LaTeCH-CLfL 2025

9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature

Proceedings of the Workshop

May 4, 2025

The LaTeCH-CLfL organizers gratefully acknowledge the support from the following sponsors.

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 979-8-89176-241-1

Introduction

Welcome to the 2025 edition of LaTeCH-CLfL! Whether you are coming back or joining us for the first time, we are delighted to have you here. This workshop, with a history of nearly two decades, continues to serve as home for a wide spectrum of discussions. This year is no exception, with a lineup of topics that span the intersection of language technology, computational linguistics and the broadly conceived humanities.

This year, in line with the general trend in computational linguistics, we see a central focus on using large language models, with innovative approaches to literary analysis and cultural studies. Papers in this area include evaluating LLM-prompting for sequence labeling in computational literary studies, using LLMs for detecting linguistic variation in Russian media, and exploring zero-shot learning for named entity recognition in historical texts. These contributions demonstrate adaptations of cutting-edge AI technologies to address classic questions in sociolinguistics and in the Humanities.

Historical language processing remains a central area of research, with papers addressing the challenges of working with historical texts and low-resource languages. Contributions in this category include matching entries in historical Swedish encyclopedias, preserving Comorian linguistic heritage through bidirectional transliteration, recovering Egyptian hieroglyphs with next-word prediction language models, and adapting multilingual embedding models to historical Luxembourgish. These papers represent the ongoing effort to extend computational methods to underrepresented languages and historical documents.

Sociopolitical text analysis has also grown in importance, with several papers examining prominent social topics such as bias, propaganda and hate speech. These include works on automated media bias detection, unveiling propagandistic strategies during the Russo-Ukrainian War, detecting gender bias in lyrics, and improving hate speech classification through cross-taxonomy dataset integration. These contributions utilize computational linguistics to observe symptoms of social issues, but also help enhance our understanding of how language shapes public discourse. This year's edition also features more innovative approaches that move beyond the classic context of sociolinguistic, such as quantitative approaches to psychological modeling, conversational AI interviewing techniques, and studies on smalltalk identification in natural conversations that reveal both psychological and social dynamics.

Finally, the computational analysis of literary texts remains a fascinating frontier. This year's papers tackle high-level topics such as scene segmentation in literary texts, relationships in fiction, poetry generation, and the dynamics of the canon – using quantitative and cutting-edge perspectives to model complex literary dynamics.

Overall, we keep seeing the growing convergence of large-scale quantitative models with deep scholarly traditions, creating a frame where cutting edge technology broadens our understanding of human language and (human, for now) culture.

There is something for everyone, all things considered. But do keep an open mind and read all papers, if you have the time. You will be glad you did.

Do not forget to visit our Web site HERE – and check out past workshops too.

It goes without saying that whatever success our workshop enjoys is due to the authors (thank you for staying with us or for trusting us the first time), and without question to the reviewers. A special shout-out to our wonderful program committee!

Yuri, Stefania, Anna, Janis, Diego, Stan

Program Committee

Chairs

Diego Alves, Saarland University
Yuri Bizzoni, Aarhus University
Stefania Degaetano-Ortlieb, Saarland University
Anna Kazantseva, National Research Council Canada
Janis Pagel, Department of Digital Humanities, University of Cologne
Stan Szpakowicz, EECS, University of Ottawa

Program Committee

Jinyeong Bak, Sungkyunkwan University
Johanna Binnewitt, Federal Institute for Vocational Education and Training
Patrick Brookshire, Academy of Sciences and Literature | Mainz
Paul Buitelaar, University of Galway
Miriam Butt, University of Konstanz
Prajit Dhar, University of Groningen
Jacob Eisenstein, Google
Anna Feldman, Montclair State University
Mark Finlayson, FIU
Francesca Frontini, Istituto di Linguistica Computazionale A. Zampolli"- ILC Consiglio Nazionale delle Ricerche - CNR
Serge Heiden, ENS de Lyon
Rebecca Hicke, Cornell University
Labiba Jahan, Southern Methodist University
Dimitrios Kokkinakis, University of Gothenburg
Stasinos Konstantopoulos, NCSR Demokritos
Maria Kunilovskaya, Saarland University
John Ladd, Washington & Jefferson College
John Lee, City University of Hong Kong
Chaya Liebeskind, Jerusalem College of Technology , Lev Academic Center
Thomas Lippincott, Johns Hopkins University
Barbara Mcgillivray, King's College London
Cara Messina, Marist University
Craig Messner, Johns Hopkins University
David Mimno, Cornell University
Vivi Nastase, University of Geneva
Borja Navarro-Colorado, University of Alicante
Pierre Nugues, Lund University
Thijs Ossenkoppele, University of Amsterdam
Andrew Piper, McGill University
Petr Plechac, Institute of Czech Literature CAS
Thierry Poibeau, LATTICE (CNRS & ENS/PSL)
Jelena Prokic, Leiden University
Georg Rehm, DFKI
Nils Reiter, University of Cologne
Pablo Ruiz Fabo, LiLPa, Université de Strasbourg
Marijn Schraagen, Utrecht University

