

# **4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH 2024)**

Miami, Florida, USA  
16 November 2024

ISBN: 979-8-3313-0850-6

**Printed from e-media with permission by:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571

**Some format issues inherent in the e-media version may also appear in this print version.**

Copyright© (2024) by the Association for Computational Linguistics  
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics  
at the address below.

Association for Computational Linguistics  
209 N. Eighth Street  
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006

Fax: 1-570-476-0860

[acl@aclweb.org](mailto:acl@aclweb.org)

**Additional copies of this publication are available from:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571 USA  
Phone: 845-758-0400  
Fax: 845-758-2633  
Email: [curran@proceedings.com](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

## Table of Contents

|  |     |
|--|-----|
| <i>Text Length and the Function of Intentionality: A Case Study of Contrastive Subreddits</i><br>Emily Sofi Ohman and Aatu Liimatta .....  | 1   |
| <i>Tracing the Genealogies of Ideas with Sentence Embeddings</i><br>Lucian Li .....  | 9   |
| <i>Evaluating Computational Representations of Character: An Austen Character Similarity Benchmark</i><br>Funing Yang and Carolyn Jane Anderson .....  | 17  |
| <i>Investigating Expert-in-the-Loop LLM Discourse Patterns for Ancient Intertextual Analysis</i><br>Ray Umphrey, Jesse Roberts and Lindsey Roberts .....   | 31  |
| <i>Extracting Relations from Ecclesiastical Cultural Heritage Texts</i><br>Giulia Cruciani .....   | 41  |
| <i>Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience</i><br>Lucija Krusic .....   | 51  |
| <i>It is a Truth Individually Acknowledged: Cross-references On Demand</i><br>Piper Vasicek, Courtni Byun and Kevin Seppi .....  | 63  |
| <i>Extracting position titles from unstructured historical job advertisements</i><br>Klara Venglarova, Raven Adam and Georg Vogeler .....  | 75  |
| <i>Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs</i><br>Natalie Hervieux, Peiran Yao, Susan Brown and Denilson Barbosa .....   | 85  |
| <i>NLP for Digital Humanities: Processing Chronological Text Corpora</i><br>Adam Pawłowski and Tomasz Walkowiak .....  | 105 |
| <i>A Multi-task Framework with Enhanced Hierarchical Attention for Sentiment Analysis on Classical Chinese Poetry: Utilizing Information from Short Lines</i><br>Quanqi Du and Veronique Hoste .....   | 113 |
| <i>Exploring Similarity Measures and Intertextuality in Vedic Sanskrit Literature</i><br>So Miyagawa, Yuki Kyogoku, Yuzuki Tsukagoshi and Kyoko Amano .....  | 123 |
| <i>Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction</i><br>Laura Manrique-Gomez, Tony Montes, Arturo Rodriguez Herrera and Ruben Manrique .....  | 132 |
| <i>Canonical Status and Literary Influence: A Comparative Study of Danish Novels from the Modern Breakthrough (1870–1900)</i><br>Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos, Kenneth Enevoldsen, Katrine Baunvig and Kristoffer Nielbo ..... | 140 |
| <i>Deciphering psycho-social effects of Eating Disorder : Analysis of Reddit Posts using Large Language Model(LLM)s and Topic Modeling</i><br>Medini Chopra, Anindita Chatterjee, Lipika Dey and Partha Pratim Das .....                                     | 156 |
| <i>Topic-Aware Causal Intervention for Counterfactual Detection</i><br>Thong Thanh Nguyen and Truc-My Nguyen .....   | 165 |

|  |     |
|--|-----|
| <i>UD for German Poetry</i>  |     |
| Stefanie Dipper and Ronja Laarmann-Quante .....  | 177 |
| <i>Molyé: A Corpus-based Approach to Language Contact in Colonial France</i>   |     |
| Rasul Dent, juliette janes, Thibault Clerice, Pedro Ortiz Suarez and Benoît Sagot .....  | 189 |
| <i>Vector Poetics: Parallel Couplet Detection in Classical Chinese Poetry</i>  |     |
| Maciej Kurzynski, Xiaotong Xu and Yu Feng .....  | 200 |
| <i>Adapting Measures of Literality for Use with Historical Language Data</i>   |     |
| Adam Roussel .....   | 209 |
| <i>Improving Latin Dependency Parsing by Combining Treebanks and Predictions</i>   |     |
| Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala and Jenna Kanerva .....  | 216 |
| <i>From N-grams to Pre-trained Multilingual Models For Language Identification</i>   |     |
| Thapelo Andrew Sindane and Vukosi Marivate .....   | 229 |
| <i>Visualising Changes in Semantic Neighbourhoods of English Noun Compounds over Time</i>  |     |
| Malak Rassem, Myrto Tsigkouli, Chris W. Jenkins, Filip Miletić and Sabine Schulte im Walde   | 240 |
| <i>SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek</i>                                      |     |
| Konstantin Schulz and Florian Deichsler .....  | 247 |
| <i>A Two-Model Approach for Humour Style Recognition</i>   |     |
| Mary Ogbuka Kenneth, Foad Khosmood and Abbas Edalat .....  | 259 |
| <i>N-gram-Based Preprocessing for Sandhi Reversion in Vedic Sanskrit</i>   |     |
| Yuzuki Tsukagoshi and Ikki Ohmukai .....   | 275 |
| <i>Enhancing Swedish Parliamentary Data: Annotation, Accessibility, and Application in Digital Humanities</i>                              |     |
| Shafqat Mumtaz Virk, Claes Ohlsson, Nina Tahmasebi, Henrik Björck and Leif Runefelt .....  | 280 |
| <i>Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams</i>                                      |     |
| Roberts Dargis, Guntis Bārzdiņš, Inguna Skadiņa and Baiba Saulite .....  | 289 |
| <i>Computational Methods for the Analysis of Complementizer Variability in Language and Literature: The Case of Hebrew "she-" and "ki"</i> |     |
| Avi Shmidman and Aynat Rubinstein .....  | 294 |
| <i>From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations</i>                      |     |
| Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo and Veronika Laippala .....            | 308 |
| <i>Testing and Adapting the Representational Abilities of Large Language Models on Folktales in Low-Resource Languages</i>                 |     |
| J. A. Meaney, Beatrice Alex and William Lamb .....   | 319 |
| <i>Examining Language Modeling Assumptions Using an Annotated Literary Dialect Corpus</i>  |     |
| Craig Messner and Thomas Lippincott .....  | 325 |

