



LREC 2026

**SIGUL 2026 Joint Workshop with ELE, EURALI, and
DCLRL:
Towards Inclusivity and Equality: Language Resources
and Technologies for Under-Resourced and
Endangered Languages @ LREC 2026**

Workshop Proceedings

Editors

**Sakriani Sakti, Claudia Soria, Maite Melero, Atul Kr
Ojha, John P. McCrae, Constantine Lignos, Chao-Hong
Liu, German Rigau Claramunt, Georg Rehm**

11–12 May, 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-79-1
EAN 9782493814791

Preface

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 SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL "Towards Inclusivity and Equality: Language Resources and Technologies for Under-Resourced and Endangered Languages at LREC 2026 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 to provide 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. We hope that the insights shared in these proceedings will inspire continued research and advocacy for the inclusion and empowerment of under-resourced languages worldwide.

Organizing Committee

- Sakriani Sakti, NAIST, Japan
- Claudia Soria, CNR-ILC, Italy
- Maite Melero, Barcelona Supercomputing Center, Spain
- Atul Kr. Ojha, University of Galway, Ireland
- John P. McCrae, University of Galway, Ireland
- Constantine Lignos, Brandeis University, USA
- Chao-Hong Liu, Potamu Research Ltd, Ireland and ITRI, Taiwan
- German Rigau Claramunt, HiTZ Center for Language Technology, University of the Basque Country, Spain
- Georg Rehm, DFKI and Humboldt-Universität zu Berlin, Germany

Program Committee

- Adrian Doyle, University of Galway, Ireland
- Ailbhe Ní Chasaide, Trinity College Dublin, Ireland
- Ali Hatami, University of Galway, Ireland
- Andras Kornai, HAS Computer and Automation Research Institute, Hungary
- Andy Murphy, Trinity College Dublin, Ireland
- Atul Kr. Ojha, Data Science Institute, Unit for Linguistic Data, University of Galway, Ireland
- Bal Krishna Bal, Department of Computer Science and Engineering, Kathmandu University, Nepal, Nepal
- Bogdan Babych, Heidelberg University, Germany
- Chao-Hong Liu, ITRI, Potamu Research, Ireland
- Chihiro Taguchi, University of Notre Dame, United States
- Constantine Lignos, Brandeis University, United States
- Crisron Rudolf G, Lucas, University College Dublin, Ireland
- Daan van Esch, Google Research, Netherlands
- Dafydd Gibbon, Universität Bielefeld, Germany
- Daniel Pimienta, Observatory of Linguistic & Cultural Diversity in the Internet -, France
- Deepak Alok, IIT-Delhi, India
- Delphine Bernhard, Lilpa, Université de Strasbourg, France

- Dewi Bryn Jones, Language Technologies Unit, Canolfan Bedwyr, Bangor University, United Kingdom
- Diana Lavrinovic, Language activist, researcher, Lithuania
- Ekaterina Vylomova, University of Melbourne, Australia
- Francis Bond, Palacky University, Czech Republic
- Georg Rehm, DFKI, Germany
- German Rigau, UPV/EHU, Spain
- Gina-Anne Levow, University of Washington, United States
- Hongchen Wu, Georgia Institute of Technology, United States
- Jenifer Andrea Vega Rodriguez, GIPSA-lab, France
- John E. Ortega, Northeastern University, United States
- John P. McCrae, Insight Centre for Data Analytics, National University of Galway, Ireland
- Joseph Mariani, LIMSI-CNRS, France
- Latifa Iben Nasr, Miracl laboratory, University of Sfax, Tunisia
- Luke Gessler, Indiana University Bloomington, United States
- Maite Melero, BSC, Spain
- Matt Coler, University of Groningen, Netherlands
- Mélanie Jouitteau, CNRS, France
- Menno van Zaanen, South African Centre for Digital Language Resources, South Africa
- Mohammad Arif Payenda, University of Agder (UiA), Norway
- Morgan Grobol, Université Paris Nanterre, France
- Neasa Ní Chiaráin, Trinity College Dublin, Ireland
- Olesea Caftanator, Moldova State University, Vladimir Andrunachievici Institute of Mathematics and Computer Science, Republic of Moldova
- Raghuram Annasamy, Google, United States
- Rajani Chulyadyo, Department of Computer Science and Engineering, Kathmandu University, Nepal, Nepal
- Ranka Stankovic, University of Belgrade - Faculty of Mining and Geology, Serbia
- Richard Sproat, Sakana.ai, Japan
- Rute Costa, CLUNL - FCSH - Universidade NOVA de Lisboa, Portugal
- Sakriani Sakti, Nara Institute of Science and Technology (NAIST), Japan
- Shyam Sunder Agrawal, KIIT College of Engineering, Gurgaon, Advisor, COE-AI- IGDTUW, Delhi, India

