631
Alberto Savi, Luca Giovannini	
Il campo semantico del fuoco nella Commedia: un caso di studio sulla pertinenza lessicale e sulla visualizzazione	638
Ruoci Song	

[8/C] ANNOTAZIONI E MARKUP

Ritagli di teatro. Annotazione semantica della rassegna stampa del Teatro Sant'Erasmus	644
Paolo Bonora	
Annotazione semantica delle formule poetiche del Medioevo germanico	651
Roberto Rosselli Del Turco, Letizia Vezzosi	
Ça y est : Annotating Arabic Texts for Teaching	658

Dal markup all'interpretazione: interrogare il linguaggio della critica verista in un corpus XML/TEI	665
Denise Bruno, Salvatore Cristofaro, Antonio Di Silvestro, Laura Mazzagufò, Daria Spampinato, Giuseppe Zappalà	

[9] SESSIONE POSTER

Archivi letterari nativi digitali: modelli descrittivi ed edizioni sperimentali	673
Elena Barchielli, Simon Willemin, Elena Spadini.	
Lexicon Philosophicum: una rivista digitale per la condivisione della conoscenza umanistica	679
Pamela Barletta, Maria Cristina Dalfino, Pietro Restaneo	
Intelligenza artificiale, reference digitale e accessibilità: l'esperienza di Alphabetic	684
Flavia Bruni – Elisabetta Castro	
Grafoteca e la costruzione digitale della memoria manoscritta: curatela, metadattazione e reti di risorse	691
Amalia Carrano	
Valorizzazione digitale del patrimonio culturale: Public-Private Partnership e partecipazione sociale	697
Annalisa Ciliberti	
I Vangeli digitali e accessibili, per un'editoria religiosa universale e senza barriere	703
Giulia D'Arcangelo	
Implementing a controlled vocabulary for medieval administrative acts: theoretical and methodological considerations	709
Chiara De Bastiani	
L'Archivio del Centro Conservazione e Restauro "La Venaria Reale": un caso studio per la diffusione dei dati tecnico-scientifici	717
Stefania De Blasi, Lorena Palmieri, Chiara Pipino	
Strumenti digitali per i percorsi di visita della Palazzina di Caccia di Stupinigi: Residenza Sabauda e Museo dell'Ammobiliamento	722
Stefania De Blasi	
Strategie di protezione, governance e sostenibilità degli archivi orali	727
Rosaria De Luca, Elvira Mercatanti, Monica Monachini	
FAIR Memories: A Workflow for the Preservation and Collaborative Reuse of Analog Oral History Collections	734
Nike del Quercio, Costanza Paolillo, Laurent Fintoni	
Ontologycore: A Formal Knowledge Graph of Internet Aesthetics	740
Anouk Flinkert, Ekaterina Krasnova, Shiho Nakamura	
Il ruolo del Test Plan per la validazione: il caso d'uso del Marketplace H2IOSC	747
Vittoria Fabiani	
Partecipazione, mediazione e conoscenza: Àndalas de Cultura quale ecosistema digitale del patrimonio culturale della Sardegna	754
Carolina Floris	
From Violin Lessons to Linked Open Data. Reconstructing Tartini's <i>Scuola delle Nazioni</i> through Digital Prosopography and Digital Editions	762
Selina Galka, Marcella Tambuscio, Rolf Wissmann, Cristina Scuderi, Georg Vogeler	

Addestramento di un modello di HTR in Transkribus per testi plurilingui di età moderna: il caso del <i>Ripulimento della lingua sarda</i> (Madau)	768
Michela Incollu, Giulia Murgia	
Modelli digitali interattivi per gli archivi teatrali storici: mappare le relazioni semantiche al Teatro Metastasio di Prato (1827–1860)	775
Matilde Innocenti	
Tra fotografia e segno: la trascrizione informatizzata dei registri storici Alinari	779
Pamela Krzemien	
Biblioteche filosofiche private e memoria culturale: http://picus.unica.it/	786
Andrea Lamberti, Giovanna Granata	
Beyond Archives: A Three-Stage Workflow for Relational Digital Libraries of Written and Oral Memory Sources	791
Giulia Lembo	
A Workflow-Centred Approach to Managing Semantic Artifacts as Linked Open Data	800
Michele Mallia, Fahad Khan, Valeria Quochi	
SEBASTIAN: uno strumento per l'organizzazione, l'analisi spaziale e la visualizzazione dei dati storici	807
Christian Marcantonio, Giuseppe Consolo	
Raccontando e valorizzando il patrimonio culturale. Dalle prime esperienze ai progetti futuri del CNR-ISEM	814
Maria Grazia Mele R., Samuel Aldo Iacolina, Luigi Serra, Giovanni Serreli	
La Digital Library e l'Ecosistema digitale per la cultura come opportunità di <i>public engagement</i>	821
Federico Meschini, Biancamaria Hermanin, Antonella Negri, Giuliano Romalli	
Strategie di engagement nella creazione e sostenibilità di un parco letterario: i risultati del progetto CHANGES	827
Sara Obbiso, Nicola Mariniello, Davide Bagnaresi, Riccardo Stracuzzi, Giuliana Benvenuti, Alessandro Iannucci	
Lo studio integrato del patrimonio materiale e immateriale per la ricostruzione delle pratiche storiche di gestione delle risorse ambientali nell'Alpujarra (Sierra Nevada): il progetto Sinergy	835
Alessandro Panetta	
<i>Reveduti, Correcti, Aprobati et Confirmati</i> : The Digital Edition of the Statutes of Ascoli	841
Michela Parma, Fabio Mariani	
Trascrivere automaticamente i registri catastali sardi del primo Novecento con Transkribus: risultati di un caso di studio	849
Andrea Pergola, Cecilia Tasca	
Ricostruzione virtuale di un patrimonio disperso: le collezioni di disegni a Genova nel XIX secolo	856
Giulia Pilosu	
Dai workflow descrittivi ai workflow applicativi tramite AEON e l'IA	860
Francesco Pinna, Emiliano Degl'Innocenti	
<i>Gamification</i> per la valorizzazione dei beni musicali. Un'ipotesi di raccolta di dati catalografici in <i>crowdsourcing</i>	867

