

NLP4DH 2025

**The 5th International Conference on Natural Language
Processing for Digital Humanities**

Proceedings of the Conference

May 3-4, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-234-3

Preface

We are delighted to welcome you to the 5th International Conference on Natural Language Processing for Digital Humanities (NLP4DH 2025), held in conjunction with the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) in Albuquerque, New Mexico.

As the intersection of computational methods and the humanities continues to evolve, the need for thoughtful, interdisciplinary dialogue has never been more important. NLP4DH provides a venue for researchers, scholars, and practitioners from both the NLP and digital humanities communities to come together and explore the unique challenges and opportunities presented by language technologies in the context of historical, cultural and social research.

This year, we received a strong set of submissions from around the world, spanning a broad spectrum of topics from corpus creation and annotation strategies for historical texts to the development of novel NLP methods tailored for underrepresented languages, genres and time periods. Many contributions also reflect on methodological and ethical questions, pushing us to think critically about the tools we build and the stories they help uncover.

The program features a mix of long and short papers that reflect the richness of the field. We are particularly proud to highlight the diverse collaborations represented in the accepted papers, which underscore the conference's commitment to cross-disciplinary exchange and open scholarship.

We are grateful to our program committee for their careful reviews and constructive feedback, and to our invited speakers for sharing their insights. We would also like to thank the NAACL 2025 organizers for their support in hosting this conference, and the broader community for its continued engagement and enthusiasm.

We hope that NLP4DH 2025 inspires new conversations, collaborations, and innovations at the intersection of NLP and the humanities.

Organizing Committee

Organizers

Mika Hämäläinen, Metropolia University of Applied Sciences

Emily Öhman, Waseda University

Yuri Bizzoni, Aarhus University

So Miyagawa, University of Tsukuba

Khalid Alnajjar, F-Secure Oyj

Program Committee

Reviewers

Hale Sirin, Johns Hopkins University
Thibault Clérice, INRIA Paris - Almanach
Noémi Ligeti-Nagy, Hungarian Research Centre for Linguistics
Anna Dmitrieva, University of Helsinki
Frederik Arnold, Humboldt Universität Berlin
Dongqi Liu, Universität des Saarlandes
Won Ik Cho, Samsung Advanced Institute of Technology
Konstantin Schulz, Humboldt Universität Berlin
Aynat Rubinstein, Hebrew University of Jerusalem
Alejandro Sierra Múnera, Hasso Plattner Institute
Tim Fischer, University of Hamburg
Shu Okabe, Technische Universität München
Ronja Laarmann-Quante, Ruhr-Universität Bochum
Yoshifumi Kawasaki, The University of Tokyo
Klara Venglarova, Universität Graz
Youngsook Song, Sionix AI
Joshua Wilbur, University of Tartu
Keito Inoshita, Shiga University
Kenichi Iwatsuki, Mirai Translate
Piper Vasicek, Brigham Young University
Mohammed Attia, Google
Laura Manrique-Gómez, Universidad de Los Andes
Craig Messner, Johns Hopkins University
Abhai Pratap Singh, Carnegie Mellon University
Balázs Indig, Eötvös Lorand University
Anton Eklund, Umeå University
Jouni Tuominen, University of Helsinki
Jesse Roberts, Tennessee Technological University
Nikita Neveditsin, St. Mary's University
William Thorne, University of Sheffield
Lev Kharlashkin, Metropolia University of Applied Sciences
Jonne Sälevä, Brandeis University
Gleb Schmidt, Radboud University
Erik Henriksson, University of Turku
Amanda Myntti, University of Turku
Erkki Mervaala, University of Helsinki
Jay Park, Nanyang Technological University
Lama Alqazlan, University of Warwick
Pascale Moreira, Aarhus University
Enrique Manjavacas Arevalo, University of Leiden
Chahan Vidal-Gorène, École Nationale des Chartes
Lucija Krusic, Karl-Franzens-Universität Graz
Lidia Pivovarova, University of Helsinki
Iana Atanassova, University of Franche-Comté
Sebastian Oliver Eck, University of Oxford
Shuo Zhang, Bose Corporation

Tomasz Walkowiak, Wrocław University of Science and Technology
Elissa Nakajima Wickham, Waseda University
Nicolas Gutehrlé, University Bourgogne Franche-Comté
Hanna-Mari Kupari, University of Turku

