

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

Copyright ELRA Language Resources Association (ELRA), 2024
These proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

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 Dođruöz, 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
Mélodie 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

<i>A Bit of a Problem: Measurement Disparities in Dataset Sizes across Languages</i> Catherine Arnett, Tyler A. Chang and Benjamin Bergen	1
<i>A Novel Corpus for Automated Sexism Identification on Social Media</i> Lutfiye Seda Mut Altin and Horacio Saggion	10
<i>Advancing Generative AI for Portuguese with Open Decoder Gervásio PT*</i> Rodrigo Santos, João Ricardo Silva, Luís Gomes, João Rodrigues and António Branco	16
<i>Assessing Pre-Built Speaker Recognition Models for Endangered Language Data</i> Gina-Anne Levow	27
<i>BERTbek: A Pretrained Language Model for Uzbek</i> Elmurod Kuriyozov, David Vilares and Carlos Gómez-Rodríguez	33
<i>Beyond Error Categories: A Contextual Approach of Evaluating Emerging Spell and Grammar Checkers</i> Pórunn Arnardóttir, Svanhvít Lilja Ingólfssdóttir, Haukur Barri Símonarson, Hafsteinn Einars- son, Anton Karl Ingason and Vilhjálmur Þorsteinsson	45
<i>Bidirectional English-Nepali Machine Translation(MT) System for Legal Domain</i> Shabdapurush Poudel, Bal Krishna Bal and Praveen Acharya	53
<i>BK3AT: Bangsamoro K-3 Children’s Speech Corpus for Developing Assessment Tools in the Bangsamoro Languages</i> Kiel D. Gonzales, Jazzmin R. Maranan, Francis Paolo D. Santelices, Edsel Jedd M. Reno- valles, Nissan D. Macale, Nicole Anne A. Palafox and Jose Marie A. Mendoza	59
<i>CorpusArièja: Building an Annotated Corpus with Variation in Occitan</i> Clamenca Poujade, Myriam Bras and Assaf Urieli	66
<i>Developing Infrastructure for Low-Resource Language Corpus Building</i> Hedwig G. Sekeres, Wilbert Heeringa, Wietse de Vries, Oscar Yde Zwagers, Martijn Wiel- ing and Goffe Th. Jensma	72
<i>Evaluating Icelandic Sentiment Analysis Models Trained on Translated Data</i> Ólafur A. Jóhannsson, Birkir H. Arndal, Eysteinn Ö. Jónsson, Stefan Olafsson and Hrafn Loftsson	79
<i>Exploring Text Classification for Enhancing Digital Game-Based Language Learning for Irish</i> Leona Mc Cahill, Thomas Baltazar, Sally Bruen, Liang Xu, Monica Ward, Elaine Uí Dhon- nchadha and Jennifer Foster	90
<i>Forget NLI, Use a Dictionary: Zero-Shot Topic Classification for Low-Resource Languages with Application to Luxembourgish</i> Fred Philippy, Shohreh Haddadan and Siwen Guo	97
<i>Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family</i> Rodrigo Santos, João Rodrigues, Luís Gomes, João Ricardo Silva, António Branco, Hen- rique Lopes Cardoso, Tomás Freitas Osório and Bernardo Leite	105

<i>Improving Language Coverage on HeLI-OTS</i> Tommi Jauhiainen and Krister Lindén	115
<i>Improving Legal Judgement Prediction in Romanian with Long Text Encoders</i> Mihai Masala, Traian Rebedea and Horia Velicu	126
<i>Improving Noisy Student Training for Low-resource Languages in End-to-End ASR Using CycleGAN and Inter-domain Losses</i> Chia-Yu Li and Ngoc Thang Vu	133
<i>Indonesian-English Code-Switching Speech Recognition Using the Machine Speech Chain Based Semi-Supervised Learning</i> Rais Vaza Man Tazakka, Dessi Lestari, Ayu Purwarianti, Dipta Tanaya, Kurniawati Azizah and Sakriani Sakti	143
<i>Inter-language Transfer Learning for Visual Speech Recognition toward Under-resourced Environments</i> Fumiya Kondo and Satoshi Tamura	149
<i>Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study</i> Wan-hua Her and Udo Kruschwitz	155
<i>Italian-Ligurian Machine Translation in Its Cultural Context</i> Christopher R. Haberland, Jean Maillard and Stefano Lusito	168
<i>Labadain-30k+: A Monolingual Tetun Document-Level Audited Dataset</i> Gabriel de Jesus and Sérgio Nunes	177
<i>Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining</i> Nikola Ljubešić, Vít Suchomel, Peter Rupnik, Taja Kuzman and Rik van Noord	189
<i>Man or Machine: Evaluating Spelling Error Detection in Danish Newspaper Corpora</i> Eckhard Bick, Jonas Nygaard Blom, Marianne Rathje and Jørgen Schack	204
<i>Managing Fine-grained Metadata for Text Bases in Extremely Low Resource Languages: The Cases of Two Regional Languages of France</i> Marianne Vergez-Couret, Delphine Bernhard, Michael Nauge, Myriam Bras, Pablo Ruiz Fabo and Carole Werner	212
<i>Mixat: A Data Set of Bilingual Emirati-English Speech</i> Maryam Khalifa Al Ali and Hanan Aldarmaki	222
<i>Multi-dialectal ASR of Armenian from Naturalistic and Read Speech</i> Malajyan Arthur, Victoria Khurshudyan, Karen Avetisyan, Hossep Dolatian and Damien Nouvel	227
<i>Multilingual Self-supervised Visually Grounded Speech Models</i> Huynh Phuong Thanh Nguyen and Sakriani Sakti	237
<i>Nepal Script Text Recognition Using CRNN CTC Architecture</i> Swornim Nakarmi, Sarin Sthapit, Arya Shakya, Rajani Chulyadyo and Bal Krishna Bal244	