Marcello Ranieri	
Tra riconoscimento disciplinare e prospettiva pubblica: un modello logico-applicativo per le conoscenze e le competenze nelle DH	874
Lisa Reggiani	
Mappe, scale e interpretazione. Per una cartografia digitale semantica dello spazio narrativo	883
Lorenzo Sabatino	
Verso un grande corpus diacronico dell'italiano: il contributo degli esempi del GDLI	893
Eva Sassolini, Sebastiana Cucurullo, Marco Biffi e Simonetta Montemagni	
Verso una classificazione funzionale dei prompt basata sulle intenzioni cognitive esplicitate dagli studenti nell'interazione con l'IA	901
Daniele Scala, Salvatore Varriale	
Il videogioco come strumento di mediazione interculturale	909
Alessia Stocco	
Edizione digitale commentata di <i>Le libere donne di Magliano</i> di Mario Tobino: primi passi, metodi e prospettive	917
Fabio Zarroli	
Il restauro testuale dei papiri nel progetto GreekSchools: CoPhiEditor e l'integrazione di modelli linguistici	923

[6/A] Modellazione testuale e linguistica computazionale



A Learning-by-Doing Approach to Spoken Data Collection: The Case of the LPSP Corpus

CLAUDIA ROBERTA COMBEI¹, GAIA ELEONORA DI RAIMONDO², SOFIA MAESTRI³, BENEDETTA ROMANAZZI⁴,
ELENA SCOTTI⁵

¹UNIVERSITÀ DEGLI STUDI DI ROMA "TOR VERGATA", ITALY - CLAUDIA.ROBERTA.COMBEI@UNIROMA2.IT

²UNIVERSITÀ DI PAVIA, ITALY - GAIAELEONORA.DIRAIMONDO01@UNIVERSITADIPAVIA.IT

³UNIVERSITÀ DI PAVIA, ITALY - SOFIA.MAESTRI01@UNIVERSITADIPAVIA.IT

⁴UNIVERSITÀ DI PAVIA, ITALY - BENEDETTA.ROMANAZZI01@UNIVERSITADIPAVIA.IT

⁵UNIVERSITÀ DI PAVIA, ITALY - ELENA.SCOTTI01@UNIVERSITADIPAVIA.IT

ABSTRACT (ENGLISH)

The paper presents an experiential learning model implemented in a 36-hour, 7-day intensive MA course in Linguistics at Collegio Ghislieri in Pavia. The course, named "Laboratory Phonetics and Speech Processing", produced the LPSP corpus, comprising two subcorpora: LPSP-IT (L1 Italian) and LPSP-ENG (L2 English). Under instructor supervision, students completed the full speech corpus construction cycle: task design, drafting and administering of informed consent and questionnaire, booth recording, speech segmentation and cleaning, data curation, documentation, licensing, and archiving. Ethics was embedded at all stages (consent in plain language, attention to sociodemographic data included, anonymization, controlled sharing). Recordings were collected in January 2026 in a phonetics laboratory, using a Blue Yeti Pro microphone at a sampling frequency of 44.1 kHz. Each of the 18 speakers completed a reading task (texts created for phoneme coverage; one speaker at a time; mono; cardioid mode; two attempts) and a find-the-difference spontaneous dialogue (paired speakers; stereo; bidirectional mode). Cleaning and segmentation resulted in 90 .wav files (total: 2 hours and 47 minutes), 45 for LPSP ENG and 45 for LPSP IT. Speakers were aged 22-27 years ($M = 24.44$, $SD = 1.50$). All were native speakers of Italian and comprised 61.11% women, 33.33% men, and 5.55% not disclosing their gender. Most participants resided in Northern Italy and reported a high level of education. This experience illustrates how attention to the technical dimensions of spoken-data collection, combined with reproducibility and ethical reflection, can transform students into competent data curators for their theses and early-stage linguistic research. The LPSP Corpus (audio recordings, stimuli, and sociodemographic information) is deposited on OSF under a CC BY-NC-SA 4.0 license and is available upon request, enabling reuse and replication of both the speech resource and the associated teaching model.

Keywords: spoken data; corpus collection; speech corpus; experiential learning

ABSTRACT (ITALIANO)

Un approccio di apprendimento esperienziale alla raccolta dei dati di parlato: il caso del corpus LPSP.

