NLP4DH 2026

The 6th International Conference on Natural Language Processing for the Digital Humanities

Proceedings of the Conference

July 4, 2026

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 979-8-89176-427-9

Preface

We are delighted to welcome you to the 6th International Conference on Natural Language Processing for the Digital Humanities (NLP4DH 2026)! This year we are co-located alongside the Annual Conference of the Association for Computational Linguistics (ACL 2026) in San Diego, California.

NLP4DH brings together researchers from natural language processing and the humanities around shared questions: how computational tools shape interpretation, how language technologies treat under-represented languages and cultural materials, and what assumptions get baked into the systems we are continuously building. With LLMs now being used by so many people, these questions are especially urgent. Thus the question we posed this year was: looking beyond NLP helping DH, in what ways can DH help NLP?

We are happy to have accepted 36 wonderful papers. They span corpus construction for low-resource and historical languages, evaluation of LLMs against literary and cultural materials, OCR quality and provenance, computational narrative analysis, and methodological reflection on what NLP can tell us about cultural texts.

Alongside the main track, NLP4DH 2026 features a special track: "Reinterpreting NLP," which invited work that examines language-model training corpora, adapts mechanistic interpretability methods to humanities questions, and brings humanistic intuition back into language-technology design. We believe the papers in this proceeding speak to this important question.

We thank our program committee for their reviewing, our invited speakers for joining us, the ACL 2026 organizers for hosting NLP4DH, and the community for keeping the conversation going. We are furthermore thank our invited speakers, Sophie Hao (Boston University) and Luca Soldaini (Allen Institute for AI).

Thank you for joining us!

All best,

The NLP4DH 2026 Program Committee

Organizing Committee

Organizers

Sil Hamilton, Cornell University
Emily Öhman, Waseda University
Rebecca M. M. Hicke, Cornell University
Yuri Bizzoni, Aarhus University
Axel Bax, Cornell University
Jacob A. Matthews, Cornell University
Mika Hämäläinen, Metropolia University of Applied Sciences

Program Committee

Program Chairs

Sil Hamilton, Cornell University
Emily Öhman, Waseda University
Rebecca M. M. Hicke, Cornell University
Yuri Bizzoni, Aarhus University
Axel Bax, Cornell University
Jacob A. Matthews, Cornell University
Mika Hämäläinen, Metropolia University of Applied Sciences

Reviewers

Maria Antoniak, University of Colorado at Boulder
Frederik Arnold, Humboldt Universität Berlin
Mohammed Attia, Google
Won Ik Cho, Samsung Advanced Institute of Technology
Thibault Clérice, INRIA Paris - Almanach
Sourav Das, Indian Institute of Information Technology Kalyani
Stefania Degaetano-Ortlieb, Universität des Saarlandes
Anna Dmitrieva, Pushkin State Russian Language Institute
Quan Duong, University of Helsinki
Sebastian Oliver Eck, University of Oxford
Pascale Feldkamp, School of Communication and Culture, Aarhus University
Sijia Ge, University of Cincinnati
Luke Gessler, Indiana University
Balázs Indig, Eötvös Lorand University
Kenichi Iwatsuki, Mirai Translate
Antti Kanner, University of Turku
Yoshifumi Kawasaki, The University of Tokyo
Katerina Korre, Athena Research and Innovation Centre
Leo Leppänen, University of Helsinki
Noémi Ligeti-Nagy, Hungarian Research Centre for Linguistics
Aatu Liimatta, University of Helsinki
Dongqi Liu, Universität des Saarlandes
Kiara M. H. Liu, Cornell University
Pierre Magistry, Institut National des Langues et Civilisations Orientales
Enrique Manjavacas, Sigma Cognition
Sanghoon Oh, Cornell University
Shu Okabe, Technische Universität München
Hugo Gonçalo Oliveira, Universidade de Coimbra
Jaihyun Park, Indiana University
Niko Tapio Partanen, University of Helsinki
Lidia Pivovarova, University of Helsinki
Aynat Rubinstein, Hebrew University of Jerusalem
Martin Ruskov, University of Milan
Alejandro Sierra Múnera, Hasso Plattner Institute
Youngsook Song, Lablup

