GlobalNLP 2025

Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models

associated with

The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025

Edited by Sudhansu Bala Das, Pruthwik Mishra, Alok Singh,
Shamsuddeen Hassan Muhammad, Asif Ekbal and Uday Kumar Dasi

12 September, 2025
Varna, Bulgaria

The Workshop on Beyond English: Natural Language Processing
for all Languages in an Era of Large Language Models
Associated with the International Conference
Recent Advances in Natural Language Processing
RANLP'2025

PROCEEDINGS

Varna, Bulgaria
12 September 2025

Online ISBN 978-954-452-105-9

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

The field of Natural Language Processing (NLP) has achieved remarkable progress in recent years, powered by the emergence of Large Language Models (LLMs) and generative AI. These advancements have significantly improved language technology for high-resource languages such as English, Mandarin, and German. However, the majority of the world's languages, including medium-resource, under-resourced, and low-resource ones—remain underserved due to limited datasets, resources, and linguistic tools.

The GlobalNLP 2025Workshop, titled "Beyond English: Natural Language Processing for All Languages in an Era of Large Language Models", was organized as part of RANLP 2025, held from 12 September 2025 in Varna, Bulgaria. This workshop provided an inclusive platform for researchers, linguists, developers, and practitioners worldwide to explore how cutting-edge NLP techniques can be extended to every language, regardless of its resource availability. We received a total of 28 paper submissions, of which a curated selection was accepted for inclusion in the proceedings following a rigorous peer-review process. The accepted papers span various domains, from cross-lingual modeling and corpus creation to NLP applications in healthcare, education, and cultural preservation. We were honored to have keynote talks by distinguished experts, including: The workshop featured two invited talks by distinguished researchers, offering complementary perspectives on multilingualism, machine translation, and the future of language technologies in medicine. The first invited speaker was Prof. Dipti Misra Sharma, Professor Emeritus at the International Institute of Information Technology (IIIT) Hyderabad, India. Her talk, titled "Multilingualism, LLMs, and Machine Translation", was delivered in the opening session from 09:00–09:45. Prof. Sharma is a pioneering figure in the field of Natural Language Processing, with a career spanning more than two decades of landmark contributions to machine translation, multilingual NLP, and linguistic resource development. At the Language Technologies Research Centre (LTRC) at IIIT Hyderabad, she has led large-scale government-funded initiatives to create corpora, treebanks, morphological analyzers, and evaluation frameworks that have become foundational resources for researchers worldwide. Her work bridges theoretical linguistics, computational modeling, and real-world deployment, with a strong focus on low-resource and morphologically rich languages. She has been instrumental in developing multilingual translation systems, interoperable linguistic tools, and pipelines for code-mixed language processing. Beyond her technical work, she has played a leading role in policy-level language technology planning in India and mentored a generation of NLP researchers. In her talk, Prof. Sharma traced the evolution of machine translation from its foundations to state-of-the-art approaches, with particular attention to the role of Large Language Models (LLMs). She emphasized both the opportunities and challenges of applying LLMs in multilingual contexts, highlighting their potential for linguistic inclusivity while underscoring the need for data-efficient, linguistically informed methods.

The second invited speaker was Prof. Michael G. Madden, Established Professor and leads the Machine Learning Research Group that he set up in 2001. His talk, "Advances in Natural Language Processing and Machine Learning for Medicine", was presented in the afternoon session from 13:25–14:10. Prof. Madden is an internationally recognized leader in machine learning, artificial intelligence, and data-driven modeling. Since founding the Machine Learning Research Group at Galway in 2001, he has produced influential work on deep learning, probabilistic reasoning, dynamic Bayesian networks, and reinforcement learning. His expertise is especially relevant to GlobalNLP 2025 through his focus on combining data-driven learning with structured background knowledge, a central challenge in adapting LLMs for specialised and multilingual applications. His career spans both academia and industry. He founded the AI spin-out company AnalyzeIQ Ltd, has served as a visiting scientist at leading institutions such as UC Berkeley, UC Irvine, and the University of Helsinki, and has fostered collaborations between academia, industry, and government. In his keynote, Prof. Madden showcased applications of NLP and machine learning in medicine, demonstrating how knowledge-aware and inclusive AI systems can be developed to support decision-making in sensitive, high-stakes domains. His talk highlighted the need

for models that are both technically advanced and socially responsible, resonating strongly with the workshop's vision of inclusive global NLP.