Table of Contents

<i>A Comparative Analysis of Word Segmentation, Part-of-Speech Tagging, and Named Entity Recognition for Historical Chinese Sources, 1900-1950</i>	
Zhao Fang, Liang-Chun Wu, Xuening Kong and Spencer Dean Stewart	1
<i>Analyzing register variation in web texts through automatic segmentation</i>	
Erik Henriksson, Saara Hellström and Veronika Laippala	7
<i>Analyzing Large Language Models' pastiche ability: a case study on a 20th century Romanian author</i>	
Anca Dinu, Andra-Maria Florescu and Liviu Dinu	20
<i>RAG-Enhanced Neural Machine Translation of Ancient Egyptian Text: A Case Study of THOTH AI</i>	
So Miyagawa	33
<i>Restructuring and visualising dialect dictionary data: Report on Erzya and Moksha materials</i>	
Jack Rueter and Niko Partanen	41
<i>Podcast Outcasts: Understanding Rumble's Podcast Dynamics</i>	
Utkucan Balci, Jay Patel, Berkan Balci and Jeremy Blackburn	48
<i>I only read it for the plot! Maturity Ratings Affect Fanfiction Style and Community Engagement</i>	
Mia Jacobsen and Ross Kristensen-McLachlan	63
<i>The AI Co-Ethnographer: How Far Can Automation Take Qualitative Research?</i>	
Fabian Retkowski, Andreas Sudmann and Alexander Waibel	73
<i>Irony Detection in Hebrew Documents: A Novel Dataset and an Evaluation of Neural Classification Methods</i>	
Avi Shmidman, Elda Weizman and Avishay Gerczuk	91
<i>Masks and Mimicry: Strategic Obfuscation and Impersonation Attacks on Authorship Verification</i>	
Kenneth Alperin, Rohan Leekha, Adaku Uchendu, Trang Nguyen, Srilakshmi Medarametla, Carlos Levya Capote, Seth Aycock and Charlie Dagli	102
<i>Song Lyrics Adaptations: Computational Interpretation of the Pentathlon Principle</i>	
Barbora Štěpánková and Rudolf Rosa	117
<i>MITRA-zh-eval: Using a Buddhist Chinese Language Evaluation Dataset to Assess Machine Translation and Evaluation Metrics</i>	
Sebastian Nehrlich, Avery Chen, Marcus Bingenheimer, Lu Huang, Rouying Tang, Xiang Wei, Leijie Zhu and Kurt Keutzer	129
<i>Effects of Publicity and Complexity in Reader Polarization</i>	
Yuri Bizzoni, Pascale Feldkamp and Kristoffer Nielbo	138
<i>PsyTEX: A Knowledge-Guided Approach to Refining Text for Psychological Analysis</i>	
Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, Gregory Webster and Damon Woodard	151
<i>Advances and Challenges in the Automatic Identification of Indirect Quotations in Scholarly Texts and Literary Works</i>	
Frederik Arnold, Robert Jäschke and Philip Kraut	179
<i>Assessing Crowdsourced Annotations with LLMs: Linguistic Certainty as a Proxy for Trustworthiness</i>	
Tianyi Li, Divya Sree and Tatiana Ringenberg	191