<i>NLP for Arbëresh: How an Endangered Language Learns to Write in the 21st Century</i> Giulio Cusenza and Çağrı Çöltekin.....	252
<i>PersianEmo: Enhancing Farsi-Dari Emotion Analysis with a Hybrid Transformer and Recurrent Neural Network Model</i> Mohammad Ali Hussiny, Mohammad Arif Payenda and Lilja Øvrelid	257
<i>Philippine Languages Database: A Multilingual Speech Corpora for Developing Systems for Low-Resource Languages</i> Rowena Cristina L. Guevara, Rhandley D. Cajote, Michael Gringo Angelo R. Bayona and Crisron Rudolf G. Lucas	264
<i>Prompting towards Alleviating Code-Switched Data Scarcity in Under-Resourced Languages with GPT as a Pivot</i> Michelle Terblanche, Kayode Olaleye and Vukosi Marivate	272
<i>Quantifying the Ethical Dilemma of Using Culturally Toxic Training Data in AI Tools for Indigenous Languages</i> Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin and Julio Nogima	283
<i>Residual Dropout: A Simple Approach to Improve Transformer's Data Efficiency</i> Carlos Escolano, Francesca De Luca Fornaciari and Maite Melero	294
<i>Resource Acquisition for Understudied Languages: Extracting Wordlists from Dictionaries for Computer-assisted Language Comparison</i> Frederic Blum, Johannes Englisch, Alba Hermida Rodriguez, Rik van Gijn and Johann-Mattis List	300
<i>Robust Guidance for Unsupervised Data Selection: Capturing Perplexing Named Entities for Domain-Specific Machine Translation</i> Seunghyun Ji, Hagai Raja Sinulingga and Darongsae Kwon.....	307
<i>Seeding Alignment between Language Technology and Indigenous Methodologies: A Decolonizing Framework for Endangered Language Revitalization</i> Craig John Carpenter, John Lyon, Miles Thorogood and Jeannette C. Armstrong.....	318
<i>Solving Failure Modes in the Creation of Trustworthy Language Technologies</i> Gianna Leoni, Lee Steven, Tūreiti Keith, Keoni Mahelona, Peter-Lucas Jones and Suzanne Duncan	325
<i>Tandem Long-Short Duration-based Modeling for Automatic Speech Recognition</i> Dalai Mengke, Yan Meng and Peter Mihajlik.....	331
<i>TELP – Text Extraction with Linguistic Patterns</i> João Cordeiro, Purificação Moura Silvano, António Leal and Sebastião Pais.....	337
<i>The First Parallel Corpus and Neural Machine Translation Model of Western Armenian and English</i> Ari Nubar Boyacıoğlu and Jan Niehues	345

<i>Tracing Linguistic Heritage: Constructing a Somali-Italian Terminological Resource through Explorers' Notebooks and Contemporary Corpus Analysis</i>	
Silvia Piccini, Giuliana Elizabeth Vilela Ruiz, Andrea Bellandi and Enrico Carniani	357
<i>Uncovering Social Changes of the Basque Speaking Twitter Community During COVID-19 Pandemic</i>	
Joseba Fernandez de Landa, Iker García-Ferrero, Ander Salaberria and Jon Ander Campos	363
<i>UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology</i>	
Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olessea Căftanatu, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz and Alina Wróblewska	372
<i>Unsupervised Outlier Detection for Language-Independent Text Quality Filtering</i>	
Jón Daðason and Hrafn Loftsson	383
<i>UzABSA: Aspect-Based Sentiment Analysis for the Uzbek Language</i>	
Sanatbek Gayratovich Matlatipov, Jaloliddin Rajabov, Elmurod Kuriyozov and Mersaid Aripov	394
<i>ViHealthNLI: A Dataset for Vietnamese Natural Language Inference in Healthcare</i>	
Huyen Nguyen, Quyen The Ngo, Thanh-Ha Do and Tuan-Anh Hoang	404
<i>Why the Unexpected? Dissecting the Political and Economic Bias in Persian Small and Large Language Models</i>	
Ehsan Barkhordar, Surendrabikram Thapa, Ashwarya Maratha and Usman Naseem .	410
<i>Work in Progress: Text-to-speech on Edge Devices for Te Reo Māori and 'Ōlelo Hawai'i</i>	
Tūreiti Keith	421