NAACL-ALP 2025

ALP 2025

Proceedings of the Second Workshop on Ancient Language Processing

associated with

The 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics
NAACL 2025

May 4, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-235-0

Preface

The Second International Workshop on Ancient Language Processing (ALP 2025), co-located with the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025), was held on May 4, 2025, in Albuquerque, New Mexico, USA. This workshop built on the success of its inaugural edition (ALP 2023 at Varna, Bulgaria) and the workshop on Ancient Language Translation (ALT 2023 at Macao, China). ALP 2025 further consolidated the global platform for scholars and practitioners exploring the intersection of natural language processing (NLP) and ancient languages—a domain rich with historical, cultural, and linguistic significance.

Ancient languages, spanning from Sumerian cuneiform (c. 3,4000 BCE) to Classical Chinese and Mayan glyphs, encapsulate humanity's earliest intellectual and cultural achievements. The ALP 2025 workshop sought to advance the application of modern NLP techniques to these languages, addressing challenges such as non-Latin scripts, transliteration variability, fragmented texts, and dialectal diversity. By fostering interdisciplinary collaboration, the workshop aimed to accelerate progress in digitizing and analyzing ancient linguistic resources, thereby unlocking new insights into human history and culture.

ALP 2025 received a diverse array of submissions covering the earliest phases of writing (proto-cuneiform), to the first languages from Mesopotamia (Akkadian, Sumerian and Akkadian), Iran (Old Persian), Egypt and Sudan (Egyptian, Demotic, and Meroitic), Anatolia (Hittite), the Mediterranean (Hebrew, Linear B, Ancient Greek), Iran (Old Persian), India (Sanskrit), East Asia (Classical Chinese, Old Tibetan, Old Japanese), and Mesoamerica (Mayan). Notable themes included: Resource development and Innovations in corpus construction, in Unicode input methods, and linguistically linked data for underrepresented scripts, and the application of Large Language Models (LLMs) in various ancient languages.

The workshop hosted two shared tasks — EvaCun 2025 (cuneiform lemmatization and text restoration analysis using LLMs) and EvaHan 2025 (classical Chinese named entity recognition) — designed to benchmark progress on unique challenges.

The workshop received 43 submissions in total. After rigorous double-blind review, the program committee accepted 33 papers: 10 long papers, 7 short papers, 2 overview papers of shared tasks, and 14 technical reports of shared tasks. Accepted works spanned methodologies from rule-based systems to deep learning, with a growing emphasis on the application of LLMs.

The ALP series reflects a rapidly expanding research community, driven by increased digitization of ancient texts and interdisciplinary interest from computational linguists, archaeologists, and philologists. This year's workshop features keynote addressed by Dr. Patrick J. Burns (NYU, ISAW, US), a pioneer in digital philology and developer of the classical language toolkits, and Dr. Donald Sturgeon (Durham University, UK), founder of the Classical Chinese Text Project. Their contributions exemplify the synergy between traditional scholarship and cutting-edge NLP.

We extend our gratitude to the ALP 2025 Program Committee for their meticulous reviews, the NAACL 2025 organizers for logistical support, and the student committee members and volunteers whose dedication ensured a seamless process. Special thanks to the shared task coordinators of EvaCun2 025 and EvaHan 2025 for designing evaluations that push the boundaries of ancient NLP.

The workshop ALP 2025 afforded the opportunity for We invited all participants to engage in dynamic discussions, share novel ideas, and contribute to a future where ancient languages are not merely preserved but actively integrated into the global ancient language processing and culture computation landscape.

Organizing Committee

Workshop Organizers

Adam Anderson, UC Berkeley, USA
Shai Gordin, Ariel University, Israel
Bin Li, Nanjing Normal University, China
Yudong Liu, Western Washington University, USA
Marco C. Passarotti, Università Cattolica del Sacro Cuore, Italy
Rachele Sprugnoli, University of Parma, Italy

