IWCLUL 2024

The 9th International Workshop on Computational Linguistics for Uralic Languages

Proceedings of the Workshop

November 28-29, 2024

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 979-8-89176-128-5

Preface

Welcome to the Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages (IWCLUL), a dedicated forum for researchers, academics and practitioners in the field of computational linguistics with a focus on Uralic languages. This year's workshop builds on the IWCLUL tradition of interdisciplinary collaboration, knowledge sharing, and a common commitment to addressing the linguistic, technical, and social challenges related to Uralic languages in the digital age.

The Uralic language family, spanning across Europe and Asia and encompassing languages as diverse as Finnish, Hungarian and the endangered Udmurt and Khanty languages, presents unique computational challenges. Many of these languages are characterized by complex morphology, agglutinative structures, and unique syntactic and phonological systems, requiring tailored approaches in computational processing and linguistic modeling. Our workshop seeks to bring attention to these challenges and foster the development of innovative solutions that not only support these languages' use in digital contexts but also contribute to their preservation and growth.

This year, IWCLUL received a record number of submissions from researchers worldwide, reflecting the growing interest and engagement in computational approaches to Uralic languages. The selected papers cover a broad spectrum of topics covering both well-studied and lesser-resourced Uralic languages. The diversity of contributions highlights the continuous evolution of the field and the range of challenges being tackled by the computational linguistics community.

We hope that these proceedings inspire continued research and collaboration in computational linguistics for Uralic languages. May the insights, methodologies, and resources shared here contribute to meaningful advances in the field and foster an inclusive future for Uralic languages in the digital landscape.

Sincerely, The IWCLUL 2024 Organizing Committee

Organizing Committee

Organizers

Mika Hämäläinen, Metropolia University of Applied Sciences
Flammie Pirinen, Arctic University of Norway
Melany Macias, Metropolia University of Applied Sciences
Mario Crespo Avila, Complutense University of Madrid

Program Committee

Laszlo Fejes, Hungarian Research Centre for Linguistics

Heiki-Jaan Kaalep, University of Tartu
Gunta Kļava, University of Latvia
Oleg Belyaev, Lomonosov Moscow State University
Trond Trosterud, The Arctic University of Norway
Linda Wiechetek, The Arctic University of Norway
Khalid Alnajjar, F-Secure Oyj
Niko Partanen, University of Helsinki
Jack Rueter, University of Helsinki
Miikka Silfverberg, University of British Columbia
Janne Kauttonen, Haaga-Helia University of Applied Sciences
Michael Rießler, University of Eastern Finland
Aleksei Dorkin, University of Tartu
Jeremy Bradley, University of Vienna
Xinqiao Zhang, UC San Diego
Irina Khomchenkova, Lomonosov Moscow State University
David Dale, Meta
Timofey Arkhangelskiy, University of Hamburg
Viktor Martinović, University of Vienna

Table of Contents

Program

Thursday, November 28, 2024

10:00 - 10:10 *Workshop Opening*

10:10 - 11:00 *Lightning Talks*

11:00 - 12:00 *Oral Session 1*

Aspect Based Sentiment Analysis of Finnish Neighborhoods: Insights from Suomi24
Laleh Davoodi, Anssi Öörni and Ville Harkke

Political Stance Detection in Estonian News Media
Lauri Lüüsi, Uku Kangur, Roshni Chakraborty and Rajesh Sharma

Scaling Sustainable Development Goal Predictions across Languages: From English to Finnish
Melany Macias, Lev Kharlashkin,, Leo Huovinen and Mika Hämäläinen

12:00 - 13:00 *Lunch*

13:00 - 14:20 *Oral Session 2*

Multilingual Approaches to Sentiment Analysis of Texts in Linguistically Diverse Languages: A Case Study of Finnish, Hungarian, and Bulgarian
Mikhail Krasitskii, Olga Kolesnikova, Liliana Chanona Hernandez, Grigori Sidorov and Alexander Gelbukh

Towards standardized inflected lexicons for the Finnic languages
Jules Bouton

DAG: Dictionary-Augmented Generation for Disambiguation of Sentences in Endangered Uralic Languages using ChatGPT
Mika Hämäläinen

Keeping Up Appearances—or how to get all Uralic languages included into bleeding edge research and software: generate, convert, and LLM your way into multilingual datasets
Flammie A Pirinen

14:20 - 14:40 *Coffee Break*

Thursday, November 28, 2024 (continued)

14:40 - 16:00 *Oral Session 3*

Towards the speech recognition for Livonian
Valts Ernštreits

Using Large Language Models to Transliterate Endangered Uralic Languages
Niko Partanen

Specialized Monolingual BPE Tokenizers for Uralic Languages Representation in Large Language Models
Iaroslav Chelombitko and Aleksey Komissarov

Leveraging Transformer-Based Models for Predicting Inflection Classes of Words in an Endangered Sami Language
Khalid Alnajjar, Mika Hämäläinen and Jack Rueter

Friday, November 29, 2024

10:00 - 11:00 *Keynote*

11:00 - 12:00 *Oral Session 4*

Compressing Noun Phrases to Discover Mental Constructions in Corpora – A Case Study for Auxiliaries in Hungarian
Balázs Indig and Tímea Borbála Bajzát

On the Role of New Technologies in the Documentation and Revitalization of Uralic Languages of Russia in Historical and Contemporary Contexts
Alexander Nazarenko

Applying the transformer architecture on the task of headline selection for Finnish news texts
Maria Adamova and Maria Khokhlova

12:00 - 13:00 *Lunch*

13:00 - 14:20 *Oral Session 5*

Kola Saami Christian Text Corpus
Michael Rießler

On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s
Aleksei Dorkin, Taido Purason and Kairit Sirts

Universal-WER: Enhancing WER with Segmentation and Weighted Substitution for Varied Linguistic Contexts
Samy Ouzerrout

On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s
Jack Rueter, Olga Erina and Nadezhda Kabaeva

14:20 - 14:40 *Coffee Break*

14:40 - 15:40 *SIGUR Business Meeting*