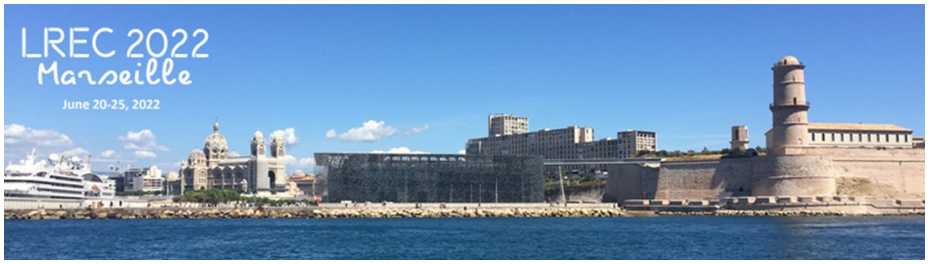

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)

PROCEEDINGS

Editors:
Atul Kr. Ojha, Sina Ahmadi, Chao-Hong Liu, John P. McCrae

Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI 2022)

Edited by:
Atul Kr. Ojha, Sina Ahmadi, Chao-Hong Liu, John P. McCrae

ISBN: 978-2-493814-07-4
EAN: 9782493814074

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Preface

Eurasia is the largest continental area comprising all of Europe and Asia. It is also home to seven families of more than 2,500 languages spoken. Despite the rich linguistic diversity in this area, the respective language communities are under-represented while their languages are low-resource, endangered and/or systematically politically oppressed in history. Others, such as Kurdish, Gilaki, Santali, Kashmiri, Laz, and Abkhaz, are not only endangered but also understudied. One interesting characteristic of these languages is the influence of communal languages on their lexicon through borrowed words and a partially shared vocabulary of phylogenetically related words (cognates). Furthermore, contact-induced similarities can be observed to some extent even in the syntax of the languages, despite typological differences across different language families. In addition, relying on a lingua franca, many of these linguistic communities are facing standardization issues, particularly in the written form of their respective languages. This commonly results in the use of other scripts by speakers of these under-resourced languages.

In line with the necessity of language technology for under-resourced and understudied languages, this workshop aims to spur the development of resources and tools for indigenous, endangered and lesser-resourced languages in Eurasia. The goal is to increase visibility and promote research for these languages in a global arena. Through collaboration between NLP researchers, language experts and linguists working for endangered languages in these communities, we aim to create language technology that will help to preserve these languages and give them a chance to receive more attention in the language processing realm.

Seeing that this is the first edition of the EURALI workshop, we are very happy to have received many submissions, on various aspects regarding Eurasian languages. In the EURALI 2022 Proceedings, 18 research papers are included, dealing with no fewer than 18 Eurasian languages. We would like to thank all the colleagues who submitted their work to the workshop, the LREC 2022 organisers, as well as reviewers for making the first EURALI workshop a success.

Workshop Chairs

Atul Kr. Ojha, Sina Ahmadi, Chao-Hong Liu and John P. McCrae

Chairs

Atul Kr. Ojha, National University of Ireland Galway, Ireland
Sina Ahmadi, National University of Ireland Galway, Ireland
Chao-Hong Liu, Potamu Research Ltd, Ireland
John P. McCrae, National University of Ireland Galway, Ireland

Program Committee:

Agata Savary, University of Paris-Saclay, France
Alina Karakanta, Fondazione Bruno Kessler (FBK) / University of Trento
Akanksha Bansal, Panlingua Language Processing LLP
Atul Kr. Ojha, National University of Ireland Galway, Ireland & Panlingua Language Processing LLP
Bharathi Raja Chakravarthi, National University of Ireland Galway, Ireland
Bogdan Babych, Heidelberg University, Germany
Chao-Hong Liu, Potamu Research Ltd
Daan van Esch, Google
Daniel Zeman, Charles University, Prague
Deepak Alok, Panlingua Language Processing LLP
Dorothee Beermann, Norwegian University of Science and Technology (NTNU)
Esha Banerjee, Google
Ekaterina Vylomova, University of Melbourne, Australia
George Rehm, DFKI GmbH, Germany
Jamal Abdul Nasir, National University of Ireland Galway, Ireland
John Ortega, New York University, USA
Jonathan Washington, Swarthmore College, USA
John P. McCrae, National University of Ireland Galway, Ireland
Joseph Mariani, LIMSI-CNRS, France
Katharina Kann, University of Colorado at Boulder, USA
Kevin Patrick Scannell, Saint Louis University
Khalid Choukri, ELDA/ELRA, France
Massimo Monaglia, University of Florence, Italy
Nicoletta Calzolari, CNR-ILC, Italy
Richard Sproat, Google, Japan
Rico Sennrich, University of Zurich, Switzerland
Ritesh Kumar, Agra University, India
Saliha Muradoglu, Australian National University, Australia
Sina Ahmadi, National University of Ireland Galway, Ireland
Sourabrata Mukherjee, Charles University, Prague
Sunipa Dev, Google
Theodorus Fransen, National University of Ireland Galway, Ireland
Valentin Malykh, Huwaei Norah's Ark lab