Jonne Sälevä, Brandeis University
Jouni Tuominen, University of Helsinki
Hisao USUI, Tokyo University of Agriculture and Technology, Tokyo Institute of Technology
Elissa Nakajima Wickham, Waseda University
Joshua Wilbur, University of Tartu
Sophie Wu, McGill University, McGill University
Shuo Zhang, Tufts University

Table of Contents

Program

Saturday, July 4, 2026

09:00 - 09:30 *Virtual Posters*

From OCR to Analysis: Tracking Correction Provenance in Digital Humanities Pipelines
Haoze Guo and Ziqi Wei

Frequency Accelerates Semantic Change: Evidence from 500 Years of Korean
Cheonkam Jeong and Yeeun Choi

100,000+ Movie Reviews from Kazakhstan: Russian, Kazakh, and Code-Switched Texts
Rustem Yeshpanov

Quantifying Text Reuse Across Three Kṛṣṇa Yajurveda Recensions: Using Multi-Algorithm Computational Collation
So Miyagawa, Kyoko Amano, Yuzuki Tsukagoshi and Yuki Kyogoku

Data Contamination in Neural Hieroglyphic Translation: A Reproducibility Study
Ammar Toutou, Abdelrahman Harb and Christine Basta

Beyond Prompt-Sensitive Emotion Words: Stable Embeddings for Tang Poetry Analysis
Linyue Zhang and Feiyue Li

Measuring Embedding Sensitivity to Authorial Style in French: Comparing Literary Texts with Language Model Rewritings
Benjamin Icard, Lila Sainero, Alice Breton, Evangelia Zve and Jean-Gabriel Ganascia

Evaluating Latin and Ancient Greek Sentence Alignment through Parallel Sentence Mining
Sebastian Reichbauer, Shu Okabe and Alexander Fraser

Beyond Genre Categories: How Narrative Pattern Coherence and Spanning Distance Shape Film Success
Zhichao Wang and ZEYU LYU

Register Mixing Is the Norm on the Web
Erik Henriksson, Alireza Razzaghi, Tuomas Lundberg, Antti Kanner and Veronika Laippala

Saturday, July 4, 2026 (continued)

PHMartialLawNER: A Tagalog Named Entity Recognition Corpus for the Philippine Martial Law Era
Abdiel Clarence Tabuzo, Vladimir Gray Velazco, Cassandra Cabral, Moneah Shaila Lacsam and Charmaine Salvador Ponay

Fluency and Faithfulness in Human and Machine Literary Translation
Sarah Griebel and Ted Underwood

Matching Meaning at Scale: Evaluating Semantic Search for 18th-Century Intellectual History through the Case of Locke
Yu Wu, Ananth Mahadevan, Filip Ginter, Michael Mathioudakis and Mikko Tolonen

Twenty's Plenty: Semantic Scaffolding and Span Architecture for 19-Label NER in Medieval Latin Charters
Tamás Kovács, Giuseppe Consolo and Georg Vogeler

Computational Modeling of Educational Theory in Low-Socioeconomic Contexts
Jadon Swearingen, Mustafa Ocal, Md Tarique Hasan Khan and Labiba Jahan

Never Care For What They Say ? Platform vs Genre Rules in Online Horror Narratives (2007–2024)
Alexandre Lionnet-Rollin and Florian Cafiero

Computational Authorship Attribution in the Children's Tales of Oscar and Constance Wilde: The Case of "The Selfish Giant"
Liviu P Dinu, Alina Iacob and Cosmin Ciotlos

Evaluating Open-Source LLMs for Text Summarization and Named Entity Recognition in Long, Unstructured Text
Pauline Kister and Miriam Schirmer

Perspectives – Interactive Document Clustering for Qualitative Data Analysis
Tim Fischer and Chris Biemann