- Silvie Cinkova, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czech Republic
- Sourabrata Mukherjee, MSR-India
- Steven Bird, Charles Darwin University, Australia
- Sylvain Kahane, Modyco, University Paris Nanterre & CNRS / Institut Universitaire de France, France
- Trond Trosterud, UiT Norgga árkálaš universitehta, Norway
- Verginica Barbu Mititelu, Research Institute for Artificial Intelligence, Romanian Academy, Romania
- Victoria Bobicev, Technical University of Moldova, Republic of Moldova
- Voula Giouli, Aristotle University of Thessaloniki / ILSP, ATHENA RC, Greece

Table of Contents

<i>How Well Do Large Language Models Reason in Under-Resourced Languages? Evidence from Vietnamese</i>	
Tuan Anh Do and Jelke Bloem	1
<i>Register Sensitivity in Scalar MT Evaluation: Evidence from Spanish–Basque Informal Discourse</i>	
Nora Aranberri	19
<i>Corpus-Linguists’ Little Helpers? Evaluating LLMs for Linguistic Annotation: The Case of Sensationalist Headlines Corpus</i>	
Petra Bago and Virna Karlić	33
<i>LLM as a Morphological Disambiguator for Belarusian: A Preliminary Study</i>	
Vladislav Poritski, Oksana Volchek and Ilia Afanasev	42
<i>Keyboards for the Endangered Idu Mishmi Language</i>	
Akhilesh Kakolu Ramarao	49
<i>SAINT: Multilingual Span-Level Interpretability for Sentiment Analysis</i>	
Seid Muhie Yimam, Tadesse Destaw Belay, Robert Geislinger, Shamsuddeen Hassan Muhammad, Adaeze Ngozi Ohuoba, Sukairaj Hafiz Imam, Abinew Ali Ayele, Martin Semmann, Chris Biemann and Serge Sharoff	54
<i>AlbanianLLMSafety: A Safety Evaluation Dataset for Large Language Models in Albanian</i>	
Wajdi Zaghouani, Kholoud Khalil Aldous and Isra Fejzullaj	66
<i>Urdu-CLEVR: A Novel Benchmark for Visual Reasoning in an Under-Resourced Linguistic Context</i>	
Sohail Ashraf, Adeel Zafar, Slawomir Nowaczyk and Ahthasham Sajid	74
<i>A Database of Romance Clitics With Speech Samples</i>	
Abdelrahim Qaddoumi, Owen Rambow, Lori Repetti and Francisco Ordóñez	83
<i>GreekCommonGen: A Benchmark for Evaluating Generative Commonsense Reasoning in Greek</i>	
Aristotelis Stamopoulos and Dimitrios Galanis	94
<i>Transfer Learning for Creole TTS: A Pilot Study on Whether Substrate Phonologies or Lexifier Vocabularies Matter More</i>	
Emmett Strickland, Marc Evrard and Valentina Fedchenko	101
<i>KZ-SafetyPrompts: A Kazakh Safety Evaluation Prompt Dataset for Large Language Models</i>	
Wajdi Zaghouani, Shimaa Amer Ibrahim, Aruzhan Muratbek, Olzhasbek Zhakenov and Adiya Akhmetzhanova	107
<i>SimLex-999 for Modern Greek</i>	
Leonidas Mylonadis and Jelke Bloem	119
<i>Quality and Appropriateness of Large Text Datasets for Irish NLP</i>	
Abigail Walsh, Mark Andrade, Jane Lauren Adkins, Ornait O’Connell, Éanna O’Connor, Ellen Rushe and Brian Davis	126

