LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

1st Workshop on Language Technologies for Historical and Ancient Languages, (LT4HALA 2020)

PROCEEDINGS

Editors: Rachele Sprugnoli and Marco Passarotti

Proceedings of the LREC 2020
1st Workshop on Language Technologies for Historical and Ancient Languages
(LT4HALA 2020)

Edited by: Rachele Sprugnoli and Marco Passarotti

ISBN: 979-10-95546-53-5
EAN: 9791095546535

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 1st Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA: https://circse.github.io/LT4HALA). The workshop was supposed to be held on May 12th 2020 in Marseille, France, co-located with the 12th Edition of the Language Resources and Evaluation Conference (LREC 2020). Unfortunately, the gravity of the Covid-19 pandemic prevented the conference from taking place. However, since the spread of the pandemic started to rise at world-level when the reviewing process and the notifications of acceptance/rejection of the proposals were just concluded, the organizers decided to publish the proceedings of both LREC 2020 and the co-located workshops as planned in May 2020, to valorize the work done by authors and reviewers, as well as to provide an overview of the state of the art in the field.

The objective of the LT4HALA workshop is to bring together scholars who are developing and/or are using Language Technologies (LTs) for historically attested languages, so 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.
Despite the current availability of large collections of digitized texts written in historical languages, such interdisciplinary collaboration is still hampered by the limited availability of annotated linguistic resources for most of the historical languages. Creating such resources is a challenge and an obligation for LTs, both to support historical linguistic research with the most updated technologies and to preserve those precious linguistic data that survived from past times.
Historical and ancient languages present several characteristics, which set them apart from modern languages, with a significant impact on LTs. Typically, historical and ancient languages lack large linguistic resources, such as annotated corpora, and data can be sparse and very inconsistent; texts present considerable orthographic variation, they can be transmitted by different witnesses and in different critical editions, they can be incomplete and scattered across a wide temporal and geographical span. This makes the selection of representative texts, and thus the development of benchmarks, very hard. Moreover, texts in machine-readable format are often the result of manuscript digitization processes during which OCR systems can cause errors degrading the quality of the documents. Another peculiarity is that most of the texts written in historical and ancient languages are literary, philosophical or documentary, therefore of a very different genre from that on which LTs are usually trained, i.e. news. This is strictly connected to the fact that the final users of LTs for historical and ancient languages are mostly humanists who expect a high accuracy of results that allows a precise analysis of linguistic data.

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 methods for automatic dating ancient texts and performing semantic analysis, processes for developing linguistic resources and performing various natural language processing (NLP) tasks, like lemmatization and semantic role labelling, and applications of machine translation and distributional semantics, speech analysis and diachronic phonology, automatic inflectional morphology and computational philology.

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 21 (note that some papers deal with more than one language), namely: Latin (5 papers), French (3), English (2), Hebrew (2), Italian (2), Spanish (2), Ancient Greek (1), Aramaic (1), Armenian (1), Georgian (1), German (1), Norwegian (1), Old Chinese (1), Portuguese (1), Romanian (1), Serbian (1), Slovene (1), Syriac (1), Vedic Sanskrit (1) and the unknown writing system of the so-called Voynich manuscript (1).

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 not anonymous. Each paper was reviewed but three independent reviewers from a program committee made of 25 scholars (12 women and 13 men) from 15 countries.
In total, we received 23 submissions from 47 authors of 13 countries: China (7 authors), France (6), Ireland (5), The Netherlands (5), Poland (5), United Stated (5), Malta (4), Belgium (3), Israel (2), Spain (2), Estonia (1), Italy (1) and Switzerland (1). After the reviewing process, we accepted 15 submissions (8 long and 7 short papers), leading to an acceptance rate of 65.22% .
Beside these 15 contributions, the program of LT4HALA would have featured also a keynote speech by Amba Kulkarni (Department of Sanskrit Studies, University of Hyderabad, India) about the challenges raised by the development of computational tools for Sanskrit. We had invited Professor Kulkarni to give a talk on this topic, because Sanskrit holds a prominent position among historical and ancient languages, being one of the oldest documented members of the Indo-European family of languages.

LT4HALA was supposed to be also the venue of the first edition of EvaLatin, the first campaign devoted to the evaluation of NLP tools for Latin (`https://circse.github.io/LT4HALA/EvaLatin`). Just because of the limited amount of data preserved for historical and ancient languages, an important role is played by evaluation practices, to understand the level of accuracy of the NLP tools used to build and analyze resources. By organizing EvaLatin, we decided to focus on Latin, considering its prominence among the ancient and historical languages, as demonstrated also by the high number of papers dealing with Latin in these proceedings. The first edition of EvaLatin focussed on two shared tasks (i.e. Lemmatization and PoS 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, 5 technical reports of EvaLatin, corresponding to as many participants, are included in these proceedings. All reports received a light review by the two of us, 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 short paper detailing some specific aspects of EvaLatin, like the composition, source, tag set and annotation criteria of the shared data.

Although we are very sorry that the LT4HALA workshop and EvaLatin could not be held, as an exciting opportunity to meet in person the authors who contributed to these proceedings, we hope that this will give us a further argument to organize a second edition of both initiatives. Indeed, as demonstrated by the good number of papers submitted to LT4HALA and participants of EvaLatin, the research field concerned is wide, diverse and lively: we will do our best to provide the scholars working in such field with a venue where they can present their work and confront with colleagues who share their research interests.

Rachele Sprugnoli
Marco Passarotti

Organizers:

Rachele Sprugnoli, Università Cattolica del Sacro Cuore (Italy)
Marco Passarotti, Università Cattolica del Sacro Cuore (Italy)

Program Committee:

Marcel Bollmann, University of Copenhagen (Denmark)
Gerlof Bouma, University of Gothenburg (Sweden)
Patrick Burns, University of Texas at Austin (USA)
Flavio Massimiliano Cecchini, Università Cattolica del Sacro Cuore (Italy)
Oksana Dereza, Insight Centre for Data Analytics (Ireland)
Stefanie Dipper, Ruhr-Universität Bochum (Germany)
Hanne Eckoff, Oxford University (UK)
Maud Ehrmann, EPFL (Switzerland)
Hannes A. Fellner, Universität Wien (Austria)
Heidi Jauhiainen, University of Helsinki (Finland)
Julia Krasselt, Zurich University of Applied Sciences (Switzerland)
John Lee, City University of Hong Kong (Hong Kong)
Chao-Lin Liu, National Chengchi University (Taiwan)
Barbara McGillivray, University of Cambridge (UK)
Beáta Megyesi, Uppsala University (Sweden)
So Miyagawa, University of Göttingen (Germany)
Joakim Nivre, Uppsala University (Sweden)
Andrea Peverelli, Università Cattolica del Sacro Cuore (Italy)
Eva Pettersson, Uppsala University (Sweden)
Michael Piotrowski, University of Lausanne (Switzerland)
Sophie Prévost, Laboratoire Lattice (France)
Halim Sayoud, USTHB University (Algeria)
Olga Scrivner, Indiana University (USA)
Amir Zeldes, Georgetown University (USA)
Daniel Zeman, Charles University (Czech Republic)

Invited Speaker:

Amba Kulkarni, University of Hyderabad (India)

Table of Contents