

VarDial 2024

**VarDial 2024 - The Eleventh Workshop on NLP for Similar
Languages, Varieties and Dialects**

Proceedings of the Workshop

June 20, 2024

©2024 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 979-8-89176-104-9

Preface

These proceedings include the 22 papers presented at the Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024), co-located with the 2024 annual conference of the North American Chapter of the Association for Computational Linguistics (NAACL). VarDial was held in Mexico City, Mexico, in a hybrid format, allowing participants to attend on-site or remotely.

Now at its eleventh edition, we are pleased to see that VarDial continues to serve the community as the main venue for researchers interested in the computational processing of diatopic language variation. The papers accepted this year address a wide range of topics, such as language variety identification, corpus creation, and machine translation. These proceedings once again illustrate the great linguistic diversity that VarDial embodies. They include work on Germanic (historical Dutch, Limburgish, Norwegian, Swiss German), Romance (Portuguese, Spanish, Occitan), and Slavic languages (South Slavic, Slovak), as well as Arabic and Nahuatl.

As in previous editions, VarDial 2024 features an evaluation campaign with two shared tasks: The DIALECT-COPA shared task on dialectal causal commonsense reasoning, and the DSL-ML shared task on multi-label classification of similar languages. Both tasks were organized for the first time this year, although DSL-ML relies on datasets built for earlier tasks. This volume includes the system description papers prepared by the participating teams, as well as a report written by the task organizers summarizing the results and findings of the evaluation campaign.

Finally, we would like to take this opportunity to thank all the shared task organizers and the participants for their hard work. We further thank the VarDial program committee members, and particularly the 14 PC members who newly joined this year, for being an important part of the workshop's success.

The VarDial workshop organizers:

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zamperri

<https://sites.google.com/view/vardial-2024>

Organizing Committee

Organizers

Tommi Jauhiainen, University of Helsinki

Nikola Ljubešić, Jožef Stefan Institute

Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence

Yves Scherrer, University of Oslo

Jörg Tiedemann, University of Helsinki

Marcos Zamperli, George Mason University

Program Committee

Program Committee

Noëmi Aepli, University of Zurich
César Aguilar, Universidad Veracruzana
Sina Ahmadi, University of Zurich
Laura Alonso Alemany, Universidad Nacional de Cordoba
Delphine Bernhard, Lilpa, Université de Strasbourg
Gabriel Bernier-Colborne, National Research Council Canada
Verena Blaschke, LMU Munich
Aoife Cahill, Dataminr
David Chiang, University of Notre Dame
Adrian-Gabriel Chifu, Aix-Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
Steven Coats, University of Oulu
Jon Dehdari, Fidelity Investments
Stefanie Dipper, Ruhr University Bochum
Mark Dras, Macquarie University
Jonathan Dunn, University of Illinois Urbana-Champaign
Pablo Gamallo, CITIUS, University of Santiago de Compostela
Cyril Goutte, National Research Council Canada
Nizar Habash, New York University Abu Dhabi
Radu Tudor Ionescu, University of Bucharest
Anjali Kantharuban, University of California, Berkeley
Ekaterina Lapshinova-Koltunski, University of Hildesheim
Lung-Hao Lee, National Yang Ming Chiao Tung University
John P. McCrae, Insight Center for Data Analytics, National University of Ireland Galway
Aleksandra Miletić, Department of Digital Humanities, University of Helsinki
Filip Miletić, University of Stuttgart
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
Petya Osenova, Sofia University St. Kl. Ohridski and IICT-BAS
Jelena Prokic, Leiden University
Christoph Purschke, University of Luxembourg
Francisco Manuel Rangel Pardo, Universitat Politècnica de València
Reinhard Rapp, University of Mainz
Tanja Samardžić, University of Zurich
Serge Sharoff, University of Leeds
Milena Slavcheva, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Aarohi Srivastava, University of Notre Dame
Joel Tetreault, Dataminr
Rob Van Der Goot, IT University of Copenhagen
Pidong Wang, Google
Taro Watanabe, Nara Institute of Science and Technology
Çağrı Çöltekin, University of Tübingen

