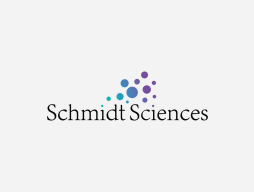
SIGHUM 2026

10th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature

Proceedings of the Workshop

March 28-29, 2026

The SIGHUM 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 979-8-89176-373-9

Introduction

Welcome to the 2026 edition of LaTeCH-CLfL!

Whether you are a seasoned participant or attending for the first time, thank you for gathering here. With twenty years of history, our workshop remains a forum for diverse and evolving conversations, and brings together work at the intersection of language technology, computational linguistics, and the Humanities in their widest sense.

This year, we are proud to present a collection of studies that goes beyond the known restrictions of Anglocentric research, and that spans a wide array of different languages. Our collection features several cutting edge works on under-resourced languages: automatic modelling of Armenian vocabulary and Armenian inscriptions, Greek Poetry rhyme generation, a quantitative study of Romanian writings, an analysis of Finnish refugees' interviews, a work on indirect speech in historical Danish and Norwegian corpora, a work on the infamous problem of reported speech in Classical Latin, and even a pipeline for Ottoman-Turkish bibliographies.

High-resource languages that are not English are also present: several works tackle fine-grained linguistic and stylistic problems in both contemporary and historical German, as well as Russian war propaganda and contemporary French literature, and linguistic diversity is central in several brand new resources, such as a cross-lingual corpus of LLM-generated film synopses, a dataset of under-resourced languages for NLP, and a dataset on digital linguistic diversity.

But a rich linguistic diversity is not the only feature of this year's workshop. Our field is still exploring and enriching LLMs' powers of analysis and their relationship with traditional Humanities' problems, applying them to a variety of high complexity tasks, from modelling domain-specific language variation to automatic annotation of Ancient Greek geographical nouns, as well as reflecting on how their development has changed corpus linguistics.

As usual, our workshop also touches several topics relevant for social and historical sciences, such as persuasion simulation, disinformation modeling, the evolution of scientific concepts, entity recognition in historical texts, personality structure retrieval from word embeddings, systems to explore noisy historical data and biographical sources.

Finally, literary concerns remain central, with studies on stylistic transfer, automatic translation of poetry, techniques to quantify translators' interventions, authorship analysis, and of course new advances on LLMs' ability of dealing with full-length novels.

Overall, researchers in our field are trying to expand the reach of computational approaches in the Humanities while keeping their linguistic, cultural, and historical diversity. Whether working with under-resourced traditions, large contemporary corpora, or literary and social-scientific questions, the contributions in this volume sketch a field that is trying to use the neck-breaking evolution of linguistic technologies to illuminate at a new scale human expression in all its complexity.

Any success this workshop enjoys is first and foremost owed to its authors. Thank you for continuing this journey, or for placing your trust in us for the first time! We want to thank our reviewers, whose care and expertise sustain the quality of the discussion. Finally our gratitude goes to the outstanding program committee for their dedication and generosity throughout the process. Enjoy!

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

Sergei Bagdasarov, Saarland University
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
Pierluigi Cassotti, University of Gothenburg
Kent Chang, UC Berkeley
Stefano De Pascale, KU Leuven
Koel Dutta Chowdhury, Saarland Informatics Campus,Saarland University
Mark Finlayson, FIU
Francesca Frontini, Istituto di Linguistica Computazionale A. Zampolli"- ILC Consiglio Nazionale delle Ricerche - CNR
Svenja Guhr, University of California at Berkeley
Matteo Guida, The University of Melbourne
Hans Ole Hatzel, Universität Hamburg
Serge Heiden, ENS de Lyon
Rebecca Hicke, Cornell University
Azwad Anjum Islam, Florida International University
Mia Jacobsen, Aarhus University
Labiba Jahan, Southern Methodist University
Dimitrios Kokkinakis, University of Gothenburg
Stasinos Konstantopoulos, NCSR Demokritos
Maria Kunilovskaya, Saarland University
John Ladd, Washington & Jefferson College
Alie Lassche, Center for Humanities Computing, Aarhus University
Chaya Liebeskind, Jerusalem College of Technology , Lev Academic Center
Thomas Lippincott, Johns Hopkins University
Barbara Mcgillivray, King's College London
Caio Mello, Luxembourg Centre for Contemporary and Digital History
Craig Messner, Johns Hopkins University
David Mimno, Cornell University
Vivi Nastase, University of Geneva
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, UniversitÃ© de Strasbourg
Marijn Schraagen, Utrecht University
Artjoms Sela, Institute of Polish Language (PAN)
Hale Sirin, Johns Hopkins University
Pia Sommerauer, Vrije Universiteit Amsterdam
Julius Steuer, Universität des Saarlandes
Elke Teich, UniversitÃ¤t des Saarlandes
Gaurish Thakkar, University of Zagreb
Ted Underwood, Univ of Illinois
Sowmya Vajjala, National Research Council
Freek Van De Velde, KU Leuven
Marieke Van Erp, KNAW Humanities Cluster
Menno Van Zaanen, South African Centre for Digital Language Resources
Francielle Vargas, University of São Paulo
Lorella Viola, Vrije Universiteit Amsterdam
Albin Zehe, University of Wuerzburg
Naitian Zhou, University of Michigan
Heike Zinsmeister, Universitaet Hamburg

