Historical Languages and AI 2026

Preliminary Book of Abstracts

Program:

Thursday, March 5, 2026

Time	Format
9.00 - 9.30	Arrival & first coffee
9.30 - 10.00	Opening & The Daidalos project
10.00 - 11.00	**Panel 1: Datasets & Data Augmentation I** 1. **ALMeD: Medieval French Gold-Standard** (Ragini Menon & Sabine Tittel, Heidelberg, 20+10min) 2. **Unified Old Church Slavonic Dataset** (Anastasija Margolina, Belgrad, 20+10min)
11.00 - 11.30	Coffee break
11.30 - 12.45	**Panel 2: Datasets & Data Augmentation II** 1. **Data Augmentation Strategies for Premodern Greek LM** (Jacob Murel et al., Princeton, 20+10 min) 2. **InviTE Corpus: Annotating Tudor Invectives** (Sophie Spliethoff et al., Bielefeld, 30+15 min)
12.45 - 13.45	Lunch
13.45 - 14.30	**Pitch your Research Idea** 1. **Mobility.AI. Detecting Mobility in Antique Sources** (András Handl, Leuven, 5min) 2. **Introducing the COALA Project** (Barabara McGillivray, London, 5min) 3. **Technical Standards for Latin and Ancient Greek** (Konstantin Schulz, Berlin, 5min)
14.30 - 16.00	**Science Speed Dating**
16.00 - 16.30	Coffee Break
16.30 - 18.00	**Panel 3: LLMs** 1. **Evaluating Latin/Greek Preverbs with LLMs** (Andrea Farina & Michele Ciletti, London, 30+15 min) 2. **LLM for Middle French Translation (online)** (Raphael Rubino et al., Genf, 30+15 min)
	Dinner

Friday, March 6 2026

Time	Format
8.30 - 9.00	Arrival & first coffee
9.00 - 11.00	**Panel 4: Specific NLP Tasks** 1. **Supervised Ancient Greek translation alignment** (Alek Keersmaekers, Leuven, 30+15 min) 2. **Topic Modeling of Latin Sermons** (Tam Johnson et al., Stockholm, 30+15 min) 3. **Evaluating Sentence Embeddings for Intertextuality** (Michael Wittweiler et al., Zürich/Konstanz, 20+10 min)
11.00 - 11.30	Coffee Break
11.30 - 13.00	**Workshop 1** **Tailored LLM for Ancient Texts** (Premshay Hermon et al., Jerusalem)
13.00 - 14.00	Lunch
14.00 - 15.30	**Workshop 2** **Logion: Machine Learning for Classical Philology** (Jacob Murel, Princeton, LINK) **Workshop 3** **Finetuning and Low-resource Languages** (Thomas Renkert & Florian Nieser, Heidelberg)
15.30 - 16.00	Coffee Break
16.00 - 17.30	**Panel 5: Applications beyond Research** 1. **Libraries and the Digital Humanities** (Katharina Ost, Düsseldorf, 30+15 min) 2. **A Classical Language AI Query Assistant** (Eleni Bozia et al., Florida, 30+15 min)
17.30 - 18.00	Wrap up & feedback

Panels:

Panel 1: Datasets & Data Augmentation I

ALMeD: Medieval French Gold-Standard (Ragini Menon & Sabine Tittel, Heidelberg, 20+10min)

Full Title: **ALMeD: A domAin-specific goLd-standard for MeDieval French**

All authors: Ragini Menon, Sabine Tittel

Abstract:

Medieval French is one of the Romance vernacular languages with a significant history of transmission and an enormous influence on other vernacular languages of the time.

The sources convey valuable knowledge about all aspects of culture. Studying this language increasingly relies on NLP approaches. There have been numerous efforts to parse medieval French; however, these efforts use corpora comprising legal texts and—primarily—literary texts as their foundation and training data. This means that, with the exception of juridical resources, scientific literature is overlooked.