|  |     |
|--|-----|
| <i>Evaluating Language Models in Location Referring Expression Extraction from Early Modern and Contemporary Japanese Texts</i>  |     |
| Ayuki Katayama, Yusuke Sakai, Shohei Higashiyama, Hiroki Ouchi, Ayano Takeuchi, Ryo Bando, Yuta Hashimoto, Toshinobu Ogiso and Taro Watanabe . . . . .                             | 331 |
| <i>Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction</i>   |     |
| Woori Jang and Seohyon Jung . . . . .  | 339 |
| <i>Text vs. Transcription: A Study of Differences Between the Writing and Speeches of U.S. Presidents</i>  |     |
| Mina Rajaei Moghadam, Mosab Rezaei, Gülşat Aygen and Reva Freedman . . . . .   | 352 |
| <i>Mitigating Biases to Embrace Diversity: A Comprehensive Annotation Benchmark for Toxic Language</i>   |     |
| Xinmeng Hou . . . . .  | 362 |
| <i>Classification of Buddhist Verses: The Efficacy and Limitations of Transformer-Based Models</i>   |     |
| Nikita Neveditsin, Ambuja Salgaonkar, Pawan Lingras and Vijay Mago . . . . .   | 377 |
| <i>Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora</i>  |     |
| Amanda Myntti, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala and Erik Henriksen . . . . .  | 386 |
| <i>Sui Generis: Large Language Models for Authorship Attribution and Verification in Latin</i>   |     |
| Svetlana Gorovaia, Gleb Schmidt and Ivan P. Yamshchikov . . . . .  | 398 |
| <i>Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration</i>   |     |
| Ryo Igarashi and So Miyagawa . . . . .   | 413 |
| <i>Exploring Large Language Models for Qualitative Data Analysis</i>   |     |
| Tim Fischer and Chris Biemann . . . . .  | 423 |
| <i>Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of RNNs, Transformers and LLMs</i>   |     |
| Chahan Vidal-Gorène, Nadi Tomeh and Victoria Khurshudyan . . . . .   | 438 |
| <i>Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference for Cost-Effective Cultural Heritage Dataset Generation</i> |     |
| William Thorne, Ambrose Robinson, Bohua Peng, Chenghua Lin and Diana Maynard . . . . .   | 450 |
| <i>Assessing Large Language Models in Translating Coptic and Ancient Greek Ostraca</i>   |     |
| Audric-Charles Wannaz and So Miyagawa . . . . .  | 463 |
| <i>The Social Lives of Literary Characters: Combining citizen science and language models to understand narrative social networks</i>  |     |
| Andrew Piper, Michael Xu and Derek Ruths . . . . .   | 472 |
| <i>Multi-word expressions in biomedical abstracts and their plain English adaptations</i>  |     |
| Sergei Bagdasarov and Elke Teich . . . . .   | 483 |
| <i>Assessing the Performance of ChatGPT-4, Fine-tuned BERT and Traditional ML Models on Moroccan Arabic Sentiment Analysis</i>   |     |
| Mohamed HANNANI, Abdelhadi Soudi and Kristof Van Laerhoven . . . . .   | 489 |

|   |     |
|---|-----|
| <i>Analyzing Pokémon and Mario Streamers' Twitch Chat with LLM-based User Embeddings</i><br>Mika Hämmäläinen, Jack Rueter and Khalid Alnajjar .....                               | 499 |
| <i>Corpus Development Based on Conflict Structures in the Security Field and LLM Bias Verification</i><br>Keito Inoshita .....  | 504 |
| <i>Generating Interpretations of Policy Announcements</i><br>Andreas Marfurt, Ashley Thornton, David Sylvan and James Henderson .....   | 513 |
| <i>Order Up! Micromanaging Inconsistencies in ChatGPT-4o Text Analyses</i><br>Erkki Mervaala and Ilona Kousa .....  | 521 |
| <i>CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration</i><br>Anton Eklund, Mona Forsman and Frank Drewes .....                           | 536 |
| <i>Empowering Teachers with Usability-Oriented LLM-Based Tools for Digital Pedagogy</i><br>Melany Vanessa Macias, Lev Kharlashkin, Leo Einari Huovinen and Mika Hämmäläinen ..... | 549 |