Table of Contents

<i>VarDial Evaluation Campaign 2024: Commonsense Reasoning in Dialects and Multi-Label Similar Language Identification</i>	
Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer and Ivan Vulić	1
<i>What Drives Performance in Multilingual Language Models?</i>	
Sina Bagheri Nezhad and Ameeta Agrawal	16
<i>Does Whisper Understand Swiss German? An Automatic, Qualitative, and Human Evaluation</i>	
Eyal Dolev, Clemens Lutz and Noëmi Aepli	28
<i>How Well Do Tweets Represent Sub-Dialects of Egyptian Arabic?</i>	
Mai Mohamed Eida, Mayar Nassar and Jonathan Dunn	41
<i>When Elote, Choclo and Mazorca are not the Same. Isomorphism-Based Perspective to the Spanish Varieties Divergences</i>	
Cristina España-Bonet, Ankur Bhatt, Koel Dutta Chowdhury and Alberto Barrón-Cedeño	56
<i>Modeling Orthographic Variation in Occitan's Dialects</i>	
Zachary William Hopton and Noëmi Aepli	78
<i>DIALECT-COPA: Extending the Standard Translations of the COPA Causal Commonsense Reasoning Dataset to South Slavic Dialects</i>	
Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik and Taja Kuzman	89
<i>The Role of Adverbs in Language Variety Identification: The Case of Portuguese Multi-Word Adverbs</i>	
Izabela Müller, Nuno Mamede and Jorge Baptista	99
<i>NoMusic - The Norwegian Multi-Dialectal Slot and Intent Detection Corpus</i>	
Petter Mæhlum and Yves Scherrer	107
<i>Understanding Position Bias Effects on Fairness in Social Multi-Document Summarization</i>	
Olubusayo Olabisi and Ameeta Agrawal	117
<i>Can LLMs Handle Low-Resource Dialects? A Case Study on Translation and Common Sense Reasoning in Šariš</i>	
Viktória Ondrejová and Marek Šuppa	130
<i>Experiments in Multi-Variant Natural Language Processing for Nahuatl</i>	
Robert Pugh and Francis M. Tyers	140
<i>Highly Granular Dialect Normalization and Phonological Dialect Translation for Limburgish</i>	
Andreas Simons, Stefano De Pascale and Karlien Franco	152
<i>Multilingual Identification of English Code-Switching</i>	
Igor Sterner	163
<i>Studying Language Variation Considering the Re-Usability of Modern Theories, Tools and Resources for Annotating Explicit and Implicit Events in Centuries Old Text</i>	
Stella Verkijk, Pia Sommerauer and Piek T. J. M. Vossen	174

<i>Language Identification of Philippine Creole Spanish: Discriminating Chavacano From Related Languages</i>	
Aileen Joan Vicente and Charibeth Cheng	188
<i>Data-Augmentation-Based Dialectal Adaptation for LLMs</i>	
Fahim Faisal and Antonios Anastasopoulos	197
<i>JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far</i>	
Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Ivan Vulić, Fabian David Schmidt and Goran Glavaš	
209	
<i>Incorporating Dialect Understanding Into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning</i>	
Benedikt Perak, Slobodan Beliga and Ana Meštrović	220
<i>One-Shot Prompt for Language Variety Identification</i>	
Nat Gillin	230
<i>Improving Multi-Label Classification of Similar Languages by Semantics-Aware Word Embeddings</i>	
The Quyen Ngo, Thi Anh Phuong Nguyen, My Linh Ha, Thi Minh Huyen Nguyen and Phuong Le-Hong	235
<i>Brandeis at VarDial 2024 DSL-ML Shared Task: Multilingual Models, Simple Baselines and Data Augmentation</i>	
Jonne Sälevä and Chester Palen-Michel	241