We present ALMeD, an annotated, semantically rich corpus (work-in-progress) of medical and surgical treatises produced in medieval French, i.e., Old and Middle French including Anglo-Norman. Our semantically disambiguated gold-standard goes beyond existing resources, focusing on medical terminology, and accounts for morphological development, evolving syntactical structures and lack of spelling normalisation in these languages.

Unified Old Church Slavonic Dataset (Anastasija Margolina, Belgrad, 20+10min)

Full Title: **Down the Slavic Memory Lane: A Unified Old Church Slavonic Corpus for Core NLP Tasks**

All authors: Anastasija Margolina

Abstract:

This work introduces a large, unified Old Church Slavonic (OCS) corpus to address the under-representation of OCS texts in a standardized digital format. The openly licensed corpus merges 45 heterogeneous sources into a single dataset containing 256,000 text segments, which amounts to 4.4 million tokens. All texts feature standardized Unicode encoding and retain their full diacritics. To demonstrate the corpus's usability for core NLP tasks, a six-class genre classifier was developed by fine-tuning a RuBERT model, achieving a weighted F1 score of 0.8.

The complete corpus, preprocessing scripts, and the fine-tuned model are made publicly available at https://huggingface.co/datasets/BlindSubmission2025/anonymous-ocs-dataset and https://huggingface.co/BlindSubmission2025/anonymous-ocs-model.

Panel 2: Datasets & Data Augmentation II

Data Augmentation Strategies for Premodern Greek LM (Jacob Murel et al., Princeton, 20+10 min)

Full title: **A Comparison of Data Augmentation Strategies for Premodern Greek Language Models**

All authors: Jacob Murel, Sarah Yuan, Barbara Graziosi

Abstract:

This paper compares standard and original data augmentation (DA) strategies for premodern Greek language models (LMs). We also propose an original DA method intended to reflect manuscript variance. We evaluate DA's impact on LM performance for downstream philological tasks: gap-filling, error detection, and correction.

Initial results show DA yields marginal gains, with our variance-based DA performing worse than noisier methods. We suggest this may stems from the degree of variation introduced, highlighting the need to balance grammatical accuracy with textual diversity in classical language LMs.

InviTE Corpus: Annotating Tudor Invectives (Sophie Spliethoff et al., Bielefeld, 30+15 min)

Full title: **The InviTE Corpus: Annotating Invectives in Tudor English Texts for Computational Modeling**

All authors: Sophie Spliethoff, Sanne Hoeken, Silke Schwandt, Sina Zarrieß, Özge Alaçam

Abstract:

In this paper, we aim at the application of Natural Language Processing (NLP) techniques to historical research endeavors, particularly addressing the study of religious invectives in the context of the Protestant Reformation in Tudor England. We outline a workflow spanning from raw data, through pre-processing and data selection, to an iterative annotation process.

As a result, we introduce the InviTE corpus – a corpus of almost 2000 Early Modern English (EModE) sentences, which are enriched with expert annotations regarding invective language throughout 16th-century England. Subsequently, we assess and compare the performance of fine-tuned BERT-based models and zero-shot prompted instruction-tuned large language models (LLMs), which highlights the superiority of models pre-trained on historical data and fine-tuned to invective detection.

Panel 3: LLMs

Evaluating Latin/Greek Preverbs with LLMs (Andrea Farina & Michele Ciletti, London, 30+15 min)

Full title: **Probing Preverbs: Evaluating Large Language Models on Latin and Ancient Greek Preverbed Motion Verbs**

All authors: Andrea Farina, Michele Ciletti

Abstract:

Preverbs, i.e., prefixes that modify verbal bases, play a central role in the semantics of Latin and Ancient Greek. Their meanings range from fully compositional to highly lexicalized, making them an ideal test case for evaluating the semantic capacities of Large Language Models (LLMs). We investigate the ability of 13 LLMs to interpret preverb semantics across a dataset of 2,834 manually annotated preverbed motion verbs, under zero-, one-, two-, and five-shot prompting conditions.

