LREC-COLING 2024

# Third Workshop on Language Technologies for Historical and Ancient Languages @LREC-COLING-2024 (LT4HALA 2024)

Workshop Proceedings

Editors
Rachele Sprugnoli and Marco Passarotti

25 May, 2024
Torino, Italia

**Proceedings of LT4HALA 2024: The Third Workshop on Language Technologies for Historical and Ancient Languages @LREC-COLING-2024**



ISBN 978-2-493814-46-3
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

# Preface

These proceedings include the papers accepted for presentation at the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024).[1] The workshop was held on May 25th 2024 in Turin, Italy, co-located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).[2]

The workshop wants to provide a venue to discuss research works on a wide range of topics concerning the building, analysis, exploitation and distribution of collections of digitized texts written in historical and ancient languages as well as of their lexica, with a specific focus on the development and application of Language Technologies (LTs) for such purposes.

The topics of the workshop are strictly bound to the peculiar characteristics of textual and lexical data for historical and ancient languages, which set them apart from modern languages, with a significant impact on the use and development of LTs for their processing and study. Among the topics covered by the workshop are issues about the digitization process of textual sources, like handling spelling variation, and detecting and correcting OCR (Optical Character Recognition) errors. Issues related to the automatic processing of various layers of metalinguistic annotation are also included. Annotation is made complex by the sparsity and inconsistency of texts that present considerable orthographic variation, are sometimes incomplete and belong to a large spectrum of literary genres. Such issues raise problems of adaptation of Natural Language Processing (NLP) tools and pipelines to address diachronic/diatopic/diastratic variation in texts, which requires to be properly evaluated.

The various LTs tasks related to the topics of LT4HALA require a strict collaboration between scholars from different disciplinary areas. In such respect, the objective of the LT4HALA workshop series is to foster cross-fertilization between the Computational Linguistics community and the areas in the Humanities dealing with historical linguistic data, e.g. historians, philologists, linguists, archaeologists and literary scholars. Such a wide and diverse range of disciplines and scholars involved in the development and use of LTs for historical and ancient languages is mirrored by the large set of topics covered by the papers published in these proceedings, which, among others, include the creation and evaluation of annotated corpora and lexical resources for historical languages, and the use of Large Language Models (together with their fine-tuning) for performing various NLP tasks, like machine translation, summarization, sentiment analysis, dependency parsing, part-of-speech tagging, named entity recognition, and authorship attribution.

As large as the number of topics discussed in the papers is that of the either ancient/dead languages or the historical varieties of modern/living ones concerned. Overall, the languages tackled in the papers published in these proceedings are the following: Latin (as the most represented language), Old English, Old Irish, Old Italian, Dutch (in historical documents), Middle French, Ancient Greek, Hebrew, XIX century Italian and English, variations of the Ancient Egyptian languages (Old, Middle, and Late Egyptian, Demotic, Coptic), Gothic, Classical Armenian, Old High German.

In the call for papers, we invited to submit proposals of different types, such as experimental papers, reproduction papers, resource papers, position papers and survey papers. We asked both for long and short papers describing original and unpublished work. We defined as suitable long papers (up to 9 pages, plus references) those that describe substantial completed research and/or report on the development of new methodologies, as well as position papers. Short papers (up to 5 pages, plus references) were instead more appropriate for reporting on works in progress or for describing a specific tool or project. We encouraged the authors

[1] https://circse.github.io/LT4HALA/2024/

[2] https://lrec-coling-2024.org

of papers reporting experimental results to make their results reproducible and the entire process of analysis replicable, by distributing the data and the tools they used. Like for LREC-COLING 2024, the submission process was double-blind. Each paper was reviewed but three independent reviewers from a program committee made of 27 scholars (13 women and 14 men) from 13 countries. In total, we received 32 submissions (against the 24 of the previous edition). After the reviewing process, we accepted 20 submissions, leading to an acceptance rate of 62.50%.

