

ML4AL 2024

1st Workshop on Machine Learning for Ancient Languages

Proceedings of the Workshop

August 15, 2024

The ML4AL organizers gratefully acknowledge the support from the following sponsors.

Diamond Tier

 Google DeepMind

Silver Tier

 **Vesuvius Challenge**

Supporting Organisations



©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-144-5

Preface by the General Chair

Welcome to the proceedings of the 1st Machine Learning for Ancient Languages (**ML4AL**) Workshop, held as part of the Annual Conference of the Association for Computational Linguistics (ACL) 2024. Taking place on August 15th, 2024, this is a hybrid event with virtual and on-site participation in Thailand.

ML4AL showcases the scientific opportunities at the intersection of the Humanities and ML, representing a unique convergence between the two and spotlighting promising directions for future endeavors within this rising field. By leveraging advances in AI and by focusing on the study and preservation of ancient texts, ML4AL aims to inspire collaboration and support research momentum in the emerging field of ML for the study of ancient languages.

On its 1st year, ML4AL received 50 submissions from a global community of researchers. The submissions concerned multiple languages, including Ancient Greek, Latin, Sumerian and Akkadian, Classical and Old Chinese, ancient Egyptian, Coptic, etc. 18 papers were accepted for oral presentation (36%) and 10 were accepted as posters (20%). The accepted submissions covered diverse topics, such as digitization, restoration, attribution, linguistic analysis, textual criticism, translation, and decipherment of ancient texts. These contributions reflect the depth and breadth of current research and highlight the innovative approaches being developed to tackle the unique challenges posed by ancient languages.

Besides the oral and poster presentations, ML4AL features two distinguished keynote talks to provide valuable perspectives on the integration of machine learning for the study of ancient texts. The talk of Dr Stephen Parsons from Educe Lab, University of Kentucky, USA concerns the virtual unwrapping of the Herculaneum Scrolls. The talk of Professor JinYeong Bak from the Department of Computer Science and Engineering, Sungkyunkwan University, South Korea focuses on monarchical ruling styles when applying ML to historical corpora.

The ML4AL Organising Committee is grateful: to the keynote speakers for their stimulating talks; the authors for their valuable contributions; the members of the Program Committee for their hard work. We would like to particularly thank our emergency reviewers, who provided very valuable expertise in a very limited time window. We would also like to extend our gratitude to the ACL 2024 Workshop Chairs for their kind assistance, and to our sponsors and supporting organization for their generous contributions. Specifically, Google DeepMind was our diamond-tier sponsor, the Vezuvius Challenge was our silver-tier sponsor, and Archimedes/Athena RC was our supporting organization.

Hopefully, the discussions and collaborations initiated at this workshop will lead to significant advancements in the study of ancient languages and foster a deeper understanding of our shared human heritage.

Sincerely,

John Pavlopoulos, General Chair

Organizing Committee

General Chair

John Pavlopoulos, Athens University of Economics and Business, Archimedes/Athena RC, Greece

Co-Chair

Thea Sommerschild, University of Nottingham, UK

Yannis Assael, Google DeepMind, UK

Shai Gordin, Ariel University, Israel

Organizing Committee

Kyunghyun Cho, NYU, CIFAR, Genentech, USA

Marco Passarotti, Università Cattolica del Sacro Cuore, Italy

Rachele Sprugnoli, Università di Parma, Italy

Yudong Liu, Western Washington University, USA

Bin Li, Nanjing Normal University, China

Adam Anderson, UC Berkeley, USA

Program Committee

Program Chairs

John Pavlopoulos, Athens University of Economics and Business, Archimedes/Athena RC, Greece
Thea Sommerschild, University of Nottingham, UK
Yannis Assael, Google DeepMind
Shai Gordin, Ariel University, Israel
Kyunghyun Cho, NYU, CIFAR, Genentech, USA
Adam Anderson, University of California, Berkeley
Yudong Liu, Western Washington University
Marco Passarotti, Università Cattolica del Sacro Cuore
Rachele Sprugnoli, University of Parma
Bin Li, Nanjing Normal University, China

