

1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)

Bangkok, Thailand
15 August 2024

ISBN: 979-8-3313-0170-5

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© (2024) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

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>Challenging Error Correction in Recognised Byzantine Greek</i>	
John Pavlopoulos, Vasiliki Kougia, Esteban Garces Arias, Paraskevi Platanou, Stepan Shabalin, Konstantina Liagkou, Emmanouil Papadatos, Holger Essler, Jean-Baptiste Camps and Franz Fischer .	1
<i>MsBERT: A New Model for the Reconstruction of Lacunae in Hebrew Manuscripts</i>	
Avi Shmidman, Ometz Shmidman, Hillel Gershuni and Moshe Koppel	13
<i>Predicate Sense Disambiguation for UMR Annotation of Latin: Challenges and Insights</i>	
Federica Gamba	19
<i>Classification of Paleographic Artifacts at Scale: Mitigating Confounds and Distribution Shift in Cuneiform Tablet Dating</i>	
Danlu Chen, Jiahe Tian, Yufei Weng, Taylor Berg-Kirkpatrick and Jacobo Myerston	30
<i>Classifier identification in Ancient Egyptian as a low-resource sequence-labelling task</i>	
Dmitry Nikolaev, Jorke Grotenhuis, Haleli Harel and Orly Goldwasser	42
<i>Long Unit Word Tokenization and Bunsetsu Segmentation of Historical Japanese</i>	
Hiroaki Ozaki, Kanako Komiya, Masayuki Asahara and Toshinobu Ogiso	48
<i>A new machine-actionable corpus for ancient text restoration</i>	
Will Fitzgerald and Justin Barney	56
<i>Lacuna Language Learning: Leveraging RNNs for Ranked Text Completion in Digitized Coptic Manuscripts</i>	
Lauren Elizabeth Levine, Cindy Tung Li, Lydia BremerMcCollum, Nicholas E. Wagner and Amir Zeldes	61
<i>Deep Learning Meets Egyptology: a Hieroglyphic Transformer for Translating Ancient Egyptian</i>	
Mattia De Cao, Nicola De Cao, Angelo Colonna and Alessandro Lenci	71
<i>Neural Lemmatization and POS-tagging models for Coptic, Demotic and Earlier Egyptian</i>	
Aleksi Sahala and Eliese-Sophia Lincke	87
<i>UFCNet: Unsupervised Network based on Fourier transform and Convolutional attention for Oracle Character Recognition</i>	
Yanan Zhou, Guoqi Liu, Yiping Yang, Linyuan Ru, Dong Liu and Xueshan Li	98
<i>Coarse-to-Fine Generative Model for Oracle Bone Inscriptions Inpainting</i>	
Shibin Wang, Wenjie Guo, Yubo Xu, Dong Liu and Xueshan Li	107
<i>Restoring Mycenaean Linear B 'A&B' series tablets using supervised and transfer learning</i>	
Katerina Papavassileiou and Dimitrios Kosmopoulos	115
<i>CuReD: Deep Learning Optical Character Recognition for Cuneiform Text Editions and Legacy Materials</i>	
Shai Gordin, Morris Alper, Avital Romach, Luis Daniel Saenz Santos, Naama Yochai and Roey Lalazar	130
<i>Towards Context-aware Normalization of Variant Characters in Classical Chinese Using Parallel Editions and BERT</i>	
Florian Kessler	141

<i>Gotta catch ‘em all!’: Retrieving people in Ancient Greek texts combining transformer models and domain knowledge</i>	
Marijke Beersmans, Alek Keersmaekers, Evelien de Graaf, Tim Van De Cruys, Mark Depauw and Margherita Fantoli	152
<i>Adapting transformer models to morphological tagging of two highly inflectional languages: a case study on Ancient Greek and Latin</i>	
Alek Keersmaekers and Wouter Mercelis	165
<i>A deep learning pipeline for the palaeographical dating of ancient Greek papyrus fragments</i>	
Graham West, Matthew I. Swindall, James H. Brusuelas and John Wallin	177
<i>UD-ETCSUX: Toward a Better Understanding of Sumerian Syntax</i>	
Kenan Jiang and Adam G Anderson	186
<i>SumTablets: A Transliteration Dataset of Sumerian Tablets</i>	
Cole Simmons, Richard Diehl Martinez and Dan Jurafsky	192
<i>Latin Treebanks in Review: An Evaluation of Morphological Tagging Across Time</i>	
Marisa Hudspeth, Brendan O’Connor and Laure Thompson	203
<i>The Metronome Approach to Sanskrit Meter: Analysis for the Rigveda</i>	
Yuzuki Tsukagoshi and Ikki Ohmukai	219
<i>Ancient Wisdom, Modern Tools: Exploring Retrieval-Augmented LLMs for Ancient Indian Philosophy</i>	
Priyanka Mandikal	224
<i>Leveraging Part-of-Speech Tagging for Enhanced Stylometry of Latin Literature</i>	
Sarah Li Chen, Patrick J. Burns, Thomas J. Bolt, Pramit Chaudhuri and Joseph P. Dexter	251
<i>Exploring intertextuality across the Homeric poems through language models</i>	
Maria Konstantinidou, John Pavlopoulos and Elton Barker	260