We find that LLMs perform moderately well overall, with GPT-5 achieving the highest F1 of 0.629. Performance improves with increased context examples and is generally higher for compositional, non-lexicalized preverbs, Ancient Greek, and earlier historical periods. Qualitative evaluation highlights systematic differences in consistency, sensitivity to lexicalization, and handling of polysemy, revealing that reasoning-enabled and larger models tend to generate more accurate and internally coherent interpretations.

These results provide insight into the capacities and limitations of current LLMs for modeling morpheme-level semantics in historical languages, with implications for philology, digital humanities, and the development of linguistically informed AI systems.

LLM for Middle French Translation (Raphael Rubino et al., Geneva, 30+15 min)

Full title: **Prompting Large Language Model for 16th Century Middle French Text Normalization and Modernization**

All authors: Raphael Rubino, Mathilde Fontanet, Sandra Coram-Mekkey, Christophe Chazalon, Pierrette Bouillon

Abstract:

This paper presents a study on 16th century Middle French text normalization and modernization through pretrained large language model (LLM) prompting. The modernization process is decomposed in several steps performed by domain experts, leading to four textual variants manually derived from the original text. We explore the use of several few-shot sampling techniques for LLM prompting and compare them to full model fine-tuning, showing a trade-off between computation time and performance on the downstream tasks, with similar hardware specifications.

The evaluation, in terms of automatic metrics, shows that the normalization task leads to lower error rates compared to modernization. All sampling methods outperform random selection of examples while surface-based sampling methods outperform the embedding-based approaches tested in our study for n-shot selection.

Panel 4: Specific NLP Tasks

Supervised Ancient Greek translation alignment (Alek Keersmaekers, Leuven, 30+15 min)

Full title: **Lexically-oriented word alignment for Ancient Greek: a learning-to-rank approach**

All authors: Alek Keersmaekers

Abstract:

This paper introduces a new approach to translation alignment, called lexically-oriented word alignment, which aligns words based on their lexical content.

This approach makes translation alignment more flexible for specific purposes, such as automatically creating annotated datasets and bilingual dictionaries. It develops a new method to perform lexically-oriented word alignment between Ancient Greek and English, based on learning-to-rank supervised machine learning and automatic phrase detection. This approach is able to outperform earlier, unsupervised techniques for Ancient Greek-English word alignment, in particular for content words.

Topic Modeling of Latin Sermons (Tam Johnson et al., Stockholm, 30+15 min)

Full title: **Topic Modeling of Latin Sermons**
All authors: Tam Johnson, Jacob Langeloh, Beáta Megyesi

Abstract:

When examining the works of influential thinkers and theologians from the late medieval period, scholarship has often shown a curious reluctance to engage with the sometimes substantial collections of sermons. Can the latest advances in Natural Language Processing (NLP), and particularly in topic modeling, help to access the contents of these vital sources, in order to chart them more efficiently and understand them better?

This study investigates the viability of using an unsupervised clustering technique to discover the various themes underlying and connecting these historical documents. Using BERTopic as the analytical tool, we examine parameter and data preprocessing alternatives and assess the impact on semantic quality of the generated topics by using standard quantitative metrics paired with a qualitative evaluation based on expert human judgments.

The best performing model achieved scores -0.031 for *topic coherence* and

.840 for *topic diversity*, while the expert human evaluator correctly identified 55% of intruder words in a word intrusion task, suggesting that automated unsupervised methods such as BERTopic could be useful for historians.

Evaluating Sentence Embeddings for Intertextuality (Michael Wittweiler et al., Zürich/Konstanz, 20+10 min)

Full title: **Context Matters: Probing the Robustness of Sentence Embeddings for Intertextuality Detection in Latin Text**

All authors: Michael Wittweiler, Marie Revellio, Julian Schelb

Abstract:

This paper explores the use of sentence embeddings for detecting intertextual references and paraphrased citations in Latin texts. We fine-tune a sentence transformer (SPhilBERTa) on a curated dataset of 544 intertextual pairs from Jerome and Lactantius, paired with candidate sentences from Virgil and other classical authors. Our analysis includes controlled modifications of target sentences to test robustness against inflectional variation, paraphrasing, and contextual changes.