<i>The evolution of relative clauses in the IcePaHC treebank</i> Anton Ingason and Johanna Mechler	202
<i>On Psychology of AI – Does Primacy Effect Affect ChatGPT and Other LLMs?</i> Mika Hämmäläinen	209
<i>The Literary Canons of Large-Language Models: An Exploration of the Frequency of Novel and Author Generations Across Gender, Race and Ethnicity, and Nationality</i> Paulina Toro Isaza and Nalani Kopp	214
<i>Moral reckoning: How reliable are dictionary-based methods for examining morality in text?</i> Ines Rehbein, Lilly Brauner, Florian Ertz, Ines Reinig and Simone Ponzetto	232
<i>Bootstrapping AI: Interdisciplinary Approaches to Assessing OCR Quality in English-Language Historical Documents</i> Samuel Backer and Louis Hyman	251
<i>Poetry in RAGs: Modern Greek interwar poetry generation using RAG and contrastive training</i> Stergios Chatzikiyriakidis and Anastasia Natsina	257
<i>Using Multimodal Models for Informative Classification of Ambiguous Tweets in Crisis Response</i> Sumiko Teng and Emily Öhman	265
<i>Transferring Extreme Subword Style Using Ngram Model-Based Logit Scaling</i> Craig Messner and Tom Lippincott	272
<i>Evaluating Large Language Models for Narrative Topic Labeling</i> Andrew Piper and Sophie Wu	281
<i>Beyond Cairo: Sa’idi Egyptian Arabic Corpus Construction and Analysis</i> Mai Mohamed Eida and Nizar Habash	292
<i>Advancing Sentiment Analysis in Tamil-English Code-Mixed Texts: Challenges and Transformer-Based Solutions</i> Mikhail Krasitskii, Olga Kolesnikova, Liliana Chanona Hernandez, Grigori Sidorov and Alexander Gelbukh	305
<i>Language use of political parties over time: Stylistic Fronting in the Icelandic Gigaword Corpus</i> Johanna Mechler, Lilja Björk Stefánsdóttir and Anton Ingason	313
<i>From Causal Parrots to Causal Prophets? Towards Sound Causal Reasoning with Large Language Models</i> Rahul Babu Shrestha, Simon Malberg and Georg Groh	319
<i>Modern Models, Medieval Texts: A POS Tagging Study of Old Occitan</i> Matthias Schöffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann and Matthias Aßenmacher	334
<i>A Data-driven Investigation of Euphemistic Language: Comparing the usage of slave and servant in 19th century US newspapers</i> Jaihyun Park and Ryan Cordell	350
<i>It’s about What and How you say it: A Corpus with Stance and Sentiment Annotation for COVID-19 Vaccines Posts on X/Twitter by Brazilian Political Elites</i> Lorena Barberia, Pedro Schmalz, Norton Trevisan Roman, Belinda Lombard and Tatiane Moraes de Sousa	365

<i>A Bit of This, a Bit of That: Building a Genre and Topic Annotated Dataset of Historical Newspaper Articles with Soft Labels and Confidence Scores</i>	
Karin Stahel, Irenie How, Lauren Millar, Luis Paterson, Daniel Steel and Kaspar Middendorf	377
<i>Development of Old Irish Lexical Resources, and Two Universal Dependencies Treebanks for Diplomatically Edited Old Irish Text</i>	
Adrian Doyle and John McCrae	393
<i>Augmented Close Reading for Classical Latin using BERT for Intertextual Exploration</i>	
Ashley Gong, Katy Gero and Mark Schiefsky	403
<i>An evaluation of Named Entity Recognition tools for detecting person names in philosophical text</i>	
Ruben Weijers and Jelke Bloem	418
<i>Testing Language Creativity of Large Language Models and Humans</i>	
Anca Dinu and Andra-Maria Florescu	426
<i>Strategies for political-statement segmentation and labelling in unstructured text</i>	
Dmitry Nikolaev and Sean Papay	437
<i>Mining the Past: A Comparative Study of Classical and Neural Topic Models on Historical Newspaper Archives</i>	
Keerthana Murugaraj, Salima Lamsiyah, Marten DURING and Martin Theobald	452
<i>A Comparative Analysis of Ethical and Safety Gaps in LLMs using Relative Danger Coefficient</i>	
Yehor Tereshchenko and Mika Hämäläinen	464
<i>Threefold model for AI Readiness: A Case Study with Finnish Healthcare SMEs</i>	
Mohammed Alnajjar, Khalid Alnajjar and Mika Hämäläinen	478
<i>AI Assistant for Socioeconomic Empowerment Using Federated Learning</i>	
Nahed Abdelgaber, Labiba Jahan, Nino Castellano, Joshua Oltmanns, Mehak Gupta, Jia Zhang, Akshay Pednekar, Ashish Basavaraju, Ian Velazquez and Zerui Ma	490
<i>Team Conversational AI: Introducing Effervesce</i>	
Erjon Skenderi, Salla-Maaria Laaksonen and Jukka Huhtamäki	502
<i>Mapping Hymns and Organizing Concepts in the Rigveda: Quantitatively Connecting the Vedic Suktas</i>	
Venkatesh Bollineni, Igor Crk and Eren Gultepe	514
<i>EduPo: Progress and Challenges of Automated Analysis and Generation of Czech Poetry</i>	
Rudolf Rosa, David Mareček, Tomáš Musil, Michal Chudoba and Jakub Landsperský	524
<i>A City of Millions: Mapping Literary Social Networks At Scale</i>	
Sil Hamilton, Rebecca Hicke, David Mimno and Matthew Wilkens	543
<i>VLG-BERT: Towards Better Interpretability in LLMs through Visual and Linguistic Grounding</i>	
Toufik Mechouma, Ismail Biskri and Serge Robert	550
<i>Historical Ink: Exploring Large Language Models for Irony Detection in 19th-Century Spanish</i>	
Kevin Cohen, Laura Manrique-Gómez and Ruben Manrique	559
<i>Insights into developing analytical categorization schemes: three problem types related to annotation agreement</i>	
Pihla Toivanen, Eetu Mäkelä and Antti Kanner	570
<i>A Comprehensive Evaluation of Cognitive Biases in LLMs</i>	
Simon Malberg, Roman Poletukhin, Carolin Schuster and Georg Groh Groh	578

