

MRL 2023

**The 3rd Workshop on Multi-lingual Representation Learning**

**Proceedings of the Workshop**

December 7, 2023

©2023 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
209 N. Eighth Street  
Stroudsburg, PA 18360  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 979-8-89176-056-1

# Organizing Committee

## Organizers

Duygu Ataman, New York University  
Hila Gonen, Meta, University of Washington  
Sebastian Ruder, Google  
David Ifeoluwa Adelani, Google Deepmind and UCL  
Gözde Gül Sahin, Koc University  
Chris Emezue, TU Munich  
Benjamin Muller, Meta  
Omer Goldman, Bar-Ilan University  
Mammad Hajili, Microsoft  
Francesco Tinner, University of Amsterdam  
Genta Indra Winata, Bloomberg

## Program Committee

### Reviewers

Saksham Bassi, New York University  
Jannis Vamvas, University of Zurich  
Ankur Bapna, Google  
Ivan Vulić, University of Cambridge  
Biao Zhang, Google  
Sneha Kudugunta, Google  
Ahmet Ustun, Cohere  
Gozde Gul Sahin, Koc University  
Duygu Ataman, New York University  
Asa Cooper Stickland, New York University  
Jonne Saleva, Brandeis University  
Richard Yuanzhe Pang, New York University  
Genta Winata, Bloomberg  
Abhinav Arora, Meta  
Constantine Lignos, Brandeis University  
Xinyan Yu, University of Southern California  
Antonios Anastasopoulos, George Mason University  
Abdullatif Koksal, LMU Munich  
Holy Lovenia, AISG

# Keynote Talk: Orhan Firat

Orhan Firat  
Google Deepmind  
2023-12-07 09:10 –

**Bio:** Orhan Firat is a senior research scientist at Google Deepmind where he works on cutting-edge technologies on scalable and multi-lingual language models.

# Keynote Talk: Katharina Kann

**Katharina Kann**  
UC Boulder  
**2023-12-07 09:50 –**

**Bio:** Katharina Kann is an assistant professor at UC Boulder and JGU Mainz and her research focuses on building natural language processing systems that work for all of the world's languages.

# Keynote Talk: Sunayana Sitaram

Sunayana Sitaram

Microsoft

2023-12-07 16:00 –

**Bio:** Sunayana Sitaram is a principal researcher at Microsoft Research India. Her research interests are broadly in democratizing AI and making LLMs more inclusive to more languages and cultures.

## Table of Contents

<i>UniBriVL: Robust Audio Representation and Generation of Audio Driven Diffusion Models</i> Sen Fang, Bowen Gao, Yangjian Wu and TeikToe Teoh . . . . .	1
<i>Meta-learning For Vision-and-language Cross-lingual Transfer</i> Hanxu Hu and Frank Keller . . . . .	12
<i>Counterfactually Probing Language Identity in Multilingual Models</i> Anirudh Srinivasan, Venkata Subrahmanyam Govindarajan and Kyle Mahowald . . . . .	24
<i>A General-Purpose Multilingual Document Encoder</i> Onur Galoğlu Robert Litschko, Robert Litschko and Goran Glavaš . . . . .	37
<i>Zero-Shot Cross-Lingual Sentiment Classification under Distribution Shift: an Exploratory Study</i> Maarten De Raedt, Semere Kiros Bitew, Frédéric Godin, Thomas Demeester and Chris Develder 50	
<i>To token or not to token: A Comparative Study of Text Representations for Cross-Lingual Transfer</i> Md Mushfiqur Rahman, Fardin Ahsan Sakib, Fahim Faisal and Antonios Anastasopoulos . . . . .	67
<i>Adapt and Prune Strategy for Multilingual Speech Foundational Model on Low-resourced Languages</i> Hyeon Soo Kim, Chung Hyeon Cho, Hyejin Won and Kyung Ho Park . . . . .	85
<i>Multilingual Word Embeddings for Low-Resource Languages using Anchors and a Chain of Related Languages</i> Viktor Hangya, Silvia Severini, Radoslav Ralev, Alexander Fraser and Hinrich Schütze . . . . .	95
<i>TalaMT: Multilingual Machine Translation for Cabécar-Bribri-Spanish</i> Alex Jones, Rolando Coto-Solano and Guillermo González Campos . . . . .	106
<i>Mergen: The First Manchu-Korean Machine Translation Model Trained on Augmented Data</i> Jean Seo, Sungjoo Byun, Minha Kang and Sangah Lee . . . . .	118
<i>Improving Cross-Lingual Transfer for Open Information Extraction with Linguistic Feature Projection</i> Youmi Ma, Bhushan Kotnis, Carolin Lawrence, Goran Glavaš and Naoaki Okazaki . . . . .	125
<i>Geographic and Geopolitical Biases of Language Models</i> Fahim Faisal and Antonios Anastasopoulos . . . . .	139
<i>Task-Based MoE for Multitask Multilingual Machine Translation</i> Hai Pham, Young Jin Kim, Subhabrata Mukherjee, David P. Woodruff, Barnabas Poczos and Hany Hassan . . . . .	164
<i>Does the English Matter? Elicit Cross-lingual Abilities of Large Language Models</i> Leonardo Ranaldi and Giulia Pucci . . . . .	173
<i>CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages</i> Gabriel Oliveira dos Santos, Diego Alysson Braga Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Da Silva, Esther Colombini, Helio Pedrini and Sandra Avila . . . . .	184
<i>Code-switching as a cross-lingual Training Signal: an Example with Unsupervised Bilingual Embedding</i> Felix Gaschi, Ilias El-Baamrani, Barbara Gendron, Parisa Rastin and Yannick Toussaint . . . . .	208