Results show that embeddings retrieve longer verbatim quotations with high accuracy and are tolerant of morphological variation and paraphrase. However, performance decreases for short allusions and when surrounding semantic contexts diverge strongly. We find that context plays a decisive role: condensing target sentences to citation material can improve retrieval, but additional contextual cues can also strengthen similarity. We conclude that combining embedding-based and n-gram methods may yield broader coverage of intertextuality in Latin, while philological interpretation remains essential.

Panel 5: Applications beyond Research

Libraries and the Digital Humanities (Katharina Ost, Düsseldorf, 30+15 min)

Full title: **Data Holdings of Academic Libraries as an Opportunity for the Digital Humanities?**

All authors: Katharina Ost

Abstract:

The paper outlines the services offered by academic libraries in the digital humanities based on four roles: as data providers, networkers, trainers, and developers.

A practical example illustrates the preparation of a digitized incunable for NLP-workflows. The case study shows that there is a considerable amount of curatorial and technical effort involved between making available page scans and achieving genuine DH reusability.

A Classical Language AI Query Assistant (Eleni Bozia et al., Florida, 30+15 min)

Full title: **Automating Classical Language Philology: The Classical Language AI Query Assistant**

All authors: Audrey Barber, Thomas Cerniglia, Eleni Bozia

Abstract:

In an effort to make classical research more accessible, the Data-Driven Humanities Research Group has developed an AI assistant focused on computational philology. Historically, researchers have manually parsed Latin and Greek texts to translate and reference ancient sources. The Classical Language AI Query Assistant aims to streamline this process by utilizing an automated querying system. Using ChatGPT-4-based research, this model was developed to answer specific grammatical and syntactic questions. School and higher-education instructors can use this assistant to enhance traditional pedagogical approaches. Additionally, scholars of other disciplines who wish to learn ancient Greek and Latin to increase their research sources can have easier access to AI-assisted language tools.

Leveraging such modern computational methods opens up several new avenues for teaching and research in classical studies.

Workshops:

Workshop 1: **Tailored LLM for Ancient Texts** (Premshay Hermon et al., Jerusalem)

Full Title: **First Steps towards a Tailored LLM for Ancient Scientific Texts (and beyond)**

Organisers: Premshay Hermon, Orly Lewis, Gideon Manelis, Gabriele Torcoletti

Beginner Level: No technical skills or knowledge is required of participants.

Requirements: Please bring your own laptop.

The workshop serves as an initial collaborative step toward developing domain-specific small LLMs for classical scholarship.

Large language models (LLMs) are advanced AI systems trained on extensive text corpora to predict and generate language by learning patterns of words, concepts, and relationships. They create internal representations that support flexible reasoning and text generation. Fine-tuning these models with carefully curated question-answer pairs from specific domains such as ancient texts allows them to specialize in recognizing historical languages, terminology, and scholarly inquiry styles. Small LLMs with fewer parameters can be effectively fine-tuned to perform specific tasks with less computational cost, making them accessible for humanistic research. This process improves their accuracy and relevance by focusing on domain-specific knowledge, producing reliable tools that assist scholars in analysing and interpreting classical literature.

The workshop will focus on a key and early step in building a tailored LLM: creating a suitable dataset for training it. Such a dataset consists of a large number of human-formulated questions and answers related to given data (e.g. an edition, a commentary, or a cluster of editions), which are then used to train the model to perform the required tasks.

For the sake of feasibility and efficiency, we will focus on sources related to ancient medical and philosophical texts in Greek and Latin. However, we perceive this workshop as a case study and model for broader LLMs for all Greco-Roman literature, or for other historical languages and scholarly fields relying on them.