In addition to paper sessions and keynote addresses, the workshop featured:

• A panel discussion on the challenges and opportunities in building multilingual LLMs.

• Interactive demo sessions showcasing NLP tools and technologies developed for diverse linguistic communities.

Core themes of the workshop included inclusivity, resource creation for under-represented languages, and the practical deployment of LLMs across domains such as education, healthcare, and cultural heritage.

• Data-efficient NLP: Transfer learning, few-shot and zero-shot methods for low-resource settings.

• Multilingual and cross-lingual modeling: Techniques adaptive to morphologically rich and typologically diverse languages.

• Semantic and ontology-driven approaches: Entity linking, semantic similarity, and knowledge graph integration.

• Resource creation and reuse: Development of sustainable corpora, tools, and evaluation benchmarks.

• Real-world impact: Applying NLP in domains such as education, healthcare, policy, and digital humanities.

• LLMs in practice: Deployments for code generation, document summarization, personalized conversational agents, and beyond.

We extend our deepest gratitude to all authors for submitting their research, and to the Program Committee members for their careful and insightful reviews. We especially thank our keynote speakers, panelists, and demonstrators for enriching the workshop with their expertise. Finally, we are thankful to the RANLP 2025 Organizing Committee for supporting and hosting this inclusive initiative. We hope that these proceedings will inspire ongoing research and collaboration toward more equitable and universal NLP.

GlobalNLP 2025 Organizing Committee

Organizing Committee and Volume Editors

• Sudhansu Bala Das (Insight Research Ireland Centre for Data Analytics, University of Galway, Ireland)
• Pruthwik Mishra (SVNIT Surat, India)
• Alok Singh (University of Oxford, UK)
• Shamsuddeen Hassan Muhammad (AfricaNLP, Imperial College London)
• Asif Ekbal (IIT Jodhpur, India)
• Uday Kumar Das (Software Engineer, Dundalk)

Program Committee

The program committee consists of distinguished researchers and practitioners from across the globe, actively working in the fields of Natural Language Processing (NLP) and Large Language Models (LLMs).
• Alexander Gelbukh (Instituto Politécnico Nacional, Mexico)
• Bidyut Kumar Patra (IIT BHU, India)
• Clarence Teo (Nanyang Technological University, Singapore)
• Gaurish Thakkar (University of Zagreb, Croatia)
• Helena Moniz (Universidade de Lisboa, Lisbon, Portugal)
• Idris Abdulmumin (DSFSI, University of Pretoria)
• Ibrahim Said Ahmad (Northeastern University)
• Juri Opitz (University of Zurich, Switzerland)
• Luan Thanh Nguyen (Vietnam National University Ho Chi Minh City, Vietnam)
• Marie-Aude Lefer (UCLouvain, Belgium)
• Mohammed Hasanuzzaman (Queen's University Belfast, UK)
• Moritz Schaeffer (Johannes Gutenberg University of Mainz, Germany)
• Muslim Jameel Sayed (Atlantic Technological University, Ireland)
• Pádraic Moran (University of Galway, Ireland)
• Paolo Rosso (Valencia Polytechnic University, Spain)
• Paul Buitelaar (University of Galway, Ireland)
• Soumik Mandal (NYU Tandon School of Engineering, USA)
• Surangika Ranathunga (Massey University, New Zealand)
• Uthayasanker Thayasivam (University of Moratuwa, Sri Lanka)

Table of Contents

Towards the Creation of a Collao Quechua–Spanish Parallel Corpus Using Optical Character Recognition
Gian Carlo Orcotoma Mormontoy, Lida Leon Nuñez and Hugo Espetia Huamanga 1