Il lavoro presenta un modello di apprendimento esperienziale implementato in un corso magistrale di linguistica (7 giorni, intensivo, 36 ore) tenuto al Collegio Ghislieri di Pavia. Nel corso, intitolato "Laboratory Phonetics and Speech Processing", è stato creato il corpus di parlato LPSP, composto da due sottocorpora: LPSP-IT (italiano L1) e LPSP-ENG (inglese L2). Sotto la supervisione della docente, i/le partecipanti hanno completato l'intero ciclo di costruzione del corpus: progettazione e somministrazione di consensi informati e questionari, registrazione in cabina, pulizia e segmentazione, documentazione, licenza, gestione e archiviazione dati. Gli aspetti etici sono stati integrati in tutte le fasi (consenso informato chiaro, attenzione ai dati sociodemografici, anonimizzazione, condivisione controllata). Le registrazioni sono state fatte a gennaio 2026 in un laboratorio di fonetica, con un microfono Blue Yeti Pro (frequenza di campionamento: 44,1 kHz). I 18 parlanti hanno svolto un compito di lettura (testi che rispettano esigenze fonologiche; mono; modalità cardioide; due tentativi) e un dialogo spontaneo di tipo trova le differenze (stereo; modalità bidirezionale). Le operazioni di pulizia e segmentazione hanno prodotto 90 file .wav (2 ore e 47 minuti), 45 per LPSP-IT e 45 per LPSP-ENG. I 18 parlanti madrelingua italiano avevano un'età compresa tra 22 e 27 anni ($M = 24,44$, $DS = 1,50$), 61,11% erano donne, 33,33% uomini e 5,55% di genere non dichiarato. La maggioranza risiedeva nel Nord Italia e aveva un livello alto di istruzione. Questa esperienza didattica mostra come l'attenzione all'etica, alla riproducibilità e agli aspetti tecnici e teorici della raccolta dei corpora orali possa trasformare gli/le studenti/esse in curatori/trici competenti di dati linguistici per tesi e altre ricerche. Il corpus LPSP è depositato su OSF con

licenza CC BY-NC-SA 4.0 ed è disponibile su richiesta per consentire il riuso dei dati e la riproducibilità del modello didattico.

Parole chiave: dati di parlato; raccolta di corpora; corpus di parlato; apprendimento esperienziale

1. INTRODUCTION¹

This paper describes a teaching and research initiative carried out within the 36-hour Master's degree in Linguistics course "Laboratory Phonetics and Speech Processing" at Collegio Ghislieri in Pavia. The course was designed to bridge theory and practice in experimental phonetics, corpus linguistics, and speech technology, by guiding students through the complete pipeline of speech science research, from task design and ethical approvals to recording, annotation, transcription, cleaning, documentation, analysis, and research dissemination (including the draft of this paper). As a course capstone outcome, the students collaboratively created the LPSP corpus (the name derives from the initials of "Laboratory Phonetics and Speech Processing"), a spoken resource comprising two subcorpora: LPSP-IT for L1 Italian and LPSP-ENG for L2 English, both containing data uttered by native speakers of Italian. The corpus can be reused for future research (including their own dissertations) and teaching. The project provided a structured experiential learning environment in which students assumed rotating, well-defined roles (e.g., task design, fieldwork, audio engineering, data management), thereby acquiring both methodological rigor and good data stewardship.

The rationale for this course is twofold. From a teaching and learning perspective, the instructor began from the premise (supported by their own experience) that linguistics students often work with pre-existing corpora and other linguistic resources without ever engaging directly with the processes of data elicitation, processing, documentation, and curation (Ma et al., 2023). This lack of exposure can mask issues of data provenance, representativeness, ethical considerations, measurement validity, and reproducibility, which typically surface only later during their Master's degree theses, at a stage when meaningful correction is often no longer feasible. By assuming responsibility for each stage of corpus creation, students confronted the practical challenges inherent in real-world spoken data collection: designing elicitation tasks that simultaneously support segmental and suprasegmental analyses, standardizing recording conditions, and ensuring that file naming, version control, and documentation aligned with typical linguistic research requirements (Wieczorkowska, 2025).

From a research-oriented perspective, the project provided experience with professional tools and techniques (e.g., Praat, forced alignment pipelines) and speech processing workflows (e.g., segmentation, feature extraction, signal inspection), while foregrounding transparent documentation practices to facilitate accurate interpretation by other potential corpus users. The combination of these two dimensions allowed students to gain both practical and technical insight into the methods that underpin research in speech science (Mauranen, 2004).

The course, and subsequently the corpus construction, integrated ethical aspects across all stages. Rather than treating ethics as a one-off compliance exercise, students engaged with it iteratively by drafting and refining informed consent forms in plain language, designing sociodemographic questionnaires that are informative yet proportionate (neither overly long nor unduly intrusive), implementing anonymization for each subcorpus, and discussing potential risks related to identifiability, data sharing, and re-use. Ethical choices informed methodological ones, for example, deciding which sociodemographic variables to collect (and which to omit), how to justify them in relation to possible research questions, and how to store, license, and distribute data responsibly (Cheng et al., 2024). This ethical reflection at all steps aimed to improve the students' understanding of privacy, diversity, potential bias, and the scope of permissible reuse, all aspects that they can carry into their own future research (e.g., their Master's dissertations).

Therefore, LPSP corpus aimed to function as both an instructional scaffold and a research contribution. The purpose of the activity was to help students become competent linguistic data curators (not only users) while producing a reusable resource for the research community useful for investigating L1 Italian and L2 English.

¹ This paper is the result of a collaborative effort by all the authors, who share responsibility for its content. For the purposes of the Italian academic system, we specify the contributions to the first draft, as follows: Sections 1 and 5 were written by C.R. Combei, Section 2 by B. Romanazzi and C.R. Combei, Section 3 by G.E. Di Raimondo and E. Scotti, and Section 4 by S. Maestri. The data visualization, revision, editing, and preparation of the final published version of the paper were carried out by C.R. Combei.

The remainder of the paper details the corpus design and data collection protocol (Section 2), summarizes corpus composition and participant profiles (Section 3), and describes data availability, licensing, and intended uses (Section 4), with the aim of allowing replication of both the corpus and the teaching model. Concluding remarks and future steps are briefly outlined at the end of the paper (Section 5).

