

VarDial 2026

**VarDial 2026 - The Thirteenth Workshop on
NLP for Similar Languages, Varieties and Dialects**

Proceedings of the Workshop

March 29, 2026

©2026 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-372-2

Preface

These proceedings include the 32 papers presented at the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026), co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL). VarDial was held in Rabat, Morocco.

This year, VarDial saw a record number of submissions, with 44 research papers submitted to the workshop, 25 of which were accepted to appear in these proceedings. We extended the programme committee to be able to handle the increased reviewing load and would like to thank all new and returning PC members for being an important part of the workshop's success!

The accepted papers focus on a large number of language varieties, including Tamil, Finnish, Arabic, Ukrainian, Greek, Galician, and Piedmontese, among others. Similarly, the NLP tasks addressed in the papers are very diverse, including topics such as syllabification and quantifying linguistic distances as well as sentiment analysis and generating texts with LLMs. Notably, six of the accepted papers focus on spoken language input.

As in previous editions, VarDial 2026 features an evaluation campaign. This year, the AMIYA shared task (Arabic Modeling In Your Accent) attracted six participating teams whose system description papers are included in these proceedings, along with a report by the shared task organizers. We thank all the shared task organizers and the participants for their hard work!

The VarDial workshop organizers:

Yves Scherrer, Noëmi Aepli, Verena Blaschke, Tommi Jauhiainen,
Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri

<http://sites.google.com/view/vardial-2026/>

Organizing Committee

Organizers:

Yves Scherrer, University of Oslo (Norway)

Noëmi Aepli, University of Pennsylvania (USA)

Verena Blaschke, LMU Munich and Munich Center for Machine Learning (Germany)

Tommi Jauhiainen, University of Helsinki (Finland)

Nikola Ljubešić, Jožef Stefan Institute and University of Ljubljana (Slovenia)

Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence (UAE)

Jörg Tiedemann, University of Helsinki (Finland)

Marcos Zampieri, George Mason University (USA)

Program Committee

Program Committee:

César Aguilar (Universidad Veracruzana, Mexico)
Sina Ahmadi (George Mason University, United States)
Jorge Baptista (University of Algarve and INESC-ID, Portugal)
Delphine Bernhard (University of Strasbourg, France)
Gabriel Bernier-Colborne (National Research Council, Canada)
David Chiang (University of Notre Dame, United States)
Steven Coats (University of Oulu, Finland)
Çağrı Çöltekin (University of Tübingen, Germany)
Stefanie Dipper (Ruhr University Bochum, Germany)
Sascha Diwersy (University of Montpellier, France)
Mark Dras (Macquarie University, Australia)
Jonathan Dunn (University of Illinois Urbana-Champaign, United States)
Fahim Faisal (George Mason University, USA)
Pablo Gamallo (University of Santiago de Compostela, Spain)
Ona de Gibert Bonet (University of Helsinki, Finland)
Rob van der Goot (IT University Copenhagen, Denmark)
Cyril Goutte (National Research Council, Canada)
Jindřich Helcl (University of Oslo, Norway)
Radu Ionescu (University of Bucharest, Romania)
Aditya Joshi (UNSW, Australia)
Hour Kaing (NICT, Japan)
Anjali Kantharuban (Carnegie Mellon University, United States)
Amr Keleg (MBZUAI, UAE)
Olli Kuparinen (Tampere University, Finland)
Taja Kuzman Pungeršek (Jožef Stefan Institute, Slovenia)
John McCrae (University of Galway, Ireland)
Aleksandra Miletić (CNRS, France)
Filip Miletić (University of Stuttgart, Germany)
John Nerbonne (University of Groningen, Netherlands and University of Freiburg, Germany)
Ekaterina Lapshinova-Koltunski (University of Hildesheim, Germany)
Lung-Hao Lee (National Yang Ming Chiao Tung University, Taiwan)
Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences, Poland)
Petya Osenova (Bulgarian Academy of Sciences, Bulgaria)
Siyao (Logan) Peng (LMU Munich, Germany)
Alistair Plum (University of Luxembourg, Luxembourg)
Jelena Prokic (Leiden University, Netherlands)
Christoph Purschke (University of Luxembourg, Luxembourg)
Alan Ramponi (FBK Trento, Italy)
Francisco Rangel (Symanto Research, Spain)
Reinhard Rapp (University of Mainz, Germany)
Tanja Samardžić (IDSIA, Switzerland)
Serge Sharoff (University of Leeds, United Kingdom)
Milena Slavcheva (Bulgarian Academy of Sciences, Bulgaria)
Aarohi Srivastava (University of Notre Dame, United States)
Joel Tetreault (Dataminr, United States)
Samia Touileb (University of Bergen, Norway)
Taro Watanabe (Google Inc., Japan)

