

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)

Dubrovnik, Croatia
5 May 2023

ISBN: 978-1-7138-7389-1

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2023) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2023)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006

Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection</i> Galo Castillo-lópez, Arij Riabi and Djamé Seddah	1
<i>Optimizing the Size of Subword Vocabularies in Dialect Classification</i> Vani Kanjirangat, Tanja Samardžić, Ljiljana Dolamic and Fabio Rinaldi	14
<i>Murreviikko - A Dialectologically Annotated and Normalized Dataset of Finnish Tweets</i> Olli Kuparinen	31
<i>Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages</i> Verena Blaschke, Hinrich Schütze and Barbara Plank	40
<i>Temporal Domain Adaptation for Historical Irish</i> Oksana Dereza, Theodorus Fransen and John P. McCrae	55
<i>Variation and Instability in Dialect-Based Embedding Spaces</i> Jonathan Dunn	67
<i>PALI: A Language Identification Benchmark for Perso-Arabic Scripts</i> Sina Ahmadi, Milind Agarwal and Antonios Anastasopoulos	78
<i>Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora</i> Taja Kuzman, Peter Rupnik and Nikola Ljubešić	91
<i>Reconstructing Language History by Using a Phonological Ontology. An Analysis of German Surnames</i> Hanna Fischer and Robert Engsterhold	104
<i>BENCHiC-lang: A Benchmark for Discriminating between Bosnian, Croatian, Montenegrin and Serbian</i> Peter Rupnik, Taja Kuzman and Nikola Ljubešić	113
<i>Comparing and Predicting Eye-tracking Data of Mandarin and Cantonese</i> Junlin Li, Bo Peng, Yu-yin Hsu and Emmanuele Chersoni	121
<i>A Measure for Linguistic Coherence in Spatial Language Variation</i> Alfred Lameli and Andreas Schönberg	133
<i>Dialect and Variant Identification as a Multi-Label Classification Task: A Proposal Based on Near-Duplicate Analysis</i> Gabriel Bernier-colborne, Cyril Goutte and Serge Leger	142
<i>Fine-Tuning BERT with Character-Level Noise for Zero-Shot Transfer to Dialects and Closely-Related Languages</i> Aarohi Srivastava and David Chiang	152
<i>Lemmatization Experiments on Two Low-Resourced Languages: Low Saxon and Occitan</i> Aleksandra Miletić and Janine Siewert	163
<i>The Use of Khislavichi Lect Morphological Tagging to Determine its Position in the East Slavic Group</i> Ilia Afanasev	174
<i>DiatopIt: A Corpus of Social Media Posts for the Study of Diatopic Language Variation in Italy</i> Alan Ramponi and Camilla Casula	187

<i>Dialect Representation Learning with Neural Dialect-to-Standard Normalization</i> Olli Kuparinen and Yves Scherrer	200
<i>VarDial in the Wild: Industrial Applications of LID Systems for Closely-Related Language Varieties</i> Fritz Hohl and Soh-eun Shim	213
<i>Two-stage Pipeline for Multilingual Dialect Detection</i> Ankit Vaidya and Aditya Kane	222
<i>Using Ensemble Learning in Language Variety Identification</i> Mihaela Gaman	230
<i>SIDLR: Slot and Intent Detection Models for Low-Resource Language Varieties</i> Sang Yun Kwon, Gagan Bhatia, Elmoatez Billah Nagoudi, Alcides Alcoba Inciarte and Muhammad Abdul-mageed	241
<i>Findings of the VarDial Evaluation Campaign 2023</i> Noëmi Aeppli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer and Marcos Zampieri	251