2. CORPUS DESIGN AND DATA COLLECTION PROTOCOL

As part of the coursework, students designed and compiled a spoken corpus, namely the LPSP corpus. This resource comprises two subcorpora: an L1-Italian subcorpus (LPSP-IT) and an L2-English subcorpus (LPSP-ENG). The students were divided into two groups, each dedicated to one subcorpus. Within each group, students were assigned specific responsibilities, including elicitation task design, drafting informed consent materials, developing and administering the sociolinguistic questionnaire, managing sociolinguistic data, operating the recording workflow, and cleaning and segmenting the audio. For the LPSP corpus 18 native speakers of Italian were recruited (see Section 3 for sociodemographic details). All 18 participants received two informed consent forms (one per subcorpus), detailing the aims of the project, data collection and storage procedures, participant rights, and data sharing plans (speakers were informed that the corpus would have been shared through the Open Science Framework platform). By signing the informed consent, speakers consented to being recorded and to anonymized data sharing with the research community. Speakers also completed one sociolinguistic questionnaire for each subcorpus to provide background information relevant for analysis and interpretation of the audio material. An anonymization protocol was implemented for both subcorpora: each speaker was assigned two distinct alphanumeric identifiers (one for L1 Italian, one for L2 English), and these identifiers were used to label recordings and link them to sociolinguistic variables. Names were not associated with the audio or the sociolinguistic datasets. Access to the re-identification key (the consent forms and named questionnaires) was restricted to one designated student and to the course instructor (the first author of this paper).

Elicitation took place in January 2026 in the recording booth of the experimental phonetics laboratory at the University of Pavia. Recordings were made with a Blue Yeti Pro microphone using Praat (Boersma, 2001) at a 44.1 kHz sampling frequency. A designated student oversaw the recording process, ensuring consistent file naming with the assigned alphanumeric codes and managing technical parameters (e.g., mono vs. stereo channel; cardioid mode vs. bidirectional polar mode). Data collection occurred over two sessions: the first afternoon was dedicated to the L1-Italian subcorpus and the second to the L2-English subcorpus. Each participant completed four recording sessions in total (two per subcorpus), corresponding to two task types: a reading task (with two attempts) and a find-the-difference task. The choice of the two tasks was dictated by previous research that demonstrated how spontaneous speech differs acoustically from read speech, especially through reduced spectral space, which significantly lowers phoneme and speech recognition accuracy (Nakamura et al., 2008).

For the reading task, speakers read aloud texts specifically constructed to cover the full phoneme inventory of the target language (Italian and British English). Each text was read twice to increase control and reliability, enabling future analyses of the realization of Italian sounds by native Italian speakers and the realization of British English sounds by Italian learners of English. The Italian text had a type-token ratio of 0.77, consisted of 8 sentences, and the mean sentence length was 15.9 words. The lexical items spanned a wide range of frequency per million, which was checked in the CORIS corpus (Rossini Favretti et al., 2002), yielding a balanced complexity profile. The English text had a type-token ratio of 0.65, contained 25 sentences and had a mean sentence length of 8.4 words. Similarly to the Italian text, lexical items for the English text were selected to cover a broad frequency range, which was checked in the British National Corpus (Burnard & Aston, 1998).

The paired task elicited semi-spontaneous dialogue using two closely related images that contained systematic similarities and differences. The English materials were designed around familiar lexical items to ensure accessibility for speakers with lower English proficiency. Paired speakers were given different versions of the image and asked to describe what they saw and identify differences, thereby prompting spontaneous conversation anchored in the visual stimuli.

Both texts and images were created by the students in charge of stimulus design. All stimuli and accompanying materials are available via the Open Science Framework (OSF) repository for the LPSP Corpus: <https://osf.io/3vu7e/>.

3. CORPUS DESCRIPTION AND STATISTICS

The complete corpus comprises 90 audio files, for a total of 2 hours 47 minutes and 48 seconds. In particular, 45 audio files are part of the LPSP-ENG subcorpus, which contains English-language audio material (for a total of 103 minutes and 18 seconds), while the other 45 audio files regard the LPSP-IT subcorpus with its Italian-language material (for a total of 64 minutes and 30 seconds).

The original recordings of the find-the-differences task and the ones of the reading task of each subcorpus were cleaned by background noises and silence using the freeware software Praat (Boersma, 2001). Specifically, the files of the reading task of both subcorpora were segmented in two different files, as participants made two attempts while recording. Therefore, the initial number of 18 files turned into 36 for the reading task and 9 for the find-the-difference task, for a total of 45 recordings in each subcorpus.

The 18 participants that took part in the corpus recording were all native Italian speakers (L1) aged from 22 to 27 ($M = 24.44$, $S.D. = 1.50$). The gender distribution in the sample included 61.11% ($n = 11$) of female speakers, 33.33% ($n = 6$) of male speakers, and 5.55% ($n = 1$) who did not prefer to disclose their gender.

Then, 72.22% ($n = 13$) of the speakers came from Northern Italy, specifically 38.88% ($n = 7$) from Lombardy, 11.11% ($n = 2$) from Liguria, 11.11% ($n = 2$) from Emilia-Romagna, and 11.11% ($n = 2$) from Piedmont. While the remaining participants, which is 27.77% ($n = 5$), came from Southern and Central Italy, namely 5.55% ($n = 1$) from Sicily, 5.55% ($n = 1$) from Campania, 5.55% ($n = 1$) from Calabria, 5.55% ($n = 1$) from Abruzzo, and 5.55% ($n = 1$) from Umbria. When they were recorded, the 83.33% ($n = 15$) of speakers resided in Northern Italy, so only 16.66% ($n = 3$) indicated they resided in the South, in the same region they were born.