<i>BiST: A Gold Standard Bangla-English Bilingual Corpus for Sentence Structure and Tense Classification with Inter-Annotator Agreement</i>	
Abdullah Al Shafi, Swapnil Kundu Argha, M. A. Moyeen, Abdul Muntakim and Shoumik Barman Polok	143
<i>LLM-Assisted Spanish Dialect Corpus Construction</i>	
Jessica Claribel RAMIREZ VIDAL, Hiroki Ouchi and Sakriani Sakti	153
<i>Structured Entity Extraction from Hawaiian Television Chyrons Using Vision-Language Models</i>	
Kelley Lynch, Owen King, Kyeongmin Rim, Gabrielle Keen, Yangyang Chen and James Pustejovsky	160
<i>Towards a general theory of linguistic diversity</i>	
Steven Bird	169
<i>Interlinear Glosses as a Multilingual Pivot for Machine Translation: An Updated Study on Turkish with Restricted Resources</i>	
Volkan Ozer, Shu Okabe and Alexander Fraser	183
<i>Benchmarking Multilingual LLM Translation Accuracy for Fuzhounese</i>	
Sue Zheng and Jelke Bloem	198
<i>Evaluating Nepali NER and POS Tagging Models on the Achhami Dialect</i>	
Samikshya Dhamala, Rishav Beejukchhen, Subresh Thakulla, Bikash Kadayat and Supriya Khadka	210
<i>From LLM Prompts to Acoustic Baselines: A Scalable Pipeline for Under-Resourced Disfluent Code-Mixed Speech</i>	
Anuran Mitra, Anirvan Chakravarty, Tapabrata Mondal and Sivaji Bandyopadhyay	222
<i>Beyond Fine-Tuning: Procrustes Alignment of Multilingual Embeddings for Low-Resource Cross-Lingual Retrieval</i>	
Ali Faheem, Muhammad Hammad, Faizad Ullah, Ahmed Hassan, Fezan Rasool and Asim Karim	233
<i>Language Identification for Low-Resource Formosan Languages</i>	
Henry Gagnier	242
<i>Rebelòt: Datasets and Token-Level Language Identification for Lombard-Italian-English Code-Mixing</i>	
Edoardo Signoroni, Emma Bednaříková and Pavel Rychly	253
<i>Fine-tuning Whisper with Spontaneous Persian Speech (SPS)</i>	
Behnoosh Namdarzadeh and Nicolas Ballier	263
<i>Small Language Models for Less-Resourced Languages in a Real-World Scenario: The Case for Catalan</i>	
Roser Saurí, Josep Sànchez-Ferreres, Lluís Padro and Josep Carmona	270
<i>AmazoniaNLP: A Survey of Extreme Low-Resource Languages in the Peruvian-Brazilian Amazon</i>	
Rodolfo Joel Zevallos, Fabrício Carraro and John E. Ortega	280

AfriVoices-KE: A Multilingual Speech Dataset for Kenyan Languages

Lilian Wanzare, Cynthia Jayne Amol, Ezekiel Maina, Nelson Odhiambo, Hope Kerubo, Leila Misula, Vivian Oloo, Rennish Mboya, Edwin Onkoba, Edward Ombui, Joseph Muguro, Ciira wa Maina, Andrew Kipkebut, Alfred Omondi Otom, Ian Ndung'u Kang'ethe, Angela Wambui Kanyi and Brian Gichana Omwenga 288

HuNeBR: A Multitask Benchmark to Evaluate LLMs' Understanding of Northeastern Brazilian Portuguese Humor

José Matheus do Nascimento Gama, David Candeia Maia, Leandro Balby Marinho, Fabio Morais and João Brunet 299

Open Machine Translation for Esperanto

Ona de Gibert and Lluís de Gibert 312

Workshop Program

11th May 2026

09:00–10:30 **Oral Session 1: LLMs for Linguistic Analysis and Evaluation**

+ ***Welcome and Introduction***

How Well Do Large Language Models Reason in Under-Resourced Languages? Evidence from Vietnamese

Tuan Anh Do and Jelke Bloem

Register Sensitivity in Scalar MT Evaluation: Evidence from Spanish–Basque Informal Discourse

Nora Aranberri

Corpus-Linguists' Little Helpers? Evaluating LLMs for Linguistic Annotation: The Case of Sensationalist Headlines Corpus

Petra Bago and Virna Karlić

LLM as a Morphological Disambiguator for Belarusian: A Preliminary Study

Vladislav Poritski, Oksana Volchek and Ilia Afanasev

10:30–11:00 ***Coffee Break***

11:00–12:00 **Poster Session**

Keyboards for the Endangered Idu Mishmi Language

Akhilesh Kakolu Ramarao

SAINT: Multilingual Span-Level Interpretability for Sentiment Analysis

Seid Muhie Yimam, Tadesse Destaw Belay, Robert Geislinger, Shamsuddeen Hassan Muhammad, Adaeze Ngozi Ohuoba, Sukairaj Hafiz Imam, Abinew Ali Ayele, Martin Semmann, Chris Biemann and Serge Sharoff

AlbanianLLMSafety: A Safety Evaluation Dataset for Large Language Models in Albanian

Wajdi Zaghouani, Kholoud Khalil Aldous and Isra Fejzullaj

Urdu-CLEVR: A Novel Benchmark for Visual Reasoning in an Under-Resourced Linguistic Context

Sohail Ashraf, Adeel Zafar, Slawomir Nowaczyk and Ahthasham Sajid

11th May 2026 (continued)

A Database of Romance Clitics With Speech Samples

Abdelrahim Qaddoumi, Owen Rambow, Lori Repetti and Francisco Ordóñez

GreekCommonGen: A Benchmark for Evaluating Generative Commonsense Reasoning in Greek

Aristotelis Stamopoulos and Dimitrios Galanis

Transfer Learning for Creole TTS: A Pilot Study on Whether Substrate Phonologies or Lexifier Vocabularies Matter More