Keynote Talk

Between Precision and Scale: Bridging Computational Methods and Humanistic Inquiry in Historical Semantics

Barbara McGillivray
King's College London

Abstract: How do we scale up the study of meaning change across centuries of texts while preserving the nuanced, culturally grounded interpretations that humanistic scholarship demands? This question sits at the heart of my research on Word Sense Disambiguation (WSD) and semantic change in historical languages, where the challenges are both computational and epistemological.
In this talk, I outline several principles that have emerged as central to my approach. First, systematic quantitative analyses of semantic phenomena require disambiguating polysemous words at scale, moving beyond qualitative observations to identify which specific senses are active in each context. Yet historical and low-resource languages resist standard techniques due to sparse data and shifting semantic boundaries. Second, expert annotation is not merely a preliminary step but the essential foundation that enables computational methods to capture culturally and temporally situated meanings. This human-in-the-loop approach ensures that automated WSD algorithms, once trained, can be deployed across large corpora while maintaining interpretive fidelity.
I illustrate these principles through work on nineteenth-century English, where rapid semantic shifts over decades reveal how industrialisation restructured entire semantic fields, and through Computational Corpus Annotation for Quantitative Analysis of Latin Lexical Semantics (COALA), an ERC-selected project that aims to trace gradual evolution across centuries of Latin textual tradition. These contrasting temporal scales demonstrate how integrating computational scalability with philological depth can transform our understanding of the relationship between linguistic and historical change, and offer a model for Digital Humanities research that neither privileges algorithmic automation nor retreats into purely qualitative analysis.

Bio:

Barbara McGillivray is Senior Lecturer in Digital and Computational Humanities in the Department of Digital Humanities of King's College, where she leads the Computational Humanities research group. She is Principal Investigator of the COALA project, successfully evaluated by the ERC. She is Editor-in-Chief of the Journal of Open Humanities Data and convenor of the MA programme in Digital Humanities at King's. Her research focusses on computational methods for the study of language change in both historical languages and contemporary data. As a Turing research fellow at the University of Cambridge and at The Alan Turing Institute she was also co-Investigator of the Living with machine project. Previously she worked as a language technologist in the Dictionaries division of Oxford University Press and as a data scientist in the Open Research Group of Springer Nature. Her most recent book is "Applying Language Technology in Humanities Research. Design, Application, and the Underlying Logic" (Palgrave Macmillan 2020).

Table of Contents

From Corpus to Concept Scheme: Developing a SKOS Vocabulary for Armenian Epigraphic Heritage
Hamest Tamrazyan, Kamal Nour and Emanuela Boros 1

Armenian AutoEpiDoc: Automated Extraction and Encoding of Armenian Inscriptions into EpiDoc TEI/XML
Hamest Tamrazyan, Emile Cornamusaz and Emanuela Boros 11

Studying Expert-ese: Profiling and Classification of Domain-Specific Language Variation in Architecture with Traditional Machine Learning and LLMs
Carmen Schacht and Renate Delucchi Danhier 16

