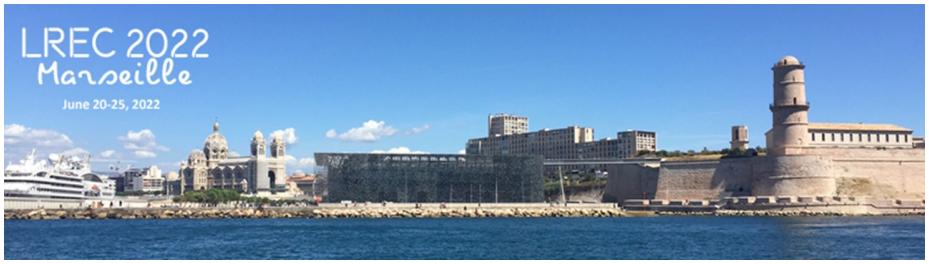

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)

PROCEEDINGS

Editors: Rachele Sprugnoli and Marco Passarotti

Proceedings of the LREC 2022
Second Workshop on Language Technologies for Historical and Ancient Languages
LT4HALA 2022

Edited by: Rachele Sprugnoli and Marco Passarotti

ISBN: 979-10-95546-78-8
EAN: 9791095546788

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Preface

These proceedings include the papers accepted for presentation at the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022).[1] The workshop was held on June 25th 2022 in Marseille, France, co-located with the 13th Edition of the Language Resources and Evaluation Conference (LREC).[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, 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 data for historical and ancient languages, which set them apart from modern languages, with a significant impact on LTs. 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 errors. Also concerned are questions about the automatic processing of various layers of metalinguistic annotation, which are 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 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, including the creation of annotated corpora and advanced computational lexical resources for historical languages, the development of models for performing various NLP tasks, the application of machine translation and linguistic analyses based on the empirical evidence provided by textual resources.

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. In total, the languages tackled in the proceedings are the following: Latin, Italian, Japanese, Chinese, Hungarian, French, Spanish, German, Portuguese, Dutch, Vedic Sanskrit, Ancient Greek (and Cypro-Greek), Ancient Hebrew, Maya, Umbrian and a set of languages of ancient Italy, namely Oscan, Faliscan, Celtic and Venetic.

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 8 pages, plus references) those that describe substantial completed research and/or report on the development of new methodologies. Short papers (up to 4 pages, plus references) were instead more appropriate for reporting on works in progress or for describing a singular tool or project. We encouraged the authors 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, the submission process was single-blind. Each paper was reviewed but three independent reviewers from a program committee made of 24 scholars (12 women and 12 men) from 16 countries. In total, we received 24 submissions from 56 authors from institutions located in 10 countries: Italy (24 authors), Japan (7 authors), Switzerland (6 authors), Germany (5 authors), United States (4 authors), Belgium (3 authors), France (3 authors), Sweden (3 authors), Denmark (1 author), Spain (1 author). After the reviewing process, we accepted 18 submissions, leading to an acceptance rate of 75%.