<i>Learning to translate by learning to communicate</i>	
C.M. Downey, Xuhui Zhou, Zeyu Liu and Shane Steinert-Threlkeld . . . . .	218
<i>Contrastive Learning for Universal Zero-Shot NLI with Cross-Lingual Sentence Embeddings</i>	
Md Kowsher, Md. Shohanur Islam Sobuj, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin and Yasuhiko Morimoto . . . . .	239
<i>UD-MULTIGENRE – a UD-Based Dataset Enriched with Instance-Level Genre Annotations</i>	
Vera Danilova and Sara Stymne . . . . .	253
<i>Embedding Structure Matters: Comparing Methods to Adapt Multilingual Vocabularies to New Languages</i>	
C.M. Downey, Terra Blevins, Nora Goldfine and Shane Steinert-Threlkeld . . . . .	268
<i>Multi-EuP: The Multilingual European Parliament Dataset for Analysis of Bias in Information Retrieval</i>	
Jinrui Yang, Timothy Baldwin and Trevor Cohn . . . . .	282
<i>Generating Continuations in Multilingual Idiomatic Contexts</i>	
Rhitabrat Pokharel and Ameeta Agrawal . . . . .	292
<i>CUNI Submission to MRL 2023 Shared Task on Multi-lingual Multi-task Information Retrieval</i>	
Jindřich Helcl and Jindřich Libovický . . . . .	302
<i>Findings of the 1st Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2023</i>	
Francesco Tinner, David Ifeoluwa Adelani, Chris Emezue, Mammad Hajili, Omer Goldman, Muhammad Farid Adilazuarda, Muhammad Dehan Al Kautsar, Aziza Mirsaidova, Müge Kural, Dylan Massey, Chiamaka Chukwuneke, Chinedu Mbonu, Damilola Oluwaseun Oloyede, Kayode Olaleye, Jonathan Atala, Benjamin A. Ajibade, Saksham Bassi, Rahul Aralikkatte, Najoung Kim and Duygu Ataman	310

# Program

## Thursday, December 7, 2023

- 09:00 - 09:10     *Opening Remarks*
- 10:30 - 11:00    *Coffee Break*
- 11:00 - 12:30    *Poster Session*
- 12:30 - 14:00    *Lunch Break*
- 14:00 - 14:30    *Shared task session*
- 14:30 - 15:30    *Best Paper Award Session*
- 15:30 - 16:00    *Coffee Break*
- 16:00 - 16:50    *Afternoon Oral Session*
- 16:50 - 17:00    *Closing Remarks*

**Friday, December 8, 2023**

09:10 - 10:30 *Morning Oral Session*