Program Committee

Alaa Mamdouh Akef, Peking University, China
Chao-Lin Liu, National Chengchi University, Taiwan
Christian M. Prager, University of Bonn, Germany
Congjun Long, Chinese Academy of Social Sciences, China
Dongbo Wang, Nanjing Agricultural University, China
Francesco Mambrini, Università Cattolica del Sacro Cuore, Milan, Italy
Gregory Crane, Tufts University, USA
Heidi Jauhiainen, University of Helsinki, Finland
Jacob Murel, Princeton University, USA
Johann-Mattis List, Max Planck Institute for Evolutionary Anthropology, Germany
Jonathan Berant, Tel Aviv University, Israel
Kyle P. Johnson, Berlin-Brandenburg Academy of Sciences, Germany
Liu Liu, Nanjing Agricultural University, China
Long long Ma, University of Chinese Academy of Sciences, China
Luis Sáenz, Ariel University/Heidelberg University, Israel/Germany
Martijn Naaijer, University of Copenhagen, Denmark
Masayuki Asahara, National Institute for Japanese Language and Linguistics, Japan
Minxuan Feng, Nanjing Normal University, China
Monica Berti, Leipzig University, Germany
Niek Veldhuis, University of California, Berkeley, USA
Orly Lewis, Hebrew University of Jerusalem, Israel
Qi Su, Peking University, China
Rachele Sprugnoli, Università degli Studi di Parma, Italy
Renfen Hu, Beijing Normal University, China
Sanhong Deng, Nanjing University, China
Si Shen, Nanjing University of Science and Technology, China
Thea Sommerschield, Ca' Foscari University of Venice, Italy
Toon Van Hal, University of Leuven, Belgium
Xuri Tang, Huazhong University of Science and Technology, China

Student Committee

Stav Klein (Ariel University/TAD Center, Tel Aviv University, Israel)
Zhixing Xu (Nanjing Normal University, China)
Bolin Chang (Nanjing Normal University, China)
Xue Zhao (Nanjing Agricultural University, China)
Ruilin Liu (Nanjing Agricultural University, China)
Jing Chen (Hong Kong Poly University, Hong Kong)
Claudia Corbetta (claudia.corbetta@unibg.it, University of Bergamo, Italy)
Danlu Chen (danlu@ucsd.edu, UC San Diego, USA)
Ercong Nie (nie@cis.lmu.de, Ludwig Maximilian University of Munich, Germany)

Table of Contents

Conference Program

9:00–9:05 **Opening**

9:05–9:35 **Invited Talk1 (Patrick)**

9:35–10:50 **Oral Reports**

Automatic Text Segmentation of Ancient and Historic Hebrew
Elisha Rosensweig, Benjamin Resnick, Hillel Gershuni, Joshua Guedalia, Nachum Dershowitz and Avi Shmidman

Integrating Semantic and Statistical Features for Authorial Clustering of Qumran Scrolls
Yonatan Lourie, Jonathan Ben-Dov and Roded Sharan

Assignment of account type to proto-cuneiform economic texts with Multi-Class Support Vector Machines
Piotr Zadworny and Shai Gordin

Accessible Sanskrit: A Cascading System for Text Analysis and Dictionary Access
Giacomo De Luca

A Dataset of Ancient Chinese Math Word Problems and an Application for Research in Historic Mathematics
Florian Keßler

10:50–11:00 **Break**

11:00–12:00 Poster

Using Cross-Linguistic Data Formats to Enhance the Annotation of Ancient Chinese Documents Written on Bamboo Slips
Michele Pulini and Johann-Mattis List

Towards an Integrated Methodology of Dating Biblical Texts: The Case of the Book of Jeremiah
Martijn Naaijer and Aren Wilson-Wright

The Development of Hebrew in Antiquity – A Computational Linguistic Study
Hallel Baitner, dimid duchovny, Lee-Ad Gottlieb, Amir Yorav, Nachum Dershowitz and Eshbal Ratzon

Evaluating Evaluation Metrics for Ancient Chinese to English Machine Translation
Eric R. Bennett, HyoJung Han, Xinchen Yang, Andrew Schonebaum and Marine Carpuat

Neural Models for Lemmatization and POS-Tagging of Earlier and Late Egyptian (Supporting Hieroglyphic Input) and Demotic
Aleksi Sahala and Eliese-Sophia Lincke

