

RANLP-ALP 2023

**Proceedings of the
Ancient Language Processing Workshop**

associated with

**The 14th International Conference on
Recent Advances in Natural Language Processing
RANLP 2023**

8 September, 2023
Varna, Bulgaria

ANCIENT LANGUAGE PROCESSING WORKSHOP
ASSOCIATED WITH THE 14TH INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING (RANLP 2023)

PROCEEDINGS

8 September 2023
Varna, Bulgaria

ISBN 978-954-452-087-8

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

The First Workshop on Ancient Language Processing (ALP-2023) is co-located with the fourteenth edition of Recent Advance in Natural Language Processing (RANLP) in Varna, Bulgaria on 8 September, 2023.

Ancient languages function as stores of the historical and cultural legacy of humanity. In recent years, significant advancements have been made through the utilization of language technologies in the analysis and interpretation of archaic languages. The purpose of the workshop is to establish a reputable platform for both scholars and practitioners, facilitating the exchange of their most recent research findings and fostering meaningful discussions.

The workshop has received thirty-six submissions covering a wide variety of ancient languages, including Ancient Chinese, Ancient Tibetan, Ancient Greek, Latin, Etruscan, Akkadian, Sumerian, Ancient Syriac, Ancient Hebrew, Basquenglish, Classical Arabic, Meroitic, Middle High German, Pali, and Sanskrit. Among these, Latin, Greek, ancient Chinese, and Sumerian are prominently featured. Within the realm of natural language processing, this collection of papers employs techniques that encompass a diverse range of methodologies, spanning from qualitative analyses and corpus construction strategies to sophisticated machine learning algorithms.

We have accepted sixteen papers for oral presentations and nine for poster presentations. The topics of the accepted submissions include: morphological analysis, POS-tagging and lemmatization, parsing, evaluation of LLMs, text annotation, corpus construction, distributional semantic models, emotion recognition, machine translation, corrupted text correction, hand-written text recognition, intertextual identification, stylistic analysis, named entity recognition, input method, and NLP pipeline systems. The quality resonating through these submissions is genuinely commendable, highlighting the excellence that characterizes the content of this workshop.

The field of ancient language processing is witnessing an expanding research community, driven by the increasing availability of ancient language resources and the growing interest of scholars with machine learning expertise in this domain. The strong turnout in the inaugural year of the ALP workshop serves as a testament to this trend. Hence, we are confident that this event will continue to thrive, fostering an environment for dynamic discussions and interdisciplinary collaborations in both the research and applications of this domain.

We extend our gratitude to the members of the Program Committee for their exhaustive reviews, including the acceptance of additional reviews. We are also appreciative of the RANLP conference chairs, whose tremendous and timely assistance proved invaluable. Last but not the least, we would like to extend our thanks to the student volunteers: Bolin Chang, Zhixiao Zhao, Yutong Zhang, Yixuan Zhang, Kaixin Yin, Feng Xie, and Zhixing Xu.

The ALP-2023 Organizers

Adam Anderson, UC Berkeley, USA

Shai Gordin, Ariel University, Israel

Stav Klein, Tel Aviv University, Israel

Bin Li, Nanjing Normal University, China

Yudong Liu, Western Washington University, USA

Marco C. Passarotti, Università Cattolica del Sacro Cuore, Italy

Organizers:

Adam Anderson (UC Berkeley, USA)
Shai Gordin (Ariel University, Israel)
Stav Klein (Tel Aviv University, Israel)
Bin Li (Nanjing Normal University, China)
Yudong Liu (Western Washington University, USA)
Marco C. Passarotti (Università Cattolica del Sacro Cuore, Italy)

Programme Committee:

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

Invited Speakers:

Dr. Chu-ren Huang (The HongKong Polytechnic University, China)
Dr. Thea Sommerschild (Ca' Foscari University of Venice, Italy)
Dr. Gabriel Stanovsky (Hebrew University of Jerusalem, Israel)