Table of Contents

<i>AMIYA Shared Task: Arabic Modeling In Your Accent at VarDial 2026</i> Nathaniel R. Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio and Kenton Murray	1
<i>Far Out: Evaluating Language Models on Slang in Australian and Indian English</i> Deniz Kaya Dilsiz, Dipankar Srirag and Aditya Joshi	18
<i>Effects of Speaker Bias in Dialect Identification and Automatic Transcription with Self-Supervised Speech Models</i> Olli Kuparinen	32
<i>OcWikiDialects: A Wikipedia Dataset With Rich Metadata for Occitan Dialect Identification</i> Oriane Nédey, Rachel Bawden, Thibault Clérice and Benoît Sagot	45
<i>Language Mixture to Develop Accurate Galician Dependency Parsers: An Exploration of Its Effects</i> Xabier Irastortza-Urbieta, José M. García-Miguel and Marcos Garcia	58
<i>Crowdsourcing Piedmontese to Test LLMs on Non-Standard Orthography</i> Gianluca Vico and Jindřch Libovický	70
<i>German-English Code-Switching in Large Language Models</i> Firat Cem Aksüt, Stefan Hillmann, Pia Knoeferle and Sebastian Möller	87
<i>Perplexity as a Metric for Dialectal Distance: A Computational Study of Greek Varieties</i> Stergios Chatzikyriakidis, Erofilis Psaltaki, Dimitrios Papadakis, Erik Henriksson and Veronika Laippala	101
<i>A Subword Embedding Approach for Variation Detection in Luxembourgish User Comments</i> Anne-Marie Lutgen, Alistair Plum and Christoph Purschke	113
<i>Onomasiological Sense Alignment Across Dialect Dictionaries. A Taxonomy-Constrained LLM Classifi- cation</i> Nathalie Mederake, Nico Urbach, Hanna Fischer and Alfred Lameli	123
<i>On the Intelligibility of Romance Language Varieties: Spanish and Portuguese in Europe and America</i> Liviu P. Dinu, Ana Sabina Uban, Teodor-George Marchitan, Ioan-Bogdan Iordache and Simona Georgescu	139
<i>Dialect Matters: Cross-Lingual ASR Transfer for Low-Resource Indic Language Varieties</i> Akriti Dhasmana, Aarohi Srivastava and David Chiang	145
<i>Ara-HOPE: Human-Centric Post-Editing Evaluation for Dialectal Arabic to Modern Standard Arabic Translation</i> Abdullah Alabdullah, Lifeng Han and Chenghua Lin	157
<i>Indic-TunedLens: Interpreting Multilingual Models in Indian Languages</i> Mihir Panchal, Deeksha Varshney, Mamta . and Asif Ekbal	172
<i>Building ASR Resources for the Hutsul Dialect of Ukrainian</i> Roman Kyslyi, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko and Zakhar Guzii . .	186

<i>From FusHa to Folk: Exploring Cross-Lingual Transfer in Arabic Language Models</i> Abdulmuizz Khalak, Abderrahmane Issam and Gerasimos Spanakis	196
<i>Extending ASR Evaluation Resources for Modern Greek Dialects</i> Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos and Stella Markantonatou	210
<i>How Should We Model the Probability of a Language?</i> Rasul Dent, Pedro Ortiz Suarez, Thibault Cl�rice and Beno�t Sagot	223
<i>Bridging Dialectal Variation: A Phonetic Transcription Tool for Tamil</i> Ahrane Mahaganapathy, Sumirtha Karunakaran, Kavitha Navakulan and Kengatharaiyer Sarveswaran	234
<i>Regional Variation in the Performance of ASR Models on Croatian and Serbian</i> Tanja Samard�i�, Peter Rupnik and Nikola Ljube�i�	242
<i>Syllable Structures Across Arabic Varieties</i> Abdelrahim Qaddoumi, Jordan Kodner, Salam Khalifa, Ellen Broselow and Owen Rambow ...	250
<i>Curriculum Learning and Pseudo-Labeling Improve the Generalization of Multi-Label Arabic Dialect Identification Models</i> Ali Mekky, Mohamed El Zeftawy, Lara Hassan, Amr Keleg and Preslav Nakov	261
<i>OpenLID-v3: Improving the Precision of Closely Related Language Identification – An Experience Report</i> Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil R�nningstad and Yves Scherrer	275
<i>Improving Dialect Robustness in Large Language Models via LoRA and Mixture-of-Experts</i> Sanjh Maheshwari, Aniket Singh Rajpoot, Oana Cocarascu and Mamta	293
<i>Evaluation Framework for Transfer Learning between Closely Related Lects: A Case Study of Lemko</i> Iliia Afanasev	304
<i>Do Large Language Models Adapt to Language Variation across Socioeconomic Status?</i> Elisa Bassignana, Mike Zhang, Dirk Hovy and Amanda Cercas Curry	317
<i>Aladdin-FTI @ AMIYA Three Wishes for Arabic NLP: Fidelity, Diglossia, and Multidialectal Generation</i> Jonathan Mutal, Perla Al Almaoui, Simon Hengchen and Pierrette Bouillon	339
<i>Maastricht University at AMIYA: Adapting LLMs for Dialectal Arabic using Fine-tuning and MBR Decoding</i> Abdulhai Alali and Abderrahmane Issam	352
<i>SDNLP at AMIYA 2026: Syrian Arabic Dialect Modeling with LoRA</i> Hasan Alkhder and Mohammad Abboush	359
<i>NUS-IDS at AMIYA/VarDial 2026: Improving Arabic Dialectness in LLMs with Reinforcement Learning</i> Sujatha Das Gollapalli, Mouad Hakam, Mingzhe Du and See-Kiong Ng	365
<i>MBZUAI at AMIYA Shared Task 2026: Adapting Open-Source LLMs for Dialectal Arabic</i> Rana Gaber, Yara Allam, Serag Amin, Ranwa Aly and Bashar Alhafni	373

A Closed-Track System for Palestinian Arabic in the AMIYA Shared Task

Khaleel Hamad and Ahmad Al-Najjar 385

