

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

Marseille, France
11-16 May 2020

Editors:

**Rachel Sprugnoli
Marco Passarotti**

ISBN: 978-1-7138-1269-2

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2020) by the Association for Computational Linguistics
All rights reserved.

Copyright for individual papers remains with the authors and are licensed under a Creative Commons 4.0 license, CC-BY-NC. (<https://creativecommons.org/licenses/by-nc/4.0/>)

Printed with permission by Curran Associates, Inc. (2020)

For permission requests, please contact the Association for Computational Linguistics at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006

Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

| | |
|---|-----|
| <i>Dating and Stratifying a Historical Corpus with a Bayesian Mixture Model</i> Oliver Hellwig | 1 |
| <i>Automatic Construction of Aramaic-Hebrew Translation Lexicon</i> Chaya Liebeskind and Shmuel Liebeskind | 10 |
| <i>Dating Ancient texts: an Approach for Noisy French Documents</i> Anaëlle Baledent, Nicolas Hiebel and Gaël Lejeune | 17 |
| <i>Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old Georgian, and Syriac</i> Chahan Vidal-Gorène and Bastien Kindt | 22 |
| <i>Computerized Forward Reconstruction for Analysis in Diachronic Phonology, and Latin to French Reflex Prediction</i> Clayton Marr and David R. Mortensen | 28 |
| <i>Using LatInfLexi for an Entropy-Based Assessment of Predictability in Latin Inflection</i> Matteo Pellegrini | 37 |
| <i>A Tool for Facilitating OCR Postediting in Historical Documents</i> Alberto Poncelas, Mohammad Aboomar, Jan Buts, James Hadley and Andy Way | 47 |
| <i>Integration of Automatic Sentence Segmentation and Lexical Analysis of Ancient Chinese based on BiLSTM-CRF Model</i> Ning Cheng, Bin Li, Liming Xiao, Changwei Xu, Sijia Ge, Xingyue Hao and Minxuan Feng | 52 |
| <i>Automatic semantic role labeling in Ancient Greek using distributional semantic modeling</i> Alek Keersmaekers | 59 |
| <i>A Thesaurus for Biblical Hebrew</i> Miriam Azar, Aliza Pahmer and Joshua Waxman | 68 |
| <i>Word Probability Findings in the Voynich Manuscript</i> Colin Layfield, Lonke van der Plas, Michael Rosner and John Abela | 74 |
| <i>Comparing Statistical and Neural Models for Learning Sound Correspondences</i> Clémentine Fourier and Benoît Sagot | 79 |
| <i>Distributional Semantics for Neo-Latin</i> Jelke Bloem, Maria Chiara Parisi, Martin Reynaert, Yvette Oortwijn and Arianna Betti | 84 |
| <i>Latin-Spanish Neural Machine Translation: from the Bible to Saint Augustine</i> Eva Martínez García and Álvaro García Tejedor | 94 |
| <i>Detecting Direct Speech in Multilingual Collection of 19th-century Novels</i> Joanna Byszuk, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa and Maciej Eder | 100 |
| <i>Overview of the EvaLatin 2020 Evaluation Campaign</i> Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini and Matteo Pellegrini | 105 |

| | |
|--|-----|
| <i>Data-driven Choices in Neural Part-of-Speech Tagging for Latin</i> Geoff Bacon | 111 |
| <i>JHUBC's Submission to LT4HALA EvaLatin 2020</i> Winston Wu and Garrett Nicolai | 114 |
| <i>A Gradient Boosting-Seq2Seq System for Latin POS Tagging and Lemmatization</i> Celano Giuseppe | 119 |
| <i>UDPipe at EvaLatin 2020: Contextualized Embeddings and Treebank Embeddings</i> Milan Straka and Jana Straková | 124 |
| <i>Voting for POS tagging of Latin texts: Using the flair of FLAIR to better Ensemble Classifiers by Example of Latin</i> Manuel Stoeckel, Alexander Henlein, Wahed Hemati and Alexander Mehler | 130 |