At the core of the workshop stand the questions and answers that participants will compos concerning the ancient texts. We will provide templates and guidelines for writing suitable questions and answers for training the LLM. Questions will pertain to content, style and terminology of ancient texts in Latin and ancient Greek. While the emphasis will be investigating these broad themes in ancient scientific and philosophical texts, the workshop will be open to those interested in ancient texts from any domain. Direct access to the ancient texts is not strictly necessary, but participants may want to consult some texts to fine-tune their questions and answers.

The organisers will provide key ancient sources – Galen, Hippocrates, Aristotle, Cicero, Lucretius, Seneca – in a searchable format, and participants may bring their own files for other texts.

Workshop 2: **Logion: Machine Learning for Classical Philology** (Jacob Murel, Princeton)

Full Title: **Logion: Machine Learning for Classical Philology**

Organisers: Jacob Murel

Beginner Level: Suitable for participants with no technical background.

Requirements: Please bring your own laptop for which you have permissions to download software, as well as a Greek or Latin text you are interested in editing.

This workshop introduces classics scholars and philologists to Logion. Logion is a no-code application that allows researchers to leverage language models to provide concrete suggestions and inspiration for common philological tasks, namely gap-filling and error detection/correction, for Latin and Greek.

The workshop begins with a 15 minute presentation about the Logion project, followed by a 10-15 minute demo of the Logion application. Participants will then be provided a walkthrough on downloading and installing the application and how to use its current features for gap-filling and error detection/correction and access software documentation. The rest of the time will be devoted for participants to explore the software and ask questions. Participants are encouraged to come with a Greek or Latin text on which they are working or interested in editing in order to explore how Logion may assist them in editing and compiling emendations. The final 20 minutes of the workshop will be devoted to discussion on in-development features and for participants to provide feedback on how the software may be improved.

Outcomes for the workshop are three-fold: participants will have an 1) opportunity to test LMs for their own philological work in a user-friendly environment without prior knowledge of code, 2) provide feedback on how the application may be improved to better assist their research, 3) foster ongoing collaborations between the Logion project and international philologists compiling critical editions of premodern texts.

Workshop 3: **Finetuning and Low-resource Languages** (Thomas Renkert & Florian Nieser, Heidelberg)

Full Title: **How to train your own LLM on historical and low-resource languages — a hands-on workshop**

Organisers: Thomas Renkert, Florian Nieser

Advanced Level:
Requirements: Base level knowledge of

1. how LLMs are structured and what fine-tuning is.
2. Jupyter Notebooks, Python, and the Linux terminal.

Deeper programming skills are not necessary.

Other requirements: Please bring your own laptop.

In this workshop, participants will learn how to train large language models on historical and low-resource languages, based on our experiences within the ParzivAI project. To

facilitate reproducible and shareable results, we will focus on the usage of open source AI models in this workshop. Participants will gain insights into the full pipeline from creating their own datasets, to fine tuning, evaluating and using their own models in user-friendly ways. We will give an outlook on how to use historical chatbots in academic teaching.

The aim of the workshop will be 1) to give a look into how datasets are created, how data is harvested, formatted and structured as well as how different pipelines for dataset creation and testing need to be organized. 2) We want to give a look into training techniques for instruction tuning (non-reasoning and reasoning), and what to consider regarding low-resource languages. 3) Furthermore, we look into different ways of using the trained LLMs (RAG, streamlit frontend, Openwebui). 4) The whole workshop will be structured as an open format with hand-on-parts, discussions, where we will show live examples of simple finetuning-tasks and testing the trained LLM.

The workshop will demonstrate methods and pipelines of facilitating datasets for low-resource languages especially regarding Middle High German as well as demonstrate training approaches and the process of choosing the correct LLM for the given task. At the end, the participants will be able to create their own finetuning pipelines and apply the concepts of the workshops to other historical languages and their own research projects.