Also, at the time of data collection, 55.55% ($n = 10$), which is more than half of the participants, reported holding a Bachelor's degree as their highest completed level of education, while being enrolled in a Master's degree program in Linguistics. Then, 33.33% ($n = 6$) were PhD students of Linguistics and reported a Master's degree as their highest completed qualification. A high school diploma was reported by 11.11% ($n = 2$) of the speakers who were enrolled in a Bachelor's degree program (respectively, one student in Classics and one in Communication and Media Studies) at the time of the recording. The high level of education of the sample reflects the academic recruitment context of the study.

Because the recording protocol included a reading task, information on visual impairments and learning disorders was collected (Ehri, 1997; Miller & Yochum, 1991). According to self-reports, 72.22% of the participants ($n = 13$) used corrective devices (glasses or contact lenses) for refractive errors such as myopia or astigmatism, whereas 27.78% ($n = 5$) reported no visual impairments. Furthermore, none of the participants (100%, $n = 18$) reported having any learning disorders.

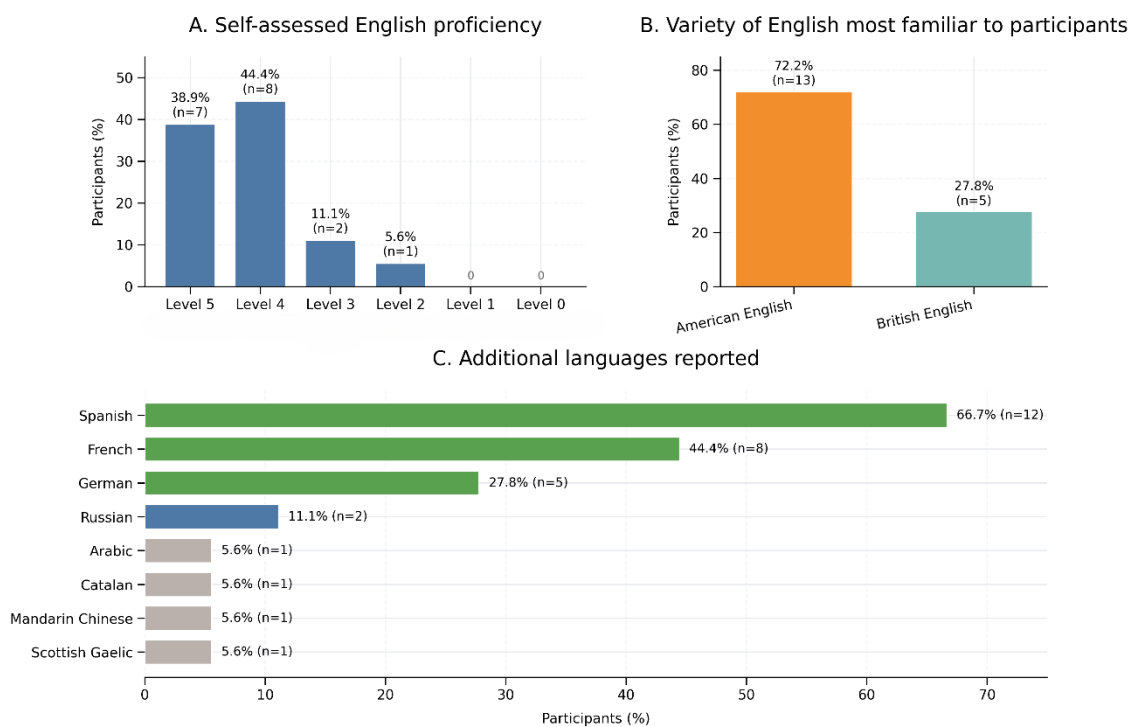


Figure 1. The speakers' L2 linguistic repertoire

The speakers' linguistic repertoire was investigated through a sociolinguistic questionnaire, in which they reported the languages they knew, besides L1 Italian, and self-assessed their proficiency on a 6-point Likert scale (i.e., 0-5). As shown in Figure 1, all participants reported knowledge of English, indeed 38.89% of them ($n = 7$) reported excellent proficiency (level 5), 44.44% ($n = 8$) advanced proficiency (level 4), and 11.11% ($n = 2$) intermediate proficiency (level 3). A further 5.55% ($n = 1$) indicated low proficiency (level 2). No participants reported very low proficiency (level 1) or no knowledge of English (level 0). In addition, American English was the variety that 72.22% ($n = 13$) of the participants were most familiar with, compared to British English of which only 27.77% ($n = 5$) had knowledge of. Moreover, 66.66% ($n = 12$) of the speakers reported knowledge of Spanish, with varying levels of proficiency. French was reported by 44.44% ($n = 8$) of the participants and German by 27.77% ($n = 5$) of them, while only 11.11% ($n = 2$) reported competence in Russian. The linguistic repertoire also included single instances (5.55%) of Arabic, Mandarin Chinese, Catalan, and Scottish Gaelic. The sociolinguistic questionnaire also investigated language use across different social domains. In interactions with friends, Italian was reported as the sole language by 66.66% of the participants ($n = 12$), while 16.66% ($n = 3$) reported using Italian in combination with English, 5.5% ($n = 1$) Italian and dialect and 11.11% ($n = 2$) Italian, English, and dialect. Furthermore, Italian was reported as the only language used at home by 77.77% ($n = 14$) of speakers, while 16.66% ($n = 3$) use Italian in combination with dialect. Only 5.5%, which corresponds to one case, uses Italian with Mandarin Chinese and Wu. Regarding interactions with siblings, 77.77% ($n = 14$) of the participants reported using Italian exclusively, while 11.11% ($n = 2$) reported using Italian and a dialect; the remaining 11.11% ($n = 2$) did not respond, probably due to the absence of siblings. Among the 18 participants, 88.88% ($n = 16$) reported Italian as the L1 of both parents, while 11.11% ($n = 2$) reported non-Italian parental L1s (Wu in one case, and Italian and Amharic in another). Finally, in interactions with romantic partners, 66.66% ($n = 12$) of participants reported using Italian exclusively, while 33.33% ($n = 6$) did not respond to this question, presumably corresponding to participants without a current partner or who did not want to disclose this information.