LT4HALA 2024 was also the venue of the third edition of EvaLatin, the campaign devoted to the evaluation of NLP tools for Latin.[3] EvaLatin was started in 2020 (co-located with the first edition of LT4HALA) considering the important role played by textual data and linguistic metadata in the study of historical and ancient languages, with a special focus on Latin due to its prominence among such languages, both for the size and for the degree of diversity of its texts. Running evaluation campaigns in such a scenario is essential to understand the level of accuracy of the NLP tools used to build and analyze resources featuring texts that show those peculiar characteristics mentioned above. The third edition of EvaLatin focused on two shared tasks (i.e. Dependency Parsing, and Emotion Polarity Detection). The Dependency Parsing task was based on the Universal Dependencies (UD) framework.[4] No specific training data was released but participants were left free to make use of any (kind of) resource they consider useful for the task, including the Latin treebanks already available in the UD collection. In this regard, one of the challenges of this task was to understand which treebank (or combination of treebanks) is the most suitable to deal with new test data. Test data included both prose and poetic texts. Also for the Emotion Polarity Detection task, no training data were released but participants were provided with an annotation sample, a manually created polarity lexicon and annotation guidelines. Again, participants were left free to pursue the approach they prefer, including unsupervised and/or cross-language ones. Test data were poetic texts from different time periods. Shared data and all the necessary evaluation scripts were distributed to participants. Participants were required to submit a technical report for each task (with all the related sub-tasks) they took part in. The maximum length of the reports was 4 pages (plus references). In total, these proceedings include 5 technical reports of EvaLatin, corresponding to as many participants (3 for the Dependency Parsing Task, and 2 for the Emotion Polarity Detection task). All reports received a light review by the organizers who checked the correctness of the format, the exactness of the results and ranking reported, as well as the overall exposition. The proceedings also feature a paper detailing some specific aspects of the third edition of EvaLatin, like dataset, annotation criteria and results of the shared tasks.

Besides EvaLatin, LT4HALA 2024 hosted also the third edition of EvaHan, an evaluation campaign of NLP tools for the Ancient Chinese language, organized by a team of scholars directed by Bin Li (Nanjing Normal University)[5] The third edition of EvaHan focused on one task, namely a joint task of Sentence Segmentation and Punctuation. The EvaHan 2024 dataset was made of texts from classical sources, notably Siku Quanshu, along with other historical texts. The dataset's processing involved initial automatic punctuation and sentence segmentation. Subsequently, these automatic outputs were corrected and refined by experts in Ancient Chinese language to ensure the highest quality of gold standard texts. All evaluation data were txt files in Unicode (UTF-8) format. The training data comprised 10 million characters sourced from the Siku Quanshu. The test data included approximately 50,000 characters of Ancient Chinese texts. Participants were allowed to submit runs following two modalities. In the closed modality, each team was allowed to use only the training data provided, and the pre-trained model XunziALLM, which is a large language model for ancient Chinese processing. In the open modality, there was no limit on the resources, data and models: annotated external

[3] https://circse.github.io/LT4HALA/2024/EvaLatin

[4] https://universaldependencies.org

[5] https://circse.github.io/LT4HALA/2024/EvaHan

data, such as the components or Pinyin of the Chinese characters, or word embeddings could be employed. Like for EvaLatin, the participants of EvaHan were required to submit a short technical report which received a light review by the organizers. Overall, these proceedings include an overview of the EvaHan campaign (authored by the organizers) and 6 technical reports, corresponding to as many participants.

We are grateful to the organizers of EvaHan, who contributed to extend the range of historical and ancient languages of the LT4HALA 2024 workshop and showed how some NLP-related issues concern ancient and historical languages per se, despite their typological differences.

Now in its third edition, LT4HALA is constantly growing both as for the number of participants and as for the quantity and diversity of the languages and topics addressed by their scholarly contributions. We are glad to realize that the field is getting bigger, yet considering that this is not surprising, as the study of ancient and historical languages has always been strictly bound to the analysis of the empirical evidence provided by texts. Processing the collections of such texts, which today are largely available in digital format, by using the most advanced LTs to perform their computational analysis, promises to advance the state of the art in the century-long study of our linguistic past. LT4HALA wants to provide a venue to support such a computational turn.

