LREC-COLING 2024

The 3rd Annual Meeting of the ELRA-ISCA Special Interest Group on Under-resourced Languages @LREC-COLING-2024 (SIGUL 2024)

Workshop Proceedings

Editors
Maite Melero, Sakriani Sakti, Claudia Soria

21-22 May, 2024
Turin, Italy

Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @LREC-COLING-2024

ISBN 978-2-493814-29-6
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

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

Message from the Workshop Chairs

Language is a fundamental aspect of human culture and expression, yet not all languages receive equal attention in the realms of research and technological development. Many languages, often referred to as under-resourced languages, lack the necessary linguistic resources and tools to fully harness the potential of modern computational and natural language processing technologies.

The Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL2024) at LREC-COLING 2024 serve as a testament to the growing awareness and commitment within the research community to address the challenges faced by these languages. This workshop aims at providing a platform for researchers, practitioners, and stakeholders to come together, share insights, and collaborate on innovative solutions to empower technological uptake for all languages equally.

In these proceedings, you will find a collection of papers that explore various facets of under-resourced languages, including data collection, annotation, machine learning techniques, and applications in fields such as machine translation, speech recognition, and information retrieval. Each contribution represents a step forward in the collective effort to bridge the digital divide and ensure linguistic diversity is preserved and celebrated in the digital age.

We extend our gratitude to the authors for their valuable contributions and to the workshop reviewers and participants for their dedication and enthusiasm. It is our hope that the insights shared in these proceedings will inspire continued research and advocacy for the inclusion and empowerment of under-resourced languages worldwide.

Maite Melero, Sakriani Sakti, Claudia Soria

Organizing Committee

Maite Melero, Barcelona Supercomputing Center, Spain
Sakriani Sakti, JAIST, Japan
Claudia Soria, CNR-ILC, Italy

Program Committee

Manex Aguirrezabal, University of Copenhagen – Center for Sprogteknologi | Center for Language Technology, Denmark
Sina Ahmadi, University of Zurich, Switzerland
Begoña Altuna, Euskal Herriko Unibertsitatea | University of the Basque Country, Spain
Raghu Annasamy, Google, USA
Antti Arppe, University of Alberta, Canada
Bal Krishna Bal, Kathmandu University, Nepal
Martin Benjamin, Kamusi Project International
Delphine Bernhard, Université de Strasbourg, France
Steven Bird, Charles Darwin University, Australia
Frederic Blum, Max-Planck Institute for Evolutionary Anthropology, Germany
Bharathi Raja Chakravarthi, University of Galway, Ireland
Rajani Chulyadyo, Kathmandu University, Nepal
Joseph Coffey, Harvard University, USA
Matt Coler, University of Groningen, Campus Fryslân, The Netherlands
Anne Dagnac, Université Toulouse Jean Jaurès, France
Arijit Das, Jadavpur University, India
A. Seza Doğruöz, Universiteit Gent, België | Ghent University, Belgium
Louis Estève, LISN, CNRS, Paris-Saclay University, France
Stefano Ghazzali, Language Technologies Unit, Prifysgol Bangor | Bangor University, UK
Itziar Gonzalez-Dios, HiTZ Basque Center for Language Technologies – Ixa, University of the Basque Country, UPV/EHU, Spain
Salima Harrat, Ecole Normale Supérieure Bouzaréah, Algeria
Lars Hellan, Norwegian University of Science and Technology, Norway
Brandi Hongell, University of Groningen, The Netherlands
Mohammad Ali Hussiny, University of Oslo (UIO), Norway
Latifa Iben Nasr, University of Sfax, Faculty of Economics and Management, Tunisia
Martin Jansche, DeepL, UK
Seunghyun Ji, Ahancompany Ltd., South Korea
Mélanie Jouitteau, IKER, CNRS, France
Oleg Kapanadze, OK.OMPLEX–Information and Language Technologies, Georgia
Ritesh Kumar, UnReaL-TecE LLP, India
Elmurod Kuriyozov, Urgench State University, Uzbekistan
Diana Lavrinovic, language activist, Lithuania
Yi Lei, University of Groningen, The Netherlands
Gina-Anne Levow, University of Washington, USA
Chia-Yu Li, University of Stuttgart, Germany
Richard Littauer, unaffiliated
Xueying Liu, University of Groningen, The Netherlands
Oier Lopez de Lacalle, University of the Basque Country, UPV/EHU, Spain
Crisron Rudolf Lucas, University College Dublin, Ireland
Teresa Lynn, Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
Nina Markl, University of Essex, UK

John P. McCrae, Insight Center for Data Analytics, National University of Ireland Galway, Ireland
Maite Melero, Barcelona Supercomputing Center, Espanya | Spain
Peter Mihajlik, Budapest University of Technology and Economics, Hungary
Alice Millour, Université Paris 8 Vincennes Saint-Denis, France
Win Pa Pa, UCS Yangon, Myanmar
Mohammad Arif Payenda, University of Agder (UiA), Norway
Daniel Pimienta, Observatory of Linguistic and Cultural Diversity on Internet, France
Sandy Ritchie, Google Research
Sakriani Sakti, NAIST, Japan
Nay San, Stanford University, USA
Joshua Schäuble, University of Groningen, The Netherlands
Erin Shi, University of Groningen, The Netherlands
Virach Sornlertlamvanich, Musashino University, Japan
Cantao Su, University of Groningen, The Netherlands
Michelle Terblanche, University of Pretoria, South Africa
Daan Van Esch, Google Research
Menno van Zaanen, South African Centre for Digital Language Resources, South Africa
Alice Vanni, University of Groningen, The Netherlands
Jenifer Vega Rodriguez, Université Grenoble Alpes, France
Thang Vu, University of Stuttgart, Germany
Yinqiu Wang, University of Groningen, The Netherlands
Yilan Wei, University of Groningen, The Netherlands)
Hongchen Wu, Georgia Institute of Technology, USA
Marcely Zanon Boito, NAVER Labs Europe, France

Table of Contents