[1] https://circse.github.io/LT4HALA/2022/

[2] https://lrec2022.lrec-conf.org/en/

LT4HALA 2022 was also the venue of the second 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 second edition of EvaLatin focussed on three shared tasks (i.e. Lemmatization, PoS Tagging, Morphological Features Tagging), each featuring three sub-tasks (i.e. Classical, Cross-Genre, Cross-Time). These sub-tasks were designed to measure the impact of genre and diachrony on NLP tools performances, a relevant aspect to keep in mind when dealing with the diachronic and diatopic diversity of Latin texts, which are spread across a time span of two millennia all over Europe. Participants were provided with shared data in the CoNLL-U format and all the necessary evaluation scripts. They 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, 2 technical reports of EvaLatin, corresponding to as many participants, are included in these proceedings. All reports received a light review by the organizers to check 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 second edition of EvaLatin, like dataset, annotation criteria and results of the shared tasks.
Besides EvaLatin, LT4HALA 2022 hosted also the first edition of EvaHan, an evaluation campaign of NLP tools for the Ancient Chinese language, organized by a team of scholars directed by Bin Li (School of Chinese Language and Literature, Nanjing Normal University), which includes Yiguo Yuan (Nanjing Normal University), Minxuan Feng (Nanjing Normal University), Chao Xu (Nanjing Normal University) and Dongbo Wang (Nanjing Agricultural University).[4] EvaHan focussed on one joint task of Word Segmentation and PoS Tagging. Test data of Ancient Chinese, which is dated back around 1000BC-221BC, were provided in raw format, featuring only Chinese characters and punctuation. The participants were provided with two sets of test data, to evaluate the accuracy rates of the systems respectively on data excerpted from the same work (the Zuozhuan book) included in the training set, without overlapping, and on data from another, yet similar, text. A pretrained model consisting in word embeddings built over a large corpus of traditional Chinese was provided as well. In total, 9 technical reports of EvaHan, corresponding to as many participants, are included in these proceedings. Like for Evalatin, all reports received a light review by the organizers of EvaHan and the proceedings include a short paper with the details of the campaign.
We are grateful to the organizers of EvaHan, who contributed to extend the range of historical and ancient languages of the LT4HALA workshop and showed how some NLP-related issues concern ancient and historical languages per se, despite their typological differences.

Rachele Sprugnoli
Marco Passarotti

[3] https://circse.github.io/LT4HALA/2022/EvaLatin
[4] https://circse.github.io/LT4HALA/2022/EvaHan

Organizers:

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

Program Committee:

Marcel Bollmann, University of Copenhagen (Denmark)
Gerlof Bouma, University of Gothenburg (Sweden)
Harry Diakoff, Alpheios Project (USA)
Stefanie Dipper, Ruhr-Universität Bochum (Germany)
Hanne Eckhoff, Oxford University (UK)
Margherita Fantoli, University of Leuven (Belgium)
Heidi Jauhiainen, University of Helsinki (Finland)
Neven Jovanovic, University of Zagreb (Croatia)
Timo Korkiakangas, University of Helsinki (Finland)
Bin Li, Nanjing Normal University (P.R. China)
Eleonora Litta, Università Cattolica del Sacro Cuore (Italy)
Chao-Lin Liu, National Chengchi University (Taiwan)
Barbara McGillivray, Turing Institute (UK)
Beáta Megyesi, Uppsala University (Sweden)
Giulia Pedonese, Università Cattolica del Sacro Cuore (Italy)
Saskia Peels, University of Groningen (The Netherlands)
Matteo Pellegrini, Università Cattolica del Sacro Cuore (Italy)
Eva Pettersson, Uppsala University (Sweden)
Sophie Prévost, Laboratoire Lattice (France)
Philippe Roelli, University of Zurich (Switzerland)
Matteo Romanello, Université de Lausanne (Switzerland)
Halim Sayoud, USTHB University (Algeria)
Dongbo Wang, Nanjing Agricultural University (P.R. China)

EvaLatin 2022 Organizers:

Rachele Sprugnoli, Università degli Studi di Parma (Italy)
Margherita Fantoli, KU Leuven (Belgium)
Flavio M. Cecchini, Università Cattolica del Sacro Cuore, Milan (Italy)
Marco Passarotti, Università Cattolica del Sacro Cuore, Milan (Italy)

EvaHan 2022 Organizers:

Bin Li, Nanjing Normal University (P.R. China)
Yiguo Yuan, 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

Conference Program

Saturday, June 25, 2022

Long and Short Papers

Identifying Cleartext in Historical Ciphers
Maria-Elena Gambardella, Beata Megyesi and Eva Pettersson

Detecting Diachronic Syntactic Developments in Presence of Bias Terms
Oliver Hellwig and Sven Sellmer

Accurate Dependency Parsing and Tagging of Latin
Sebastian Nehrdich and Oliver Hellwig

Annotating "Absolute" Preverbs in the Homeric and Vedic Treebanks
Luca Brigada Villa, Erica Biagetti and Chiara Zanchi

CHJ-WLSP: Annotation of 'Word List by Semantic Principles' Labels for the Corpus of Historical Japanese
Masayuki Asahara, Nao Ikegami, Tai Suzuki, Taro Ichimura, Asuko Kondo, Sachi Kato and Makoto Yamazaki