Reviewers

Masayuki Asahara
John Bodel
Flavio Cecchini
Heo Chul
Claudia Corbetta
Angelo Del Grosso
Mark Depauw
Hanne Eckhoff
Margherita Fantoli
Ethan Fetaya
Theodorus Fransen
Federica Gamba
Petra Heřmánková
Marietta Horster
Renfen Hu
Federica Iurescia
Kyle Johnson
Alek Keersmaekers
Els Lefever
Chaya Liebeskind
Eliese-Sophia Lincke
Chao-Lin Liu
Liu Liu
Jiaming Luo
Massimo Maiocchi
Isabelle Marthot-Santaniello
Barbara McGillivray
M. Monroe
Alex Mullen
Chiara Palladino
Chanjun Park
Matteo Pellegrini
Edoardo Ponti

Mladen Popović
Jonathan Prag
Avital Romach
Edgar Roman-Rangel
Matteo Romanello
William Seales
Andrew Senior
Si Shen
Gustav Smidt
Richard Sproat
Gabriel Stanovsky
Silvia Stopponi
Qi Su
Matthew Swindall
Xuri Tang
Charlotte Tupman
Haneul Yoo
Chongsheng Zhang

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, John Wallin, Francesca Maltomini, Marius Gerhardt, Marzia D’Angelo and John F. 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, Prमित 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
<i>Detecting Narrative Patterns in Biblical Hebrew and Greek</i>	
Hope McGovern, Hale Sirin, Tom Lippincott and Andrew Caines	269

Program

Thursday, August 15, 2024

09:15 - 09:30 *Introduction*

09:30 - 10:30 *Session 1*

Towards Context-aware Normalization of Variant Characters in Classical Chinese Using Parallel Editions and BERT

Florian Kessler

Ancient Wisdom, Modern Tools: Exploring Retrieval-Augmented LLMs for Ancient Indian Philosophy

Priyanka Mandikal

A new machine-actionable corpus for ancient text restoration

Will Fitzgerald and Justin Barney

Lacuna Language Learning: Leveraging RNNs for Ranked Text Completion in Digitized Coptic Manuscripts

Lauren Elizabeth Levine, Cindy Tung Li, Lydia BremerMcCollum, Nicholas E. Wagner and Amir Zeldes

A deep learning pipeline for the palaeographical dating of ancient Greek papyrus fragments

Graham West, Matthew I. Swindall, James H. Brusuelas, John Wallin, Francesca Maltomini, Marius Gerhardt, Marzia D'Angelo and John F. Wallin

Coarse-to-Fine Generative Model for Oracle Bone Inscriptions Inpainting

Shibin Wang, Wenjie Guo, Yubo Xu, Dong Liu and Xueshan Li

10:30 - 11:00 *Coffee and Posters I*

Exploring intertextuality across the Homeric poems through language models

Maria Konstantinidou, John Pavlopoulos and Elton Barker

The Metronome Approach to Sanskrit Meter: Analysis for the Rigveda

Yuzuki Tsukagoshi and Ikki Ohmukai

UD-ETCSUX: Toward a Better Understanding of Sumerian Syntax

Kenan Jiang and Adam G Anderson

Thursday, August 15, 2024 (continued)

Adapting transformer models to morphological tagging of two highly inflectional languages: a case study on Ancient Greek and Latin

Alek Keersmaekers and Wouter Mercelis

Gotta catch 'em all!": Retrieving people in Ancient Greek texts combining transformer models and domain knowledge

Marijke Beersmans, Alek Keersmaekers, Evelien de Graaf, Tim Van De Cruys, Mark Depauw and Margherita Fantoli