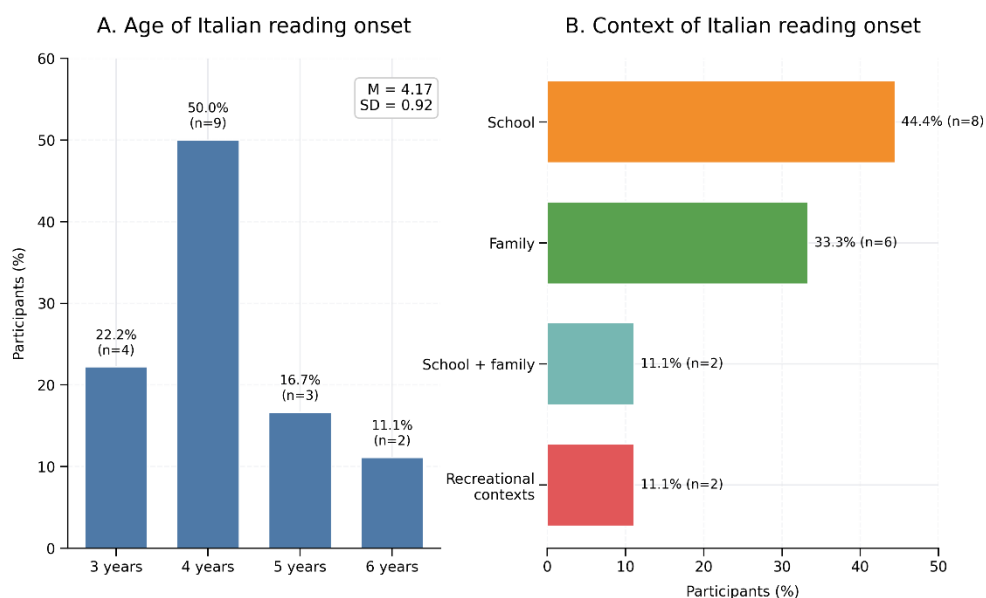


Figure 2. The speakers' L1 Italian reading onset and context

Speakers had to report in the sociolinguistic forms the age and the contexts of Italian reading onset for the LPSP-IT subcorpus (Modiano, 1968; Kamhi-Stein, 2003), whereas for the LPSP - ENG subcorpus the age and contexts of English learning onset and the frequency of spoken English usage (Marian et al., 2007; Unsworth, 2013). Figure 2 presents the data from the sociolinguistic form of the LPSP-IT subcorpus, showing that 50% ($n = 9$) of participants claimed they had started reading Italian at the age of 4, 22.22% ($n = 4$) at the age of 3, 16.66% ($n = 3$) at the age of 5 and only 11.11% ($n = 2$) at the age of 6 ($M = 4.16$, $SD = 0.92$). Regarding the contexts of reading onset, among the total number of participants, 44.44% ($n = 8$) started to read Italian at school, 33.33% ($n = 6$) with their family, 11.11% ($n = 2$) both at school and with their family and 11.11% ($n = 2$) reported recreational contexts as first onset.

As far as the LPSP-ENG is concerned, as shown in Figure 3, 22.22% ($n = 4$) started to learn English at the age of 5, 22.22% ($n = 4$) at the age of 6, 16.66% ($n = 3$) at the age of 4, 11.11% ($n = 2$) since birth (although they do not consider themselves native speakers of English), 11.11% ($n = 2$) at the age of 10, 5.55% ($n = 1$) at the age of 2, 5.55% ($n = 1$) at the age of 7 and 5.55% ($n = 1$) at the age of 15 ($M = 5.55$, $SD = 3.55$). On the other hand, among 18 participants, 83.33% ($n = 15$) had learned English at school, 11.11% ($n = 2$) in a family context and 5.55% ($n = 1$) in an informal context. In addition, 33.33% ($n = 6$) of the participants use English a few times in a month, 27.77% ($n = 5$) every day, 22.22% ($n = 4$) a few times a week, 11.11% ($n = 2$) once a week and the 5.55% ($n = 1$) less than once a month.