Prompt Balance Matters: Understanding How Imbalanced Few-Shot Learning Affects Multilingual Sense Disambiguation in LLMs
Deshan Koshala Sumanathilaka, Nicholas Micallef and Julian Hough 7

Development of a Low-Cost Named Entity Recognition System for Odia Language using Deep Active Learning
Tusarkanta Dalai, Tapas Kumar Mishra, Pankaj Kumar Sa, Prithviraj Mohanty, Chittaranjan Swain and Ajit Kumar Nayak 16

Non-Contextual BERT or FastText? A Comparative Analysis
Abhay Shanbhag, Suramya Jadhav, Amogh Thakurdesai, Ridhima Bhaskar Sinare and Raviraj Joshi 27

Kantika: A Knowledge-Radiant Framework for Dermatology QA using IR-CoT and RAPTOR-Augmented Retrieval
Deep Das, Vikram Mehrolia, Rahul Dixit and Rohit Kumar 34

GeistBERT: Breathing Life into German NLP
Raphael Scheible-Schmitt and Johann Frei 42

Identifying Contextual Triggers in Hate Speech Texts Using Explainable Large Language Models
Dheeraj Kodati and Bhuvana Sree Lakkireddy 51

PortBERT: Navigating the Depths of Portuguese Language Models
Raphael Scheible-Schmitt, Henry He and Armando B. Mendes 59

Quality Matters Measuring the Effect of Human-Annotated Translation Quality on English-Slovak Machine Translation
Matúš Kleštinec and Daša Munková 72

Spatio-Temporal Mechanism in Multilingual Sentiment Analysis
Adarsh Singh Jadon, Vivek Tiwari, Chittaranjan Swain and Deepak Kumar Dewangan 82

Automatic Animacy Classification for Latvian Nouns
Ralfs Brutāns and Jelke Bloem 90

Bootstrapping a Sentence-Level Corpus Quality Classifier for Web Text using Active Learning
Maximilian Bley, Thomas Eckart and Christopher Schröder 98

Fine-Grained Arabic Offensive Language Classification with Taxonomy, Sentiment, and Emotions
Natalia Vanetik, Marina Litvak and Chaya Liebeskind 110

Measuring Prosodic Richness in LLM-Generated Responses for Conversational Recommendation
Darshna Parmar and Pramit Mazumdar 120

Assessing the Accuracy of AI-Generated Idiom Translations
Marijana Gasparovic, Marija Brala Vukanovic and Marija Brkic Bakaric 131

From Pixels to Prompts: Evaluating ChatGPT-4o in Face Recognition, Age Estimation, and Gender Classification
Jashn Jain, Praveen Kumar Chandaliya and Dhruti P. Sharma 141

DRISHTI: Drug Recognition and Integrated System for Helping the visually Impaired with Tag-based Identification
Sajeeb Das, Srijit Paul, Ucchas Muhury, Akib Jayed Islam, Dhruba Jyoti Barua, Sultanus Salehin and Prasun Datta 149

What Language(s) Does Aya-23 Think In? How Multilinguality Affects Internal Language Representations
Katharina A. T. T. Trinley, Toshiki Nakai, Tatiana Anikina and Tanja Baeumel 159

FedCliMask: Context-Aware Federated Learning with Ontology-Guided Semantic Masking for Clinical NLP
Srijit Paul, Sajeeb Das, Ucchas Muhury, Akib Jayed Islam, Dhruba Jyoti Barua, Sultanus Salehin and Prasun Datta 172

A study on the language independent stemmer in the Indian language IR
Siba Sankar Sahu and Sukomal Pal 181

Checklist Engineering Empowers Multilingual LLM Judges
Mohammad Ghiasvand Mohammadkhani and Hamid Beigy 190

C A N C E R: Corpus for Accurate Non-English Cancer-related Educational Resources
Anika Harju, Asma Shakeel, Tiantian He, Tianqi Xu and Aaro Harju 197

Conference Program

Towards the Creation of a Collao Quechua–Spanish Parallel Corpus Using Optical Character Recognition
Gian Carlo Orcotoma Mormontoy, Lida Leon Nuñez and Hugo Espetia Huamanga