Bringing Suzhou Numerals into the Digital Age: A Dataset and Recognition Study on Ancient Chinese Trade Records
Ting-Lin Wu, Zih-Ching Chen, Chen-Yuan Chen, Pi-Jhong Chen and Li-Chiao Wang

The Historian's Fingerprint: Computational Stylometric Analysis of the Zuo Commentary and Discourses of the States
Wenjie Hua

12:00–12:30 Lunch

13:30–14:00 Invited Talk 2 (Sturgeon)

14:00–15:30 EvaHan

Overview of EvaHan2025: The First International Evaluation on Ancient Chinese Named Entity Recognition
Bin Li, Bolin Chang, Ruilin Liu, Xue Zhao, Si Shen, Lihong Liu, Yan Zhu, Zhixing Xu, Weiguang QU and Dongbo Wang

Exploring the Application of 7B LLMs for Named Entity Recognition in Chinese Ancient Texts
yicheng zhu and han bi

Construction of NER Model in Ancient Chinese: Solution of EvaHan 2025 Challenge
Yi Lu and Minyi Lei

LLM's Weakness in NER Doesn't Stop It from Enhancing a Stronger SLM
Weilu Xu, Renfei Dang and Shujian Huang

Named Entity Recognition in Context: Edit_Dunhuang team Technical Report for Evahan2025 NER Competition
Colin Brisson, Ayoub Kahfy, Marc Bui and Frédéric Constant

Simple Named Entity Recognition (NER) System with RoBERTa for Ancient Chinese
Yunmeng Zhang, Meiling Liu, Hanqi Tang, Shige Lu and Lang Xue

Make Good Use of GujiRoBERTa to Identify Entities in Ancient Chinese
Lihan Lin, Yiming Wang, Jiachen Li, Huan Ouyang and Si Li

GRoWE: A GujiRoBERTa-Enhanced Approach to Ancient Chinese NER via Word-Word Relation Classification and Model Ensembling
tian Xia, yilin wang, xinkai Wang, Qun Zhao and Menghui Yang

When Less Is More: Logits-Constrained Framework with RoBERTa for Ancient Chinese NER
Wenjie Hua and Shenghan Xu

Multi-Strategy Named Entity Recognition System for Ancient Chinese
Wenxuan Dong and Meiling Liu

Multi-Domain Ancient Chinese Named Entity Recognition Based on Attention-Enhanced Pre-trained Language Model
Qi Zhang, Zhiya Duan, Shijie Ma, ShengYu Liu, Zibo Yuan and RuiMin Ma

15:30–15:40 Break

15:40–16:20 EvaCun

EvaCun 2025 Shared Task: Lemmatization and Token Prediction in Akkadian and Sumerian using LLMs
Shai Gordin, Aleksi Sahala, Shahar Spencer and Stav Klein

Lemmatization of Cuneiform Languages Using the ByT5 Model
Pengxiu Lu, Yonglong Huang, Jing Xu, Minxuan Feng and Chao Xu

Finetuning LLMs for EvaCun 2025 token prediction shared task
Josef Jon and Ondřej Bojar

Beyond Base Predictors: Using LLMs to Resolve Ambiguities in Akkadian Lemmatization
Frederick Riemenschneider

A Low-Shot Prompting Approach to Lemmatization in the EvaCun 2025 Shared Task
John Sbur, Brandi Wilkins, Elizabeth Paul and Yudong Liu

16:20–17:35 Oral Reports

From Clay to Code: Transforming Hittite Texts for Machine Learning
Emma Yavasan and Shai Gordin

Towards Ancient Meroitic Decipherment: A Computational Approach
Joshua N. Otten and Antonios Anastasopoulos

Detecting Honkadori based on Waka Embeddings
Hayato Ogawa, Kaito Horio and Daisuke Kawahara

Incorporating Lexicon-Aligned Prompting in Large Language Model for Tangut–Chinese Translation
Yuxi Zheng and Jingsong Yu

ParsiPy: NLP Toolkit for Historical Persian Texts in Python
Farhan Farsi, Parnian Fazel, Sepand Haghighi, Sadra Sabouri, Farzaneh Goshtasb, Nadia Hajipour, Ehsaneddin Asgari and Hossein Sameti

17:35 Closing