LREC-COLING 2024

The 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia @LREC-COLING-2024 (EURALI)

Workshop Proceedings

Editors
Atul Kr. Ojha, Sina Ahmadi, Silvie Cinková, Theodorus Fransen, Chao-Hong Liu and John P. McCrae

25 May, 2024
Torino, Italia

Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia @LREC-COLING-2024 (EURALI)

ISBN 978-2-493814-33-3
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Introduction

Being the largest continental area on Earth, Eurasia is a hub of more than 2,018 languages from very diverse language families and sub-families, including Afro-Asiatic (Semitic), Austroasiatic, Caucasian, Chukchi-Kamchatkan, Dravidian, Eskimo–Aleut, Indo-European, Japonic, Koreanic, Mongolic, Nivkh, Sino-Tibetan, Tai-Kradai, Turkic, Tungusic, Uralic, and Yeniseian. At the same time, various language communities in Eurasia are under-represented, minoritized, endangered and systematically politically oppressed. Many languages, including Kurdish, Gilaki, Santali, Kashmiri, Laz, and Abkhaz, suffer from a lack of linguistic resources and thus are immediately at risk of digital extinction. Others, such as Shabaki, Talysh, Domari, Korbet, and Bawm, are under-researched in general and run the risk of vanishing completely in the absence of increased support.

Aligned with the pressing need to cultivate language technology for indigenous, endangered, and under-resourced languages across Eurasia, the EURALI workshop is dedicated to catalyzing the development of resources and tools. Our objective is to enhance visibility and foster research for these languages on a global scale. We view the current rapid advancements in language and speech technology, particularly the remarkable progress in large language models, as a unique opportunity for these languages. Moreover, by fostering collaboration among researchers, language experts, and linguists engaged with endangered languages within these communities, our aim is to forge language technology solutions that contribute to the preservation of these languages and elevate their prominence within the realm of language processing.

This year, the EURALI workshop returns for its second edition, set against the vibrant backdrop of LREC-COLING 2024. It offers a thrilling opportunity for our community to reconnect and synergize efforts. However, the presence of numerous concurrent workshops has had a modest impact on our submission numbers compared to EURALI's debut at LREC 2022. The eight selected submissions nonetheless encompass a wide array of aspects and challenges within language technology for Eurasian languages as a whole, with a particular focus on Mambai, Standard Tibetan, Persian, Cantonese, and Khroskyabs.

We extend our gratitude to colleagues who submitted their work to the workshop, the organizers of LREC-COLING 2024, and our dedicated and diligent reviewers; your contributions and support have been vital in making the second EURALI workshop a resounding success.

Workshop Chairs

Atul Kr. Ojha, Sina Ahmadi, Silvie Cinková, Theodorus Fransen, Chao-Hong Liu and John P. McCrae

Workshop Chairs

Atul Kr. Ojha, University of Galway, Galway (Ireland)
Sina Ahmadi, University of Zurich, Zurich (Switzerland)
Silvie Cinková, Charles University, Prague (Czech Republic)
Theodorus Fransen, Università Cattolica del Sacro Cuore, Milan (Italy)
Chao-Hong Liu, Potamu Research Ltd, Dublin (Ireland)
John P. McCrae, University of Galway, Galway (Ireland)

Program Committee:

Abigail Walsh, Dublin City University, Dublin (Ireland)
Aiala Rosá, Universidad de la República - Uruguay, Montevideo (Uruguay)
A. Seza Doğruöz, Ghent University, Ghent (Belgium)
Alina Karakanta, University of Leiden, Leiden (Netherlands)
Alina Wróblewska, Institute of Computer Science, Jana Kazimierza, Warszawa (Poland)
Bogdan Babych, Heidelberg University, Heidelberg (Germany)
Çağrı Çöltekin, University of Tübingen, Tübingen (Germany)
Chao-Hong Liu, Potamu Research Ltd, Dublin (Ireland)
Chihiro Taguchi, the University of Notre Dame, Notre Dame (USA)
Daan van Esch, Google, Amsterdam (Netherlands)
Daniel Zeman, Charles University, Prague (Czech Republic)
Deepak Alok, IIT-Delhi, Delhi (India)
Ekaterina Vylomova, University of Melbourne, Melbourne (Australia)
Elizabeth Sherly, Kerala University of Digital Sciences, Innovation and Technology (India)
George Rehm, DFKI GmbH, Berlin (Germany)
Hiwa Asadpour, Goethe University, Frankfurt (Germany)
Joakim Nivre, Uppsala University, Uppsala (Sweden)
John E. Ortega, New York University (USA)
John P. McCrae, University of Galway, Galway (Ireland)
Jonathan Washington, Swarthmore College, Swarthmore (USA)
Joseph Mariani, LIMSI-CNRS, Paris (France)
Kaja Dobrovoljc, University of Ljubljana, Ljubljana (Slovenia)
Khalid Choukri, ELDA/ELRA, Paris (France)
Luke D. Gessler, University of Colorado at Boulder (USA)
Maitrey Mehta, University of Utah, Utah (USA)
Marie-Catherine de Marneffe, Université catholique de Louvain, Louvain (Belgium)
Mayank Jobanputra, University of Tübingen, Tübingen (Germany)
Olesea Caftanatov, Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chişinău (Moldova)
Ranka Stanković, University of Belgrade, Belgrade (Serbia)
Rico Sennrich, University of Zurich, Zurich (Switzerland)
Ritesh Kumar, Agra University, Agra (India)
Rute Costa, the Universidade NOVA de Lisboa, Lisbon (Portugal)
Saliha Muradoglu, Australian National University, Canberra (Australia)
Sarah Moeller, University of Florida, Gainesville, FL (USA)
Silvie Cinkovà, Charles University, Prague (Czech Republic)
Sina Ahmadi, University of Zurich, Zurich (Switzerland)
Stella Markantonatou, Athena RC, Athens (Greece)
Sourabrata Mukherjee, Charles University, Prague (Czech Republic)
Sylvain Kahane, University Paris Nanterre (France)

Valentin Malykh, MTS AI / ITMO University
Verginica Barbu Mititelu, Research Institute for Artificial Intelligence, Bucharest (Romania)
Victoria Bobicev, University of Moldova, Chișinău (Moldova)
Voula Giouli, Institute for Language and Speech Processing, Athens (Greece)

Table of Contents

Conference Program

Saturday, May 25, 2024

09:00–10:05 Inaugural Session

09:00–09:10 ***Welcome***

09:10–10:05 *Keynote talk*
TBD

10:05–10:30 Oral Session-I

10:05–10:30 *Low-Resource Machine Translation through Retrieval-Augmented LLM Prompting: A Study on the Mambai Language*
Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo and Ekaterina Vylomova

10:30–11:00 Coffee break and Poster Session

10:30–11:00 *Improved Neural Word Segmentation for Standard Tibetan*
Collin J. Brown

10:30–11:00 *Open Text Collections as a Resource for Doing NLP with Eurasian Languages*
Sebastian Nordhoff, Christian Döhler and Mandana Seyfeddinipur

10:30–11:00 *The Extraction and Fine-grained Classification of Written Cantonese Materials through Linguistic Feature Detection*
Chaak-ming Lau, Mingfei Lau and Ann Wai Huen To

10:30–11:00 *Neural Mining of Persian Short Argumentative Texts*
Mohammad Yeghaneh Abkenar and Manfred Stede

Saturday, May 25, 2024 (continued)

11:00–12:15 Oral Session-II

11:00–11:25 *Endangered Language Preservation: A Model for Automatic Speech Recognition Based on Khroskyabs Data*
Ruiyao Li and Yunfan Lai

11:25–11:50 *This Word Mean What: Constructing a Singlish Dictionary with ChatGPT*
Siew Yeng Chow, Chang-Uk Shin and Francis Bond

11:50–12:15 *An Evaluation of Language Models for Hyperpartisan Ideology Detection in Persian Twitter*
Sahar Omidi Shayegan, Isar Nejadgholi, Kellin Pelrine, Hao Yu, Sacha Levy, Zachary Yang, Jean-François Godbout and Reihaneh Rabbany

12:15–13:00 Panel Discussion

12:55–13:00 *Valedictory Session*
Workshop Chairs