Artjoms Sela, Institute of Polish Language (PAN)
Hale Sirin, Johns Hopkins University
Pia Sommerauer, Vrije Universiteit Amsterdam
Elke Teich, Universität des Saarlandes
Laure Thompson, Princeton University
Ulrich Tiedau, University College London
Ted Underwood, University of Illinois
Menno Van Zaanen, South African Centre for Digital Language Resources
Lorella Viola, Vrije Universiteit Amsterdam
Rob Voigt, Northwestern University
Sophie Wu, McGill University
Albin Zehe, University of Wuerzburg
Heike Zinsmeister, Universitaet Hamburg

Keynote Talk

Computational Humanities as Cultural Seismography

Tom Lippincott
Johns Hopkins University

Abstract: How do we move between machine learning and humanistic inquiry without losing our balance? There's no single right answer, but in this talk I'll enumerate a handful of principles that have emerged as useful guidelines for my group, and how they connect to several ongoing projects in computational cultural studies. These principles include a strong dispreference for pretrained LLMs, an emphasis on deep cross-training, and research considerations closely tied to cognitive science. Beyond the specifics, I hope the talk will be a useful example for junior researchers who are beginning to characterize their own agenda and communicate with potential stakeholders across engineering and the humanities.

Bio:

We are delighted to welcome **Tom Lippincott** as our invited speaker at the LaTeCH-CLfL workshop. Tom is an Associate Research Professor at Johns Hopkins University, where he also serves as Director of Digital Humanities with a primary appointment in the Alexander Grass Humanities Institute. His work bridges the gap between machine learning and the humanities, bringing advanced computational techniques—particularly deep neural architectures—into dialogue with scholarship in literature, history, and archaeology.

Tom holds secondary appointments in the Department of Computer Science and the Center for Language and Speech Processing, and the Data Science and AI Institute. Before joining Johns Hopkins, he was research faculty at Columbia University's Center for Computational Learning Systems, following doctoral work at the University of Cambridge and undergraduate studies in Philosophy and Computer Science at the University of Chicago.

His current research focuses on the development of machine learning models, tools, and practices that can reinforce, expand, or challenge received understanding of human culture activities. He has published influential work on authorship attribution and stylistic analysis, including computational investigations into the Pauline epistles and the Documentary Hypothesis of the Hebrew Bible. Earlier in his career, Tom contributed to unsupervised learning of morphology and syntax, including work that received a Best Paper award at COLING 2016.

In addition to his work on Bayesian modeling and domain variation in scientific literature, Tom has also made significant contributions to social media analysis, language identification, and the development of resources for low-resource languages.

With his deep interdisciplinary expertise and commitment to building bridges between computational methods and humanistic inquiry, Tom brings a unique perspective to our workshop.

Table of Contents

Program

Sunday, May 4, 2025

08:30 - 09:50 *Talks I*

Evaluating LLM-Prompting for Sequence Labeling Tasks in Computational Literary Studies
Axel Pichler, Janis Pagel and Nils Reiter

Prompting the Past: Exploring Zero-Shot Learning for Named Entity Recognition in Historical Texts Using Prompt-Answering LLMs
Crina Tudor, Beata Megyesi and Robert Östling

Generation of Russian Poetry of Different Genres and Styles Using Neural Networks with Character-Level Tokenization
Ilya Koziev and Alena Fenogenova

Why Novels (Don't) Break Through: Dynamics of Canonicity in the Danish Modern Breakthrough (1870-1900)
Alie Lassche, Pascale Feldkamp, Yuri Bizzoni, Katrine Baunvig and Kristoffer Nielbo

09:50 - 10:30 *Poster teasers for online posters and Q&A*

10:30 - 11:00 *Coffee Break*

11:00 - 12:00 *Invited Talk by Tom Lippincott: 'Computational Humanities as Cultural Seismography'*

12:00 - 13:30 *Lunch*

13:30 - 14:50 *Poster session on site*

14:50 - 15:30 *Talks II*

Embedded Personalities: Word Embeddings and the Big FivePersonality Model
Oliver Müller and Stefania Degaetano-Ortlieb

Prompting the Past: Exploring Zero-Shot Learning for Named Entity Recognition in Historical Texts Using Prompt-Answering LLMs
Crina Tudor, Beata Megyesi and Robert Östling

Sunday, May 4, 2025 (continued)

15:30 - 16:00 *Coffee Break*

16:00 - 17:00 *Talks III*

'... like a needle in a haystack": Annotation and Classification of Comparative Statements
Pritha Majumdar, Franziska Pannach, Arianna Graciotti and Johan Bos

LLMs for Translation: Historical, Low-Resourced Languages and Contemporary AI Models
Merve Tekgürler

Matching and Linking Entries in Historical Swedish Encyclopedias
Simon Börjesson, Erik Ersmark and Pierre Nugues

17:00 - 17:30 *Conclusion and SIGHUM Business Meeting*