

Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)

Valencia, Spain
3 April 2017

ISBN: 978-1-5108-3872-7

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© (2017) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2017)

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>Findings of the VarDial Evaluation Campaign 2017</i> Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer and Noëmi Aepli	1
<i>Dialectometric analysis of language variation in Twitter</i> Gonzalo Donoso and David Sanchez	16
<i>Computational analysis of Gondi dialects</i> Taraka Rama, Çağrı Çöltekin and Pavel Sofroniev	26
<i>Investigating Diatopic Variation in a Historical Corpus</i> Stefanie Dipper and Sandra Waldenberger	36
<i>Author Profiling at PAN: from Age and Gender Identification to Language Variety Identification (invited talk)</i> Paolo Rosso	46
<i>The similarity and Mutual Intelligibility between Amharic and Tigrigna Varieties</i> Tekabe Legesse Feleke	47
<i>Why Catalan-Spanish Neural Machine Translation? Analysis, comparison and combination with standard Rule and Phrase-based technologies</i> Marta R. Costa-jussà	55
<i>Kurdish Interdialect Machine Translation</i> Hossein Hassani	63
<i>Twitter Language Identification Of Similar Languages And Dialects Without Ground Truth</i> Jennifer Williams and Charlie Dagli	73
<i>Multi-source morphosyntactic tagging for spoken Rusyn</i> Yves Scherrer and Achim Rabus	84
<i>Identifying dialects with textual and acoustic cues</i> Abualsoud Hanani, Aziz Qaroush and Stephen Taylor	93
<i>Evaluating HeLI with Non-Linear Mappings</i> Tommi Jauhiainen, Krister Lindén and Heidi Jauhiainen	102
<i>A Perplexity-Based Method for Similar Languages Discrimination</i> Pablo Gamallo, Jose Ramon Pichel and Iñaki Alegria	109
<i>Improving the Character Ngram Model for the DSL Task with BM25 Weighting and Less Frequently Used Feature Sets</i> Yves Bestgen	115
<i>Discriminating between Similar Languages with Word-level Convolutional Neural Networks</i> Marcelo Criscuolo and Sandra Maria Aluisio	124
<i>Cross-lingual dependency parsing for closely related languages - Helsinki's submission to VarDial 2017</i> Jörg Tiedemann	131

<i>Discriminating between Similar Languages Using a Combination of Typed and Untyped Character N-grams and Words</i>	
Helena Gomez, Iliia Markov, Jorge Baptista, Grigori Sidorov and David Pinto	137
<i>Tübingen system in VarDial 2017 shared task: experiments with language identification and cross-lingual parsing</i>	
Çağrı Çöltekin and Taraka Rama	146
<i>When Sparse Traditional Models Outperform Dense Neural Networks: the Curious Case of Discriminating between Similar Languages</i>	
Maria Medvedeva, Martin Kroon and Barbara Plank	156
<i>German Dialect Identification in Interview Transcriptions</i>	
Shervin Malmasi and Marcos Zampieri	164
<i>CLUZH at VarDial GDI 2017: Testing a Variety of Machine Learning Tools for the Classification of Swiss German Dialects</i>	
Simon Clematide and Peter Makarov	170
<i>Arabic Dialect Identification Using iVectors and ASR Transcripts</i>	
Shervin Malmasi and Marcos Zampieri	178
<i>Discriminating between Similar Languages using Weighted Subword Features</i>	
Adrien Barbaresi	184
<i>Exploring Lexical and Syntactic Features for Language Variety Identification</i>	
Chris van der Lee and Antal van den Bosch	190
<i>Learning to Identify Arabic and German Dialects using Multiple Kernels</i>	
Radu Tudor Ionescu and Andrei Butnaru	200
<i>Slavic Forest, Norwegian Wood</i>	
Rudolf Rosa, Daniel Zeman, David Mareček and Zdeněk Žabokrtský	210