<i>AI with Emotions: Exploring Emotional Expressions in Large Language Models</i> Shin-nosuke Ishikawa and Atsushi Yoshino	614
<i>Fearful Falcons and Angry Llamas: Emotion Category Annotations of Arguments by Humans and LLMs</i> Lynn Greschner and Roman Klinger	628
<i>HateImgPrompts: Mitigating Generation of Images Spreading Hate Speech</i> Vineet Kumar Khullar, Venkatesh Velugubantla, Bhanu Prakash Reddy Rella, Mohan Krishna Mannava and MSVPJ Sathvik	647

Program

Saturday, May 3, 2025

- 09:00 - 09:10 *Opening words*
- 09:10 - 10:30 *Session 1*
- 09:10 - 09:30 *Mapping Hymns and Organizing Concepts in the Rigveda: Quantitatively Connecting the Vedic Suktas*
- 09:30 - 09:50 *Historical Ink: Exploring Large Language Models for Irony Detection in 19th-Century Spanish*
- 09:50 - 10:10 *Modern Models, Medieval Texts: A POS Tagging Study of Old Occitan*
- 10:10 - 10:30 *Lightning talks*
- 10:30 - 11:00 *Coffee break*
- 11:00 - 12:20 *Session 2*
- 11:00 - 11:20 *Transferring Extreme Subword Style Using Ngram Model-Based Logit Scaling*
- 11:20 - 11:40 *Advances and Challenges in the Automatic Identification of Indirect Quotations in Scholarly Texts and Literary Works*
- 11:40 - 12:00 *PsyTEx: A Knowledge-Guided Approach to Refining Text for Psychological Analysis*
- 12:00 - 12:20 *The evolution of relative clauses in the IcePaHC treebank*
- 12:20 - 14:00 *Lunch break*
- 14:00 - 15:20 *Session 3*
- 14:00 - 14:20 *A Bit of This, a Bit of That: Building a Genre and Topic Annotated Dataset of Historical Newspaper Articles with Soft Labels and Confidence Scores*
- 14:20 - 14:40 *Song Lyrics Adaptations: Computational Interpretation of the Pentathlon Principle*

Saturday, May 3, 2025 (continued)

- 14:40 - 15:00 *Masks and Mimicry: Strategic Obfuscation and Impersonation Attacks on Authorship Verification*
- 15:00 - 15:20 *Augmented Close Reading for Classical Latin using BERT for Intertextual Exploration*
- 15:20 - 16:00 *Coffee break*
- 16:00 - 16:30 *Poster session*
- 16:00 - 16:30 *Analyzing register variation in web texts through automatic segmentation*
- 16:00 - 16:30 *From Causal Parrots to Causal Prophets? Towards Sound Causal Reasoning with Large Language Models*
- 16:00 - 16:30 *AI with Emotions: Exploring Emotional Expressions in Large Language Models*
- 16:00 - 16:30 *RAG-Enhanced Neural Machine Translation of Ancient Egyptian Text: A Case Study of THOTH AI*
- 16:00 - 16:30 *Effects of Publicity and Complexity in Reader Polarization*
- 16:00 - 16:30 *On Psychology of AI – Does Primacy Effect Affect ChatGPT and Other LLMs?*
- 16:00 - 16:30 *A City of Millions: Mapping Literary Social Networks At Scale*
- 16:00 - 16:30 *The Literary Canons of Large-Language Models: An Exploration of the Frequency of Novel and Author Generations Across Gender, Race and Ethnicity, and Nationality*

