7th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial '2020)

Held online due to COVID-19

Barcelona, Spain
13 December 2020

ISBN: 978-1-7138-2842-6

Printed from e-media with permission by:

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

CURRAN ASSOCIATES INC.
proceedings
.com

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

Copyright© (2020) by the Association for Computational Linguistics
All rights reserved.
Copyright for individual papers remains with the authors and are licensed under a Creative Commons 4.0 license, CC-BY. (https://creativecommons.org/licenses/by/4.0/)

Printed with permission by Curran Associates, Inc. (2021)

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

A Report on the VarDial Evaluation Campaign 2020
Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer and Marcos Zampieri . . . 1

ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German
Iuliia Nigmatulina, Tannon Kew and Tanja Samardzic 15

LSDC - A comprehensive dataset for Low Saxon Dialect Classification
Janine Siewert, Yves Scherrer, Martijn Wieling and Jörg Tiedemann 25

Machine-oriented NMT Adaptation for Zero-shot NLP tasks: Comparing the Usefulness of Close and Distant Languages
Amirhossein Tebbifakhr, Matteo Negri and Marco Turchi 36

Character Alignment in Morphologically Complex Translation Sets for Related Languages
Michael Gasser, Binyam Ephrem Seyoum and Nazareth Amlesom Kifle 47

Bilingual Lexicon Induction across Orthographically-distinct Under-Resourced Dravidian Languages
Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E. O'Connor and John P. McCrae 57

Building a Corpus for the Zaza–Gorani Language Family
Sina Ahmadi 70

Dealing with dialectal variation in the construction of the Basque historical corpus
Ainara Estarrona, Izaskun Etxeberria, Ricardo Etxepare, Manuel Padilla-Moyano and Ander Soraluze 79

Recycling and Comparing Morphological Annotation Models for Armenian Diachronic-Variational Corpus Processing
Chahan Vidal-Gorène, Victoria Khurshudyan and Anaïd Donabédian-Demopoulos 90

Neural Machine Translation for translating into Croatian and Serbian
Maja Popović, Alberto Poncelas, Marija Brkic and Andy Way 102

A Tokenization System for the Kurdish Language
Sina Ahmadi 114

Rediscovering the Slavic Continuum in Representations Emerging from Neural Models of Spoken Language Identification
Badr M. Abdullah, Jacek Kudera, Tania Avgustinova, Bernd Möbius and Dietrich Klakow 128

A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments
Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade and Jean Sibille 140

Vulgaris: Analysis of a Corpus for Middle-Age Varieties of Italian Language
Andrea Zugarini, Matteo Tiezzi and Marco Maggini 150

Towards Augmenting Lexical Resources for Slang and African American English
Alyssa Hwang, William R. Frey and Kathleen McKeown 160

Uralic Language Identification (ULI) 2020 shared task dataset and the Wanca 2017 corpora
Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen and Krister Lindén 173

Dialect Identification under Domain Shift: Experiments with Discriminating Romanian and Moldavian
Çağrı Çöltekin 186

Applying Multilingual and Monolingual Transformer-Based Models for Dialect Identification
Cristian Popa and Vlad Ștefănescu 193

HeLju@VarDial 2020: Social Media Variety Geolocation with BERT Models
Yves Scherrer and Nikola Ljubešić 202

A dual-encoding system for dialect classification
Petru Rebeja and Dan Cristea 212

Experiments in Language Variety Geolocation and Dialect Identification
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén 220

Exploring the Power of Romanian BERT for Dialect Identification
George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel and Traian Rebedea232

Combining Deep Learning and String Kernels for the Localization of Swiss German Tweets
Mihaela Gaman and Radu Tudor Ionescu 242

ZHAW-InIT - Social Media Geolocation at VarDial 2020
Fernando Benites, Manuela Hürlimann, Pius von Däniken and Mark Cieliebak 254

Discriminating between standard Romanian and Moldavian tweets using filtered character ngrams
Andrea Ceolin and Hong Zhang 265

Challenges in Neural Language Identification: NRC at VarDial 2020
Gabriel Bernier-Colborne and Cyril Goutte 273

Geolocation of Tweets with a BiLSTM Regression Model
Piyush Mishra 283