Emmett Strickland, Marc Evrard and Valentina Fedchenko

KZ-SafetyPrompts: A Kazakh Safety Evaluation Prompt Dataset for Large Language Models

Wajdi Zaghouani, Shimaa Amer Ibrahim, Aruzhan Muratbek, Olzhasbek Zhakenov and Adiya Akhmetzhanova

SimLex-999 for Modern Greek

Leonidas Mylonadis and Jelke Bloem

Quality and Appropriateness of Large Text Datasets for Irish NLP

Abigail Walsh, Mark Andrade, Jane Lauren Adkins, Ornait O’Connell, Éanna O’Connor, Ellen Rushe and Brian Davis

BiST: A Gold Standard Bangla-English Bilingual Corpus for Sentence Structure and Tense Classification with Inter-Annotator Agreement

Abdullah Al Shafi, Swapnil Kundu Argha, M. A. Moyeen, Abdul Muntakim and Shoumik Barman Polok

LLM-Assisted Spanish Dialect Corpus Construction

Jessica Claribel RAMIREZ VIDAL, Hiroki Ouchi and Sakriani Sakti

Structured Entity Extraction from Hawaiian Television Chyrons Using Vision-Language Models

Kelley Lynch, Owen King, Kyeongmin Rim, Gabrielle Keen, Yangyang Chen and James Pustejovsky

11th May 2026 (continued)

12:00–13:00 **Oral Session 2: Rethinking Linguistic Diversity in the Digital Age**

Towards a general theory of linguistic diversity

Steven Bird

13:00–14:00 **Lunch Break**

14:00–15:40 **Oral Session 3: Modeling and Adaptation in Low-Resource Settings**

Interlinear Glosses as a Multilingual Pivot for Machine Translation: An Updated Study on Turkish with Restricted Resources

Volkan Ozer, Shu Okabe and Alexander Fraser

Benchmarking Multilingual LLM Translation Accuracy for Fuzhounese

Sue Zheng and Jelke Bloem

Evaluating Nepali NER and POS Tagging Models on the Achhami Dialect

Samikshya Dhamala, Rishav Beejukchhen, Subresh Thakulla, Bikash Kadayat and Supriya Khadka

From LLM Prompts to Acoustic Baselines: A Scalable Pipeline for Under-Resourced Disfluent Code-Mixed Speech

Anuran Mitra, Anirvan Chakravarty, Tapabrata Mondal and Sivaji Bandyopadhyay

Beyond Fine-Tuning: Procrustes Alignment of Multilingual Embeddings for Low-Resource Cross-Lingual Retrieval

Ali Faheem, Muhammad Hammad, Faizad Ullah, Ahmed Hassan, Fezan Rasool and Asim Karim

16:00–16:30 **Coffee Break**

16:30–17:30 *Invited Talk*

11th May 2026 (continued)

Oral Session 4: Language Identification and Variation

Language Identification for Low-Resource Formosan Languages

Henry Gagnier

Rebelòt: Datasets and Token-Level Language Identification for Lombard-Italian-English Code-Mixing

Edoardo Signoroni, Emma Bednaříková and Pavel Rychly

12th May 2026

9:00–10:30 Oral Session 5: Language Resources and Scalable Systems

Fine-tuning Whisper with Spontaneous Persian Speech (SPS)

Behnoosh Namdarzadeh and Nicolas Ballier

Small Language Models for Less-Resourced Languages in a Real-World Scenario: The Case for Catalan

Roser Saurí, Josep Sànchez-Ferrerres, Lluís Padro and Josep Carmona

AmazoniaNLP: A Survey of Extreme Low-Resource Languages in the Peruvian-Brazilian Amazon

Rodolfo Joel Zevallos, Fabrício Carraro and John E. Ortega

AfriVoices-KE: A Multilingual Speech Dataset for Kenyan Languages

Lilian Wanzare, Cynthia Jayne Amol, Ezekiel Maina, Nelson Odhiambo, Hope Kerubo, Leila Misula, Vivian Oloo, Rennish Mboya, Edwin Onkoba, Edward Ombui, Joseph Muguro, Ciira wa Maina, Andrew Kipkebut, Alfred Omondi Otom, Ian Ndung'u Kang'ethe, Angela Wambui Kanyi and Brian Gichana Omwenga

10:30–11:00 Coffee Break

11:00–12:00 Invited Talk

12th May 2026 (continued)

12:00–13:00 Oral Session 6: Translation, Benchmarks and Emerging Tasks

HuNeBR: A Multitask Benchmark to Evaluate LLMs' Understanding of Northeastern Brazilian Portuguese Humor

José Matheus do Nascimento Gama, David Candeia Maia, Leandro Balby Marinho, Fabio Moraes and João Brunet

Open Machine Translation for Esperanto

Ona de Gibert and Lluís de Gibert