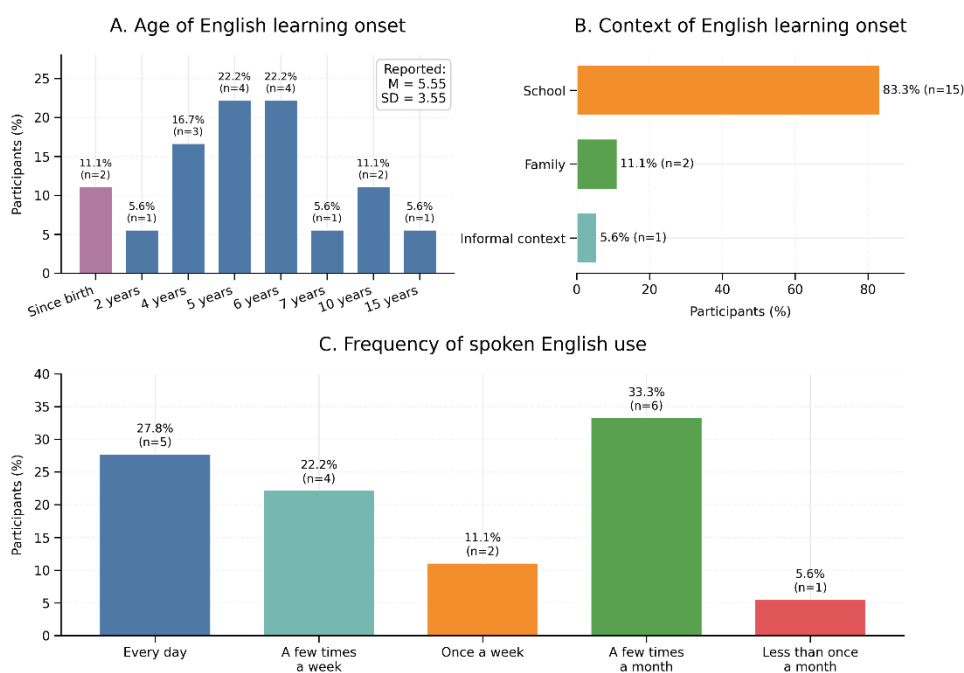


Figure 3. The speakers' L2 English learning onset and context

4. DATA AVAILABILITY AND USABILITY

All corpus anonymized data were uploaded to the OSF repository as a private project and are available for teaching and research purposes, upon reasonable request, by registered users of the platform. The repository was assigned the DOI [10.17605/OSF.IO/3VU7E](https://doi.org/10.17605/OSF.IO/3VU7E) and is available at the following OSF link: <https://osf.io/3vu7e/>. The project contains two main folders, LPSP-ENG and LPSP-IT, both of which include the audio files for the read speech task and for the spontaneous speech task (find-the-difference) in .wav format. The name of each audio name consists of an alphanumeric string which is the result of the following information:

- Alphanumeric ID code of the speaker.
- RS (read speech) or FD (find-the-difference), depending on the task performed by the recorded speaker.
- Only for RS, the number of the reading attempt (trial 1 or trial 2).

Folders also include all stimuli in .pdf format and sociodemographic data from the questionnaires in .xlsx format, which report information about the speakers.

All available data are released under the CC BY-NC-SA 4.0 International License and may be downloaded and used for non-commercial research and teaching purposes only. Potential applications include studies of linguistic variation and other sociophonetic aspects, teaching activities in speech processing (e.g., audio segmentation, acoustic analysis, feature extraction), L2 acquisition research, conversation analysis, and other related fields. Use of the data for commercial purposes is strictly prohibited. Any ethical violations will be reported to the relevant authorities and addressed in accordance with applicable regulations.

5. CONCLUSIONS AND FUTURE WORK

In this work we showed how a structured experiential learning approach can train MA students in Linguistics to manage the full speech corpora lifecycle, from the design of elicitation tasks to sharing the data and ethical considerations. The course deliverable, namely the LPSP corpus (with its L1 Italian and L2 English subcorpora), transformed the coursework into a reusable resource for the linguistic community, for teaching and research across phonetics, speech processing, and other related fields. Limitations include the modest sample size of the corpus (18 speakers), single-institution setting, constrained time frame, and a task design that prioritized control over ecological data collection. Future research will aim to broaden the diversity of participants in terms of sociodemographic characteristics, expand the range of spontaneous speech contexts, and incorporate full transcription of all audio materials