Sunday, May 4, 2025

- 09:00 - 09:30 *Online Posters*
- 09:00 - 09:30 *Testing Language Creativity of Large Language Models and Humans*
- 09:00 - 09:30 *Analyzing Large Language Models' pastiche ability: a case study on a 20th century Romanian author*
- 09:00 - 09:30 *VLG-BERT: Towards Better Interpretability in LLMs through Visual and Linguistic Grounding*
- 09:00 - 09:30 *Mining the Past: A Comparative Study of Classical and Neural Topic Models on Historical Newspaper Archives*
- 09:00 - 09:30 *A data-driven investigation of euphemistic language: Comparing the usage of slave and servant in 19th century US newspapers*
- 09:00 - 09:30 *Using Multimodal Models for Informative Classification of Ambiguous Tweets in Crisis Response*
- 09:00 - 09:30 *Evaluating Large Language Models for Narrative Topic Labeling*
- 09:00 - 09:30 *It's about What and How you say it: A Corpus with Stance and Sentiment Annotation for COVID-19 Vaccines Posts on X/Twitter by Brazilian Political Elites*
- 09:00 - 09:30 *Development of Old Irish Lexical Resources, and Two Universal Dependencies Treebanks for Diplomatically Edited Old Irish Text*
- 09:00 - 09:30 *Insights into developing analytical categorization schemes: three problem types related to annotation agreement*
- 09:00 - 09:30 *Team Conversational AI: Introducing Effervesce*
- 09:00 - 09:30 *EduPo: Progress and Challenges of Automated Analysis and Generation of Czech Poetry*
- 09:00 - 09:30 *Fearful Falcons and Angry Llamas: Emotion Category Annotations of Arguments by Humans and LLMs*
- 09:00 - 09:30 *Moral reckoning: How reliable are dictionary-based methods for examining morality in text?*
- 09:00 - 09:30 *AI Assistant for Socioeconomic Empowerment Using Federated Learning*

Sunday, May 4, 2025 (continued)

- 09:00 - 09:30 *HateImgPrompts: Mitigating Generation of Images Spreading Hate Speech*
- 09:00 - 09:30 *Assessing Crowdsourced Annotations with LLMs: Linguistic Certainty as a Proxy for Trustworthiness*
- 09:00 - 09:30 *An evaluation of Named Entity Recognition tools for detecting person names in philosophical text*
- 09:00 - 09:30 *MITRA-zh-eval: Using a Buddhist Chinese Language Evaluation Dataset to Assess Machine Translation and Evaluation Metrics*
- 09:00 - 09:30 *Strategies for political-statement segmentation and labelling in unstructured text*
- 09:00 - 09:30 *Poetry in RAGs: Modern Greek interwar poetry generation using RAG and contrastive training*
- 09:00 - 09:30 *Advancing Sentiment Analysis in Tamil-English Code-Mixed Texts: Challenges and Transformer-Based Solutions*
- 09:00 - 09:30 *Restructuring and visualising dialect dictionary data: Report on Erzya and Moksha materials*
- 09:00 - 09:30 *A Comparative Analysis of Ethical and Safety Gaps in LLMs using Relative Danger Coefficient*
- 09:00 - 09:30 *Threefold model for AI Readiness: A Case Study with Finnish Healthcare SMEs*
- 09:30 - 10:30 *NLP4DH business meeting*
- 10:30 - 11:00 *Coffee break*
- 11:00 - 12:20 *Session 4*
- 11:00 - 11:20 *Irony Detection in Hebrew Documents: A Novel Dataset and an Evaluation of Neural Classification Methods*
- 11:20 - 11:40 *Podcast Outcasts: Understanding Rumble's Podcast Dynamics*
- 11:40 - 12:00 *The AI Co-Ethnographer: How Far Can Automation Take Qualitative Research?*

Sunday, May 4, 2025 (continued)

- 12:00 - 12:20 *A Comprehensive Evaluation of Cognitive Biases in LLMs*
- 12:20 - 14:00 *Lunch break*
- 14:00 - 15:20 *Session 5*
- 14:00 - 14:20 *Bootstrapping AI: Interdisciplinary Approaches to Assessing OCR Quality in English-Language Historical Documents*
- 14:20 - 14:40 *A Comparative Analysis of Word Segmentation, Part-of-Speech Tagging, and Named Entity Recognition for Historical Chinese Sources, 1900-1950*
- 14:40 - 15:00 *Beyond Cairo: Sa'idi Egyptian Arabic Literary Corpus Construction and Analysis*
- 15:00 - 15:20 *Language use of political parties over time: Stylistic Fronting in the Icelandic Gigaword Corpus*
- 15:20 - 16:00 *Coffee break*