11:00 - 11:30 *Keynote Talk by Dr Stephen Parsons, University of Kentucky, USA*

11:30 - 12:30 *Session 2*

Classification of Paleographic Artifacts at Scale: Mitigating Confounds and Distribution Shift in Cuneiform Tablet Dating

Danlu Chen, Jiahe Tian, Yufei Weng, Taylor Berg-Kirkpatrick and Jacobo Myerston

Latin Treebanks in Review: An Evaluation of Morphological Tagging Across Time

Marisa Hudspeth, Brendan O'Connor and Laure Thompson

Leveraging Part-of-Speech Tagging for Enhanced Stylometry of Latin Literature

Sarah Li Chen, Patrick J. Burns, Thomas J. Bolt, Prमित Chaudhuri and Joseph P. Dexter

SumTablets: A Transliteration Dataset of Sumerian Tablets

Cole Simmons, Richard Diehl Martinez and Dan Jurafsky

UFCNet: Unsupervised Network based on Fourier transform and Convolutional attention for Oracle Character Recognition

Yanan Zhou, Guoqi Liu, Yiping Yang, Linyuan Ru, Dong Liu and Xueshan Li

Long Unit Word Tokenization and Bunsetsu Segmentation of Historical Japanese

Hiroaki Ozaki, Kanako Komiya, Masayuki Asahara and Toshinobu Ogiso

12:30 - 13:45 *Lunch Break*

14:00 - 14:30 *Keynote Talk by Professor JinYeong Bak, Sungkyunkwan University, South Korea*

Thursday, August 15, 2024 (continued)

14:30 - 15:30 *Session 3*

CuReD: Deep Learning Optical Character Recognition for Cuneiform Text Editions and Legacy Materials

Shai Gordin, Morris Alper, Avital Romach, Luis Daniel Saenz Santos, Naama Yochai and Roey Lalazar

Neural Lemmatization and POS-tagging models for Coptic, Demotic and Earlier Egyptian

Aleksi Sahala and Eliese-Sophia Lincke

Challenging Error Correction in Recognised Byzantine Greek

John Pavlopoulos, Vasiliki Kougia, Esteban Garces Arias, Paraskevi Platanou, Stepan Shabalin, Konstantina Liagkou, Emmanouil Papadatos, Holger Essler, Jean-Baptiste Camps and Franz Fischer

MsBERT: A New Model for the Reconstruction of Lacunae in Hebrew Manuscripts

Avi Shmidman, Ometz Shmidman, Hillel Gershuni and Moshe Koppel

Deep Learning Meets Egyptology: a Hieroglyphic Transformer for Translating Ancient Egyptian

Mattia De Cao, Nicola De Cao, Angelo Colonna and Alessandro Lenci

Classifier identification in Ancient Egyptian as a low-resource sequence-labelling task

Dmitry Nikolaev, Jorke Grotenhuis, Haleli Harel and Orly Goldwasser

15:30 - 16:00 *Coffee and Posters II*

Detecting Narrative Patterns in Biblical Hebrew and Greek

Hope McGovern, Hale Sirin, Tom Lippincott and Andrew Caines

Restoring Mycenaean Linear B 'A&B' series tablets using supervised and transfer learning

Katerina Papavassileiou and Dimitrios Kosmopoulos

Application of Machine Learning to the Critical Edition of Ancient Greek Inscriptions: Ithaca and the Corpus of Oracular Inscriptions of Dodona

Elena Martín González

A Dataset for Metaphor Detection in Early Medieval Hebrew Poetry

Michael Toker, Oren Mishali, Ophir Münz-Manor, Benny Kimelfeld and Yonatan Belinkov

Thursday, August 15, 2024 (continued)

Predicate Sense Disambiguation for UMR Annotation of Latin: Challenges and Insights

Federica Gamba

16:00 - 17:00 *Round Table*

17:00 - 17:30 *Best Paper Award*

17:30 - 17:45 *Closing Remarks*