09:30 - 09:45 *Opening Remarks*

09:45 - 10:15 *Invited Talk: Sophie Hao (Boston University)*

10:30 - 11:30 *Coffee Break*

Saturday, July 4, 2026 (continued)

11:30 - 13:00 *Session 1: Historical and Multilingual Texts*

Scaling Sentence Similarity for Classical Tibetan with Automatic Annotations
Shay Cohen, Jingyi Yang, Gal Rabinovitz, Sonam Choden, Ofir Shtrosberg, Nicola Bajetta, Goody Ben Horin, Rebecca Sundén, Omri Drori, Sonam Jamtsho, Dorji Wangchuk, Kfir Bar, Orna Almogi and Shai Fine

Modeling the "Dalet" Clitic in Historical Hebrew Texts: A New Prefix-Segmented BERT Model and Stylistic Analysis
Rachel Tal, Cheyn Shmuel Shmidman and Avi Shmidman

Tracing Thematic Change in Early English-Language Science Fiction, 1818-1930
Jonathan Gordon

From Traditional Taggers to LLMs: A Comparative Study of POS Tagging for Medieval Romance Languages
Matthias Schöffel and Esteban Garces Arias

Lost in Translation? Exploring the Shift in Grammatical Gender from Latin to Occitan
Ahan Chatterjee, Matthias Schöffel, Matthias Aßenmacher, Marinus Wiedner and Esteban Garces Arias

In Search of Lost Adventure Novels: Supervised Genre Retrieval and Corpus Refinement in Gallica
Jean Barré

13:00 - 14:00 *Lunch Break*

14:00 - 14:30 *Invited Talk: Luca Soldaini (Allen Institute for AI)*

14:30 - 15:30 *Session 2: LLMs and Cultural Analysis*

Directional Alignment and Narrative Agency in Human–LLM Co-Writing
Halfdan Nordahl Fundal and Yuri Bizzoni

MADRAG: Multi-Agent Debate with Retrieval-Augmented Generation for Training-Free Analytic Essay Scoring
Ali Keramati, Shiyuan Zhou, Sharad Mehrotra and Mark Warschauer

Saturday, July 4, 2026 (continued)

Narrative Landscape: Mapping Narrative Dispositions Across LLMs
Donghoon Jung, Jiwoo Choi, Songeun Chae and Seohyon Jung

Between Whispers and Screams: Loudness Standard Deviation as a Proxy for Explicit Content Detection in US Romance Novels
Svenja Guhr

15:30 - 16:30 *Coffee Break*

16:30 - 17:30 *In-Person Posters*

Narrative Landscape: Mapping Narrative Dispositions Across LLMs
Donghoon Jung, Jiwoo Choi, Songeun Chae and Seohyon Jung

Prompting the Past: Linguistic Transformations and Cultural Accuracy in AI-Generated Image Reconstructions for Multivocal Cultural Heritage
Ravini Wimalasuriya, Lea Krause and Gert-Jan Burgers

Temporal Text Classification with Large Language Models
Nishat Raihan and Marcos Zampieri

Modeling the "Dalet" Clitic in Historical Hebrew Texts: A New Prefix-Segmented BERT Model and Stylistic Analysis
Rachel Tal, Cheyn Shmuel Shmidman and Avi Shmidman

Scaling Sentence Similarity for Classical Tibetan with Automatic Annotations
Shay Cohen, Jingyi Yang, Gal Rabinovitz, Sonam Choden, Ofir Shtrosberg, Nicola Bajetta, Goody Ben Horin, Rebecca Sundén, Omri Drori, Sonam Jamtsho, Dorji Wangchuk, Kfir Bar, Orna Almogi and Shai Fine

Directional Alignment and Narrative Agency in Human–LLM Co-Writing
Halfdan Nordahl Fundal and Yuri Bizzoni

Bias Mitigation in Hiring-Related NLP: Interactions Between Masking, Rewriting, and Adversarial Debiasing
Alexandre Puttick and Rami El-Wazzi

Tracing Thematic Change in Early English-Language Science Fiction, 1818-1930
Jonathan Gordon