Table of Contents

<i>Training and Evaluation of Named Entity Recognition Models for Classical Latin</i> Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys and Margherita Fantoli	1
<i>Sentence Embedding Models for Ancient Greek Using Multilingual Knowledge Distillation</i> Kevin Krahn, Derrick Tate and Andrew C. Lamicela	13
<i>A Transformer-based parser for Syriac morphology</i> Martijn Naaijer, Constantijn Sikkels, Mathias Coeckelbergs, Jisk Attema and Willem Th. Van Peursen	23
<i>Graecia capta ferum victorem cepit. Detecting Latin Allusions to Ancient Greek Literature</i> Frederick Riemenschneider and Anette Frank	30
<i>Larth: Dataset and Machine Translation for Etruscan</i> Gianluca Vico and Gerasimos Spanakis	39
<i>Evaluation of Distributional Semantic Models of Ancient Greek: Preliminary Results and a Road Map for Future Work</i> Silvia Stopponi, Nilo Pedrazzini, Saskia Peels, Barbara McGillivray and Malvina Nissim	49
<i>Latin Morphology through the Centuries: Ensuring Consistency for Better Language Processing</i> Federica Gamba and Daniel Zeman	59
<i>Cross-Lingual Constituency Parsing for Middle High German: A Delexicalized Approach</i> Ercong Nie, Helmut Schmid and Hinrich Schütze	68
<i>Can Large Language Model Comprehend Ancient Chinese? A Preliminary Test on ACLUE</i> Yixuan Zhang and Haonan Li	80
<i>Unveiling Emotional Landscapes in Plautus and Terentius Comedies: A Computational Approach for Qualitative Analysis</i> Davide Picca and Caroline Richard	88
<i>Morphological and Semantic Evaluation of Ancient Chinese Machine Translation</i> Kai Jin, Dan Zhao and Wuying Liu	96
<i>A tailored Handwritten-Text-Recognition System for Medieval Latin</i> Philipp Koch, Gilary Vera Nuñez, Esteban Garces Arias, Christian Heumann, Matthias Schöffel, Alexander Häberlin and Matthias Assenmacher	103
<i>Evaluating Existing Lemmatisers on Unedited Byzantine Greek Poetry</i> Colin Swaelens, Ilse De Vos and Els Lefever	111
<i>Vector Based Stylistic Analysis on Ancient Chinese Books: Take the Three Commentaries on the Spring and Autumn Annals as an Example</i> Yue Qi, Liu Liu, Bin Li and Dongbo Wang	117
<i>A Joint Model of Automatic Word Segmentation and Part-Of-Speech Tagging for Ancient Classical Texts Based on Radicals</i> Bolin Chang, Yiguo Yuan, Bin Li, Zhixing Xu, Minxuan Feng and Dongbo Wang	122

<i>Introducing an Open Source Library for Sumerian Text Analysis</i> Hansel Guzman-Soto and Yudong Liu	133
<i>Coding Design of Oracle Bone Inscriptions Input Method Based on "ZhongHuaZiKu" Database</i> Dongxin Hu	138
<i>Word Sense Disambiguation for Ancient Greek: Sourcing a training corpus through translation alignment</i> Alek Keersmaekers, Wouter Mercelis and Toon Van Hal	148
<i>Enhancing State-of-the-Art NLP Models for Classical Arabic</i> Tariq Yousef, Lisa Mischer, Hamid Reza Hakimi and Maxim Romanov	160
<i>Logion: Machine-Learning Based Detection and Correction of Textual Errors in Greek Philology</i> Charlie Cowen-Breen, Creston Brooks, Barbara Graziosi and Johannes Haubold	170
<i>Classical Philology in the Time of AI: Exploring the Potential of Parallel Corpora in Ancient Languages</i> Tariq Yousef, Chiara Palladino and Farnoosh Shamsian	179
<i>Using Word Embeddings for Identifying Emotions Relating to the Body in a Neo-Assyrian Corpus</i> Ellie Bennett and Aleksí Sahala	193
<i>A Neural Pipeline for POS-tagging and Lemmatizing Cuneiform Languages</i> Aleksi Sahala and Krister Lindén	203
<i>Tibetan Dependency Parsing with Graph Convolutional Neural Networks</i> Bo An	213
<i>On the Development of Interlinearized Ancient Literature of Ethnic Minorities: A Case Study of the Interlinearization of Ancient Written Tibetan Literature</i> Congjun Long and Bo An	222

Program

Friday, Sept 8, 2023

9:00–9:10 *Opening Remarks*

9:10–9:40 **Invited Talk** by Chu-Ren Huang: *Processing of a ‘living’ ancient language: Issues and Insights from Chinese*

9:40–10:55 **Oral Session 1:**

9:40–9:55 *Training and Evaluation of Named Entity Recognition Models for Classical Latin*
Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys and Margherita Fantoli

9:55–10:10 *Sentence Embedding Models for Ancient Greek Using Multilingual Knowledge Distillation*
Kevin Krahn, Derrick Tate and Andrew C. Lamicela

10:10–10:25 *A Transformer-based parser for Syriac morphology*
Martijn Naaijer, Constantijn Sikkel, Mathias Coeckelbergs, Jisk Attema and Willem Th. Van Peursen