Table of Contents

Conference Program

Monday, June 20, 2022

09:00–10:00 Inagural Session

09:00–09:10 *Welcome*
Workshop Chairs

09:10–10:00 *Keynote talk*
Dr. Jonathan Washington

10:00–10:30 Oral Session-I

10:00–10:30 *NLP Pipeline for Annotating (Endangered) Tibetan and Newar Varieties*
Christian Faggionato, Nathan Hill and Marieke Meelen

10:30–11:00 Coffee break/Poster and Demo session

10:30–11:00 *Towards an Ontology for Toponyms in Nepalese Historical Documents*
Sabine Tittel

10:30–11:00 *Semiautomatic Speech Alignment for Under-Resourced Languages*
Juho Leinonen, Niko Partanen, Sami Virpioja and Mikko Kurimo

10:30–11:00 *How to Digitize Completely: Interactive Geovizualization of a Sketch Map from the Kuzmina Archive*
Elena Lazarenko and Aleksandr Riaposov

10:30–11:00 *Word Class Based Language Modeling: A Case of Upper Sorbian*
Isidor Maier, Johannes Kuhn, Frank Duckhorn, Ivan Kraljevski, Daniel Sobe, Matthias Wolff and Constanze Tschöpe

10:30–11:00 *Bringing Together Version Control and Quality Assurance of Language Data with LAMA*
Aleksandr Riaposov, Elena Lazarenko and Timm Lehmberg

10:30–11:00 *Automatic Verb Classifier for Abui (AVC-abz)*
Frantisek Kratochvil, George Saad, Jiří Vomlel and Václav Kratochvíl

Monday, June 20, 2022 (continued)

10:30–11:00 *Dialogue Act and Slot Recognition in Italian Complex Dialogues*
Irene Sucameli, Michele De Quattro, Arash Eshghi, Alessandro Suglia and Maria Simi

10:30–11:00 *Digital Resources for the Shughni Language*
Yury Makarov, Maksim Melenchenko and Dmitry Novokshanov

10:30–11:00 *German Dialect Identification and Mapping for Preservation and Recovery*
Aynalem Tesfaye Misganaw and Sabine Roller

10:30–11:00 *Exploring Transfer Learning for Urdu Speech Synthesis*
Sahar Jamal, Sadaf Abdul Rauf and Quratulain Majid

10:30–11:00 *Towards Bengali WordNet Enrichment using Knowledge Graph Completion Techniques*
Sree Bhattacharyya and Abhik Jana

10:30–11:00 *Enriching Hindi WordNet Using Knowledge Graph Completion Approach*
Sushil Awale and Abhik Jana

10:30–11:00 *A Digital Swedish-Yiddish/Yiddish-Swedish Dictionary: A Web-Based Dictionary that is also Available Offline*
Magnus Ahltorp, Jean Hessel, Gunnar Eriksson, Maria Skeppstedt and Rickard Domeij

10:30–11:00 *An Online Dictionary for Dialects of North Frisian*
Michael Wehar and Tanno Hüttenrauch

10:30–11:00 *Towards a Unified Tool for the Management of Data and Technologies in Field Linguistics and Computational Linguistics - LiFE*
Siddharth Singh, Ritesh Kumar, Shyam Ratan and Sonal Sinha

Monday, June 20, 2022 (continued)

11:10–11:50 Panel Discussion

11:50–12:50 Oral Session-II

11:50–12:20 *Universal Dependencies Treebank for Tatar: Incorporating Intra-Word Code-Switching Information*
Chihiro Taguchi, Sei Iwata and Taro Watanabe

12:20–12:50 *Preparing an endangered language for the digital age: The Case of Judeo-Spanish*
Alp Öktem, Rodolfo Zevallos, Yasmin Moslem, Özgür Güneş Öztürk and Karen Gerson Şarhon

12:50–13:00 ***Valedictory Session***