Rachele Sprugnoli
Marco Passarotti

**Workshop Organizers:**

Rachele Sprugnoli, Università degli Studi di Parma (Italy)
Marco Passarotti, Università Cattolica del Sacro Cuore di Milano (Italy)

**Program Committee:**

Adam Anderson, FactGrid Cuneiform Project (USA)
Yannis Assael, Google DeepMind (UK)
Monica Berti, University of Leipzig (Germany)
Luca Brigada Villa, Università di Bergamo (Italy)
Flavio Massimiliano Cecchini, Katholieke Universiteit Leuven (Belgium)
Claudia Corbetta, Università degli Studi di Bergamo (Italy)
Margherita Fantoli, Katholieke Universiteit Leuven (Belgium)
Federica Gamba, Charles University (Czech Republic)
Shai Gordin, Ariel University (Israel)
Timo Korkiakangas, University of Helsinki (Finland)
Federica Iurescia, Università Cattolica del Sacro Cuore di Milano (Italy)
Bin Li, Nanjing Normal University (P.R. China)
Eleonora Litta, Università Cattolica del Sacro Cuore di Milano (Italy)
Yudong Liu, Western Washington University (USA)
Francesco Mambrini, Università Cattolica del Sacro Cuore di Milano (Italy)
Barbara McGillivray, Turing Institute (UK)
Chiara Palladino, Furman University (USA)
John Pavlopoulos, Athens University of Economics and Business (Greece)
Giulia Pedonese, Istituto di Linguistica Computazionale, CNR-ILC (Italy)
Matteo Pellegrini, Università Cattolica del Sacro Cuore di Milano (Italy)
Eva Pettersson, Uppsala University (Sweden)
Sophie Prévost, Laboratoire Lattice (France)
Thea Sommerschield, University of Nottingham (UK)
James Tauber, Eldarion (USA)
Alan Thomas, University of Sheffield (UK)
Toon Van Hal, Katholieke Universiteit Leuven (Belgium)
Tariq Yousef, University of Southern Denmark (Denmark)

**EvaLatin 2024 Organizers:**

Rachele Sprugnoli, Università degli Studi di Parma (Italy)
Federica Iurescia, Università Cattolica del Sacro Cuore di Milano (Italy)
Marco Passarotti, Università Cattolica del Sacro Cuore di Milano (Italy)

**EvaHan 2024 Organizers:**

Bin Li, Nanjing Normal University (P.R. China)
Bolin Chang, Nanjing Normal University (P.R. China)
Minxuan Feng, Nanjing Normal University (P.R. China)
Chao Xu, Nanjing Normal University (P.R. China)
Dongbo Wang, Nanjing Agricultural University (P.R. China)

# Table of Contents

# Workshop Program

**Saturday, May 25, 2024**

**+** ***Long and short workshop papers***

*Goidelex: A Lexical Resource for Old Irish*
Cormac Anderson, Sacha Beniamine and Theodorus Fransen

*Developing a Part-of-speech Tagger for Diplomatically Edited Old Irish Text*
Adrian Doyle and John P. McCrae

*From YCOE to UD: Rule-based Root Identification in Old English*
Luca Brigada Villa and Martina Giarda

*Too Young to NER: Improving Entity Recognition on Dutch Historical Documents*
Vera Provatorova, Marieke van Erp and Evangelos Kanoulas

*Towards Named-Entity and Coreference Annotation of the Hebrew Bible*
Daniel G. Swanson, Bryce D. Bussert and Francis Tyers

*LiMe: A Latin Corpus of Late Medieval Criminal Sentences*
Alessanda Clara Carmela Bassani, Beatrice Giovanna Maria Del Bo, Alfio Ferrara, Marta Luigina Mangini, Sergio Picascia and Ambra Stefanello

*The Rise and Fall of Dependency Parsing in Dante Alighieri's Divine Comedy*
Claudia Corbetta, Marco Passarotti and Giovanni Moretti

*Unsupervised Authorship Attribution for Medieval Latin Using Transformer-Based Embeddings*
Loic De Langhe, Orphee De Clercq and Veronique Hoste

*"To Have the 'Million' Readers Yet": Building a Digitally Enhanced Edition of the Bilingual Irish-English Newspaper an Gaodhal (1881-1898)*
Oksana Dereza, Deirdre Ní Chonghaile and Nicholas Wolf

*Introducing PaVeDa – Pavia Verbs Database: Valency Patterns and Pattern Comparison in Ancient Indo-European Languages*
Silvia Luraghi, Alessio Palmero Aprosio, Chiara Zanchi and Martina Giuliani

**Saturday, May 25, 2024 (continued)**