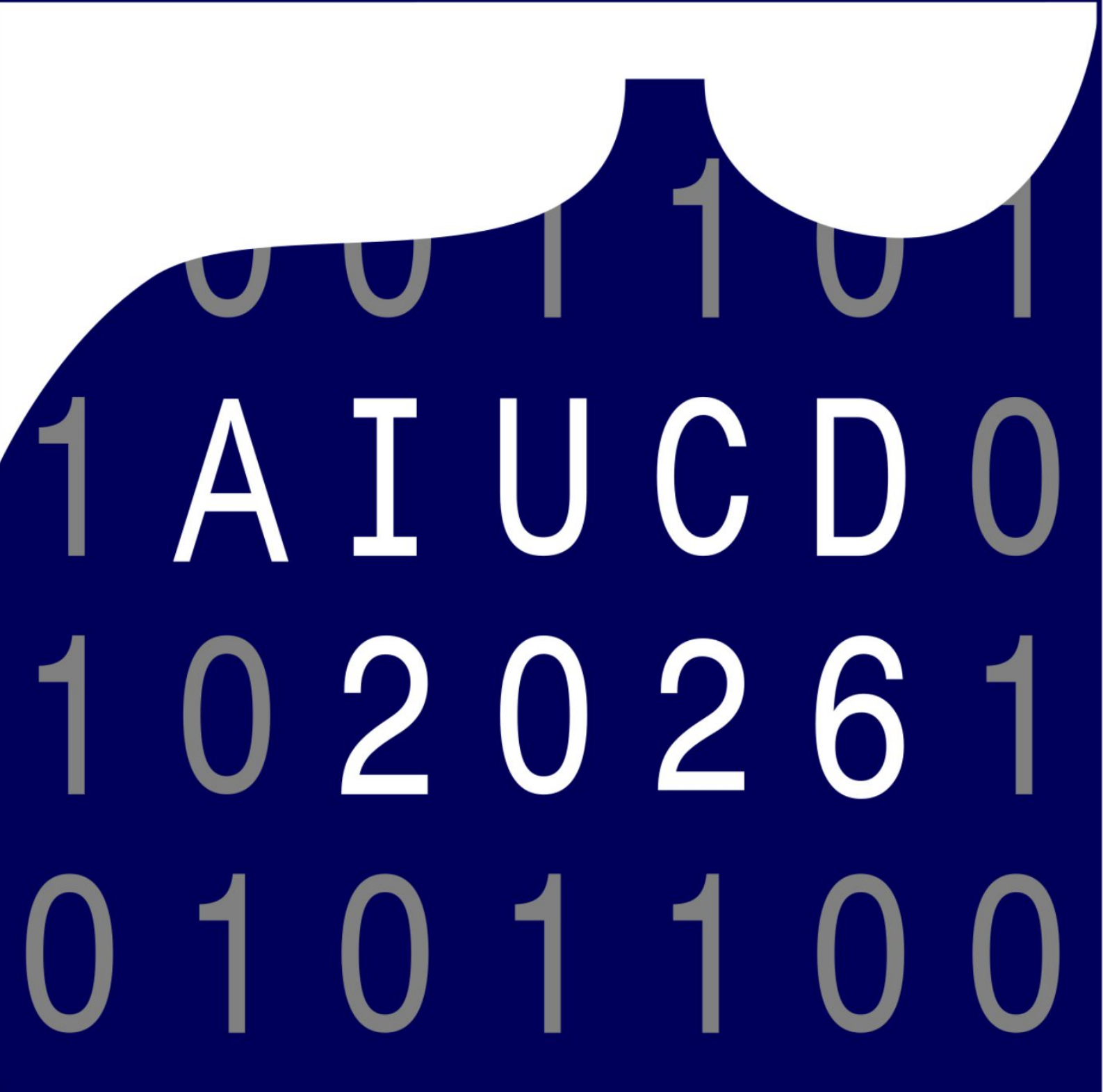
alongside multilayer phonetic and phonological annotation, work that is currently being carried out by MA students in Computational Linguistics at the University of Rome "Tor Vergata". All in all, when considered together, we believe that this activity encouraged students to move from being mere users of corpora toward a more reflective and competent engagement with linguistic data collection and curation.

ACKNOWLEDGEMENTS

The first author is thankful to Collegio Ghislieri for hosting the intensive course "Laboratory Phonetics and Speech Processing" and to Chiara Zanchi for promoting it. In addition to the authors of this paper, the following students, listed in alphabetical order, contributed to the collection and processing of the corpus: Matilde Bazzurro, Peter Bellarosa, Anna Giuliana Carlini, Emanuele Maria Ferrero, Francesco Logozzo, Elisa Mattiolo, Assunta Napoletano, Filippo Zana. The authors gratefully acknowledge their efforts, as well as the speakers who generously agreed to be recorded.

REFERENCES

- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International*, 5(9–10), 341–345.
- Burnard, L., & Aston, G. (1998). *The BNC handbook: Exploring the British National Corpus*. Edinburgh University Press.
- Cheng, L., Han, J. & Nasirov, J. (2024). Ethical considerations related to personal data collection and reuse: trust and transparency in language and speech technologies. *International Journal of Legal Discourse*, 9(2), 217-235. <https://doi.org/10.1515/ijld-2024-2010>.
- Ehri, L. C. (1997). Sight word learning in normal readers and dyslexics. In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 163–189). Routledge.
- Kamhi-Stein, L. D. (2003). Reading in two languages: How attitudes toward home language and beliefs about reading affect the behaviors of "underprepared" L2 college readers. *TESOL Quarterly*, 37(1), 35–71. <https://doi.org/10.2307/3588465>.
- Ma, Q., Chiu, M. M., Lin, S., & Mendoza, N. B. (2023). Teachers' perceived corpus literacy and their intention to integrate corpora into classroom teaching: A survey study. *ReCALL*, 35(1), 19–39. <https://doi.org/10.1017/S0958344022000180>.
- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4), 940–967. [https://doi.org/10.1044/1092-4388\(2007/067\)](https://doi.org/10.1044/1092-4388(2007/067)).
- Mauranen, A. (2004). Speech corpora in the classroom. In G. Aston, S. Bernardini & D. Stewart (Ed.), *Corpora and Language Learners* (pp. 195–211). John Benjamins Publishing Company. <https://doi.org/10.1075/scl.17.14mau>.
- Miller, S. D., & Yochum, N. (1991). Asking students about the nature of their reading difficulties. *Journal of Reading Behavior*, 23(4), 465–485. <https://doi.org/10.1080/10862969109547754>.
- Modiano, N. (1968). National or mother language in beginning reading: A comparative study. *Research in the Teaching of English*, 2(1), 32–43. <https://doi.org/10.58680/rte196820263>.
- Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2), 171–184. <https://doi.org/10.1016/j.csl.2007.07.003>.
- Rossini Favretti, R., Tamburini, F., & De Santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *A rainbow of corpora: Corpus linguistics and the languages of the world* (pp. 27–38). Lincom Europa.
- Unsworth, S. (2013). Assessing Age of Onset Effects in (Early) Child L2 Acquisition. *Language Acquisition*, 20(2), 74–92. <https://doi.org/10.1080/10489223.2013.766739>.
- Wieczorkowska, A. (2025). Methodology for Obtaining High-Quality Speech Corpora. *Applied Sciences*, 15(4), 1848. <https://doi.org/10.3390/app15041848>.



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI



Centro interdisciplinare per
l'Umanistica Digitale

ASSOCIAZIONE per
l'INFORMATICA UMANISTICA
e la CULTURA DIGITALE