CroCoSyn: A Cross-Lingual and Cross-Model Corpus of LLM-Generated Film Synopses
Louis Escouflaire 30

Identity Without Action: Rethinking Collective Action Models in Disinformation Research
Lorella Viola 36

Weakly Supervised Named Entity Recognition for Historical Texts
Marco Sorbi, Laurent Moccozet and Stephane Marchand-Maillet 48

Invisible Speakers? Gender Disparity in German AI Discourse and Its Reflection in Language Models
Milena Belosevic 66

GlobLingDiv: A global dataset linking linguistic diversity and digital support to reveal landscapes with under-resourced languages for NLP
Katharina Zeh, Hannes Essfors, Juliane Benson, Lale Tüver, Andreas Baumann and Hannes A. Fellner 80

LLMs Got Rhyme? Hybrid Phonological Filtering for Greek Poetry Rhyme Detection and Generation
Stergios Chatzikyriakidis and Anastasia Natsina 87

Style as Signature: Profile-Based Authorship Verification of Mihai Eminescu's Journalistic Corpus
Ioana-Roxana Boriceanu and Liviu Dinu 102

Measuring Social Integration Through Participation: Categorizing Organizations and Leisure Activities in the Displaced Karelians Interview Archive using LLMs
Joonatan Laato, Veera Schroderus, Jenna Kanerva, Jenni Kauppi, Virpi Lummaa and Filip Ginter 111

Catalogues as Data: Interpretable NLP Pipelines for Ottoman-Turkish Bibliographies
Mark Hill, Ayse Bulus and Paul Spence 128

Stylistic Transfer from Annotator Communities to Large Language Models
Jay Chooi 135

Modeling Changing Scientific Concepts with Complex Networks: A Case Study on the Chemical Revolution
Sofia Aguilar Valdez and Stefania Degaetano-Ortlieb 146

Speaking on Their Behalf: Detecting Indirect Speech in Historical Danish and Norwegian Texts
Ali Al-Laith, Alexander Conroy, Kirstine Degn, Jens Bjerring-Hansen and Daniel Hershcovich 157

Harder than Finding the Lost Sheep? Towards Automatically Suggesting Deliberate Metaphor Annotations in German Sermons
Ronja Laarmann-Quante and Stefanie Dipper 164

Semantic Factor Analysis: Validating Personality Structure Recovery from empirically-mediated Word Embeddings
Oliver Müller 176

Quantitative Analysis of Rhyme and Metre in LLM-generated Translations of Poetry
Jan-Felix Klumpp 189

WikiLingDiv: a dataset for quantifying digital linguistic diversity using Wikipedia page views
Hannes Essfors and Andreas Baumann 202

Modeling Linguistic Imprints of War Propaganda in a Russian Wikipedia Fork: A Comparative Analysis with the Original Wikipedia
Anastasiia Vestel and Stefania Degaetano-Ortlieb 212

Stylometric Approach to AI-generated Texts. An Analysis of Contemporary French-Language Literature
Adam Pawłowski and Tomasz Walkowiak 221

Degree Zero of Translation: Using Interlinear Baselines to Quantify Translator Intervention
Maciej Rapacz and Aleksander Smywiński-Pohl 227

How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG
Donghan Bian, Marie Puren and Florian Cafiero 241

Detecting reported speech as a token classification task: an application to Classical Latin?
Agustin Dei 251

Narrative in Short German Prose: A Multi-Phenomenon Dataset for Computational Literary Analysis
Hans Ole Hatzel, Haimo Stiemer, Evelyn Gius and Chris Biemann 257

Sense-Based Annotation of Geographical Nouns in Ancient Greek and Latin: A Diachronic Study with LLMs
Andrea Farina, Michele Ciletti, Barbara Mcgillivray and Andrea Ballatore 266

Evaluating Humanities Theory Alignment in Large Language Models: Incremental Prompting and Statistical Assessment
Axel Pichler and Janis Pagel 280

Too Long, Didn't Model: Decomposing LLM Long Context Understanding With Novels
Sil Hamilton, Rebecca Hicke, Mia Ferrante, Matthew Wilkens and David Mimno 295

AI Corpus Linguist: More than a Year of Experience
Jiří Milička and Tomáš Machálek 305