10:25–10:40 *Graecia capta ferum victorem cepit. Detecting Latin Allusions to Ancient Greek Literature*
Frederick Riemenschneider and Anette Frank

10:40–10:55 *Larth: Dataset and Machine Translation for Etruscan*
Gianluca Vico and Gerasimos Spanakis

Friday, Sept 8, 2023 (continued)

10:55–11:15 Coffee break

11:15–12:00 Oral Session 2:

11:15–11:30 *Evaluation of Distributional Semantic Models of Ancient Greek: Preliminary Results and a Road Map for Future Work*

Silvia Stopponi, Nilo Pedrazzini, Saskia Peels, Barbara McGillivray and Malvina Nissim

11:30–11:45 *Latin Morphology through the Centuries: Ensuring Consistency for Better Language Processing*

Federica Gamba and Daniel Zeman

11:45–12:00 *Cross-Lingual Constituency Parsing for Middle High German: A Delexicalized Approach*

Ercong Nie, Helmut Schmid and Hinrich Schütze

12:00–13:00 Poster Session

Can Large Language Model Comprehend Ancient Chinese? A Preliminary Test on ACLUE

Yixuan Zhang and Haonan Li

Unveiling Emotional Landscapes in Plautus and Terentius Comedies: A Computational Approach for Qualitative Analysis

Davide Picca and Caroline Richard

Morphological and Semantic Evaluation of Ancient Chinese Machine Translation

Kai Jin, Dan Zhao and Wuying Liu

A tailored Handwritten-Text-Recognition System for Medieval Latin

Philipp Koch, Gily Vera Nuñez, Esteban Garces Arias, Christian Heumann, Matthias Schöffel, Alexander Häberlin and Matthias Assenmacher

Evaluating Existing Lemmatisers on Unedited Byzantine Greek Poetry

Colin Swaelens, Ilse De Vos and Els Lefever

Vector Based Stylistic Analysis on Ancient Chinese Books: Take the Three Commentaries on the Spring and Autumn Annals as an Example

Yue Qi, Liu Liu, Bin Li and Dongbo Wang

Friday, Sept 8, 2023 (continued)

A Joint Model of Automatic Word Segmentation and Part-Of-Speech Tagging for Ancient Classical Texts Based on Radicals

Bolin Chang, Yiguo Yuan, Bin Li, Zhixing Xu, Minxuan Feng and Dongbo Wang

Introducing an Open Source Library for Sumerian Text Analysis

Hansel Guzman-Soto and Yudong Liu

Coding Design of Oracle Bone Inscriptions Input Method Based on “ZhongHuaZiKu” Database

Dongxin Hu

13:00–14:30 Lunch break

14:30–15:00 **Invited Talk** by Thea Sommerschild: *When the past meets the future at Odessus*

15:00–16:00 Oral Session 3:

15:00–15:15 *Word Sense Disambiguation for Ancient Greek: Sourcing a training corpus through translation alignment*

Alek Keersmaekers, Wouter Mercelis and Toon Van Hal

15:15–15:30 *Enhancing State-of-the-Art NLP Models for Classical Arabic*

Tariq Yousef, Lisa Mischer, Hamid Reza Hakimi and Maxim Romanov

15:30–15:45 *Logion: Machine-Learning Based Detection and Correction of Textual Errors in Greek Philology*

Charlie Cowen-Breen, Creston Brooks, Barbara Graziosi and Johannes Haubold

15:45–16:00 *Classical Philology in the Time of AI: Exploring the Potential of Parallel Corpora in Ancient Languages*

Tariq Yousef, Chiara Palladino and Farnoosh Shamsian

Friday, Sept 8, 2023 (continued)

16:00–16:20 Coffee break

16:20–17:20 Oral Session 4:

16:20–16:35 *Using Word Embeddings for Identifying Emotions Relating to the Body in a Neo-Assyrian Corpus*
Ellie Bennett and Aleksí Sahala

16:35–16:50 *A Neural Pipeline for POS-tagging and Lemmatizing Cuneiform Languages*
Aleksi Sahala and Krister Lindén

16:50–17:05 *Tibetan Dependency Parsing with Graph Convolutional Neural Networks*
Bo An

17:05–17:20 *On the Development of Interlinearized Ancient Literature of Ethnic Minorities: A Case Study of the Interlinearization of Ancient Written Tibetan Literature*
Congjun Long and Bo An

17:20–17:50 **Invited Talk** by Gabriel Stanovsky: *Harnessing Multilingual Models for Ancient Language Processing*

17:50–18:00 Open Discussion and Closing