Prompt Balance Matters: Understanding How Imbalanced Few-Shot Learning Affects Multilingual Sense Disambiguation in LLMs
Deshan Koshala Sumanathilaka, Nicholas Micallef and Julian Hough

Development of a Low-Cost Named Entity Recognition System for Odia Language using Deep Active Learning
Tusarkanta Dalai, Tapas Kumar Mishra, Pankaj Kumar Sa, Prithviraj Mohanty, Chittaranjan Swain and Ajit Kumar Nayak

Non-Contextual BERT or FastText? A Comparative Analysis
Abhay Shanbhag, Suramya Jadhav, Amogh Thakurdesai, Ridhima Bhaskar Sinare and Raviraj Joshi

Kantika: A Knowledge-Radiant Framework for Dermatology QA using IR-CoT and RAPTOR-Augmented Retrieval
Deep Das, Vikram Mehrolia, Rahul Dixit and Rohit Kumar

GeistBERT: Breathing Life into German NLP
Raphael Scheible-Schmitt and Johann Frei

Identifying Contextual Triggers in Hate Speech Texts Using Explainable Large Language Models
Dheeraj Kodati and Bhuvana Sree Lakkireddy

PortBERT: Navigating the Depths of Portuguese Language Models
Raphael Scheible-Schmitt, Henry He and Armando B. Mendes

Quality Matters Measuring the Effect of Human-Annotated Translation Quality on English-Slovak Machine Translation
Matúš Kleštinec and Daša Munková

Spatio-Temporal Mechanism in Multilingual Sentiment Analysis
Adarsh Singh Jadon, Vivek Tiwari, Chittaranjan Swain and Deepak Kumar Dewangan

Automatic Animacy Classification for Latvian Nouns
Ralfs Brutāns and Jelke Bloem

Bootstrapping a Sentence-Level Corpus Quality Classifier for Web Text using Active Learning
Maximilian Bley, Thomas Eckart and Christopher Schröder

Fine-Grained Arabic Offensive Language Classification with Taxonomy, Sentiment, and Emotions
Natalia Vanetik, Marina Litvak and Chaya Liebeskind

Measuring Prosodic Richness in LLM-Generated Responses for Conversational Recommendation
Darshna Parmar and Pramit Mazumdar

Assessing the Accuracy of AI-Generated Idiom Translations
Marijana Gasparovic, Marija Brala Vukanovic and Marija Brkic Bakaric

From Pixels to Prompts: Evaluating ChatGPT-4o in Face Recognition, Age Estimation, and Gender Classification
Jashn Jain, Praveen Kumar Chandaliya and Dhruti P. Sharma

DRISHTI: Drug Recognition and Integrated System for Helping the visually Impaired with Tag-based Identification
Sajeeb Das, Srijit Paul, Ucchas Muhury, Akib Jayed Islam, Dhruba Jyoti Barua, Sultanus Salehin and Prasun Datta

What Language(s) Does Aya-23 Think In? How Multilinguality Affects Internal Language Representations
Katharina A. T. T. Trinley, Toshiki Nakai, Tatiana Anikina and Tanja Baeumel

FedCliMask: Context-Aware Federated Learning with Ontology-Guided Semantic Masking for Clinical NLP
Srijit Paul, Sajeeb Das, Ucchas Muhury, Akib Jayed Islam, Dhruba Jyoti Barua, Sultanus Salehin and Prasun Datta

A study on the language independent stemmer in the Indian language IR
Siba Sankar Sahu and Sukomal Pal

Checklist Engineering Empowers Multilingual LLM Judges
Mohammad Ghiasvand Mohammadkhani and Hamid Beigy

C A N C E R: Corpus for Accurate Non-English Cancer-related Educational Resources
Anika Harju, Asma Shakeel, Tiantian He, Tianqi Xu and Aaro Harju

Integrating Large Language Models for Comprehensive Study and Sentiment Analysis of Student Feedback
Jana Kuzmanova, Katerina Zdravkova and Ivan Chorbev