Generative Information Extraction from Biographical Sources
Robin Winkle, Manfred Stede and Jörn Kreutel 311

WikiFirst: A Genre-Fixed, Content-controlled Corpus for Evaluating Content Effects in Authorship Analysis
Dung Nguyen, G. Çağatay Sat, Evgeny Pyshkin and John Blake 323

Measuring the Symbolic Power of Languages with LLM-based Multilingual Persuasion Simulation
Yin Jou Huang and Fei Cheng 328

Program

Saturday, March 28, 2026

14:00 - 14:05 *Welcome*

14:05 - 14:35 *Linguistic Diversity & Resources*

GlobLingDiv: A global dataset linking linguistic diversity and digital support to reveal landscapes with under-resourced languages for NLP
Katharina Zeh, Hannes Essfors, Juliane Benson, Lale Tüver, Andreas Baumann and Hannes A. Fellner

14:35 - 15:30 *Poster Teasers*

WikiLingDiv: a dataset for quantifying digital linguistic diversity using Wikipedia page views
Hannes Essfors and Andreas Baumann

Stylistic Transfer from Annotator Communities to Large Language Models
Jay Chooi

LLMs Got Rhyme? Hybrid Phonological Filtering for Greek Poetry Rhyme Detection and Generation
Stergios Chatzikyriakidis and Anastasia Natsina

Invisible Speakers? Gender Disparity in German AI Discourse and Its Reflection in Language Models
Milena Belosevic

How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG
Donghan Bian, Marie Puren and Florian Cafiero

From Corpus to Concept Scheme: Developing a SKOS Vocabulary for Armenian Epigraphic Heritage
Hamest Tamrazyan, Kamal Nour and Emanuela Boros

Armenian AutoEpiDoc: Automated Extraction and Encoding of Armenian Inscriptions into EpiDoc TEI/XML
Hamest Tamrazyan, Emile Cornamusaz and Emanuela Boros

Detecting reported speech as a token classification task: an application to Classical Latin?
Agustin Dei

Saturday, March 28, 2026 (continued)

CroCoSyn: A Cross-Lingual and Cross-Model Corpus of LLM-Generated Film Synopses
Louis Escouflaire

15:30 - 16:15 *Poster Session (GatherTown)*

16:15 - 16:30 *Closing*

Sunday, March 29, 2026

09:00 - 10:30 *Historical, Diachronic & Cultural Language Data*

Weakly Supervised Named Entity Recognition for Historical Texts
Marco Sorbi, Laurent Moccozet and Stephane Marchand-Maillet

Sense-Based Annotation of Geographical Nouns in Ancient Greek and Latin: A Diachronic Study with LLMs
Andrea Farina, Michele Ciletti, Barbara Mcgillivray and Andrea Ballatore

Measuring Social Integration Through Participation: Categorizing Organizations and Leisure Activities in the Displaced Karelians Interview Archive using LLMs
Joonatan Laato, Veera Schroderus, Jenna Kanerva, Jenni Kauppi, Virpi Lummaa and Filip Ginter

10:30 - 11:00 *Coffee Break*

11:00 - 12:00 *Invited Talk by Barbara McGillivray: 'LLMs, diachrony, humanities theory, and methodological perspectives'*

12:00 - 12:30 *Identity*

Identity Without Action: Rethinking Collective Action Models in Disinformation Research
Lorella Viola

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Information Extraction & Theory Alignment*

Generative Information Extraction from Biographical Sources
Robin Winkle, Manfred Stede and Jörn Kreutel

Evaluating Humanities Theory Alignment in Large Language Models: Incremental Prompting and Statistical Assessment
Axel Pichler and Janis Pagel

15:00 - 15:30 *Coffee Break*

Sunday, March 29, 2026 (continued)

15:30 - 16:30 *Poster session*

16:30 - 17:30 *Literary Language*

Narrative in Short German Prose: A Multi-Phenomenon Dataset for Computational Literary Analysis
Hans Ole Hatzel, Haimo Stiemer, Evelyn Gius and Chris Biemann

Quantitative Analysis of Rhyme and Metre in LLM-generated Translations of Poetry
Jan-Felix Klumpp

17:30 - 17:45 *Closing Session & SIGHUM Business Meeting*