The IKUVINA Treebank
Mathieu Dehouck

Machine Translation of 16Th Century Letters from Latin to German
Lukas Fischer, Patricia Scheurer, Raphael Schwitter and Martin Volk

A Treebank-based Approach to the Supprema Constructio in Dante's Latin Works
Flavio Massimiliano Cecchini and Giulia Pedonese

From Inscriptions to Lexica and Back: A Platform for Editing and Linking the Languages of Ancient Italy
Valeria Quochi, Andrea Bellandi, Fahad Khan, Michele Mallia, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi and Cesare Zavattari

BERToldo, the Historical BERT for Italian
Alessio Palmero Aprosio, Stefano Menini and Sara Tonelli

Saturday, June 25, 2022 (continued)

In Search of the Flocks: How to Perform Onomasiological Queries in an Ancient Greek Corpus?
Alek Keersmaekers and Toon Van Hal

Contextual Unsupervised Clustering of Signs for Ancient Writing Systems
Michele Corazza, Fabio Tamburini, Miguel ValÃ©rio and Silvia Ferrara

Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations
Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi and Simonetta Montemagni

Automatic Translation Alignment for Ancient Greek and Latin
Tariq Yousef, Chiara Palladino, David J. Wright and Monica Berti

Handling Stress in Finite-State Morphological Analyzers for Ancient Greek and Ancient Hebrew
Daniel Swanson and Francis Tyers

From Inscription to Semi-automatic Annotation of Maya Hieroglyphic Texts
Cristina Vertan and Christian Prager

Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and Bert-based Models.
Sergio Torres Aguilar

Linguistic Annotation of Neo-Latin Mathematical Texts: A Pilot-Study to Improve the Automatic Parsing of the Archimedes Latinus
Margherita Fantoli and Miryam de Lhoneux

Saturday, June 25, 2022 (continued)

EvaHan Technical Reports

The First International Ancient Chinese Word Segmentation and POS Tagging Bake-off: Overview of the EvaHan 2022 Evaluation Campaign
Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang QU and Dongbo Wang

Automatic Word Segmentation and Part-of-Speech Tagging of Ancient Chinese Based on BERT Model
Yu Chang, Peng Zhu, Chaoping Wang and Chaofan Wang

Ancient Chinese Word Segmentation and Part-of-Speech Tagging Using Data Augmentation
Yanzhi Tian and Yuhang Guo

BERT 4EVER@EvaHan 2022: Ancient Chinese Word Segmentation and Part-of-Speech Tagging Based on Adversarial Learning and Continual Pre-training
Hailin Zhang, Ziyu Yang, Yingwen Fu and Ruoyao Ding

Construction of Segmentation and Part of Speech Annotation Model in Ancient Chinese
Longjie Jiang, Qinyu C. Chang, Huyin H. Xie and Zhuying Z. Xia

Simple Tagging System with RoBERTa for Ancient Chinese
Binghao Tang, Boda Lin and Si Li

The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS
Pengyu Wang and Zhichen Ren

Data Augmentation for Low-resource Word Segmentation and POS Tagging of Ancient Chinese Texts
Yutong Shen, Jiahuan Li, Shujian Huang, Yi Zhou, Xiaopeng Xie and Qinxin Zhao

A Joint Framework for Ancient Chinese WS and POS Tagging Based on Adversarial Ensemble Learning
Shuxun Yang

Glyph Features Matter: A Multimodal Solution for EvaHan in LT4HALA2022
Wei Xinyuan, Liu Weihao, Qing Zong , Zhang Shao Qing and Baotian Hu

Saturday, June 25, 2022 (continued)

EvaLatin Technical Reports

Overview of the EvaLatin 2022 Evaluation Campaign
Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli and Giovanni Moretti

An ELECTRA Model for Latin Token Tagging Tasks
Wouter Mercelis and Alek Keersmaekers

Transformer-based Part-of-Speech Tagging and Lemmatization for Latin
Krzysztof WrÃ³bel and Krzysztof Nowak