

1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects 2014

Held at COLING 2014

**Dublin, Ireland
23 August 2014**

ISBN: 978-1-63439-221-1

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© (2014) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2014)

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: 845-758-2634
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Corpus-based Study and Identification of Mandarin Chinese Light Verb Variations</i> Chu-Ren Huang, Jingxia Lin, Menghan JIANG and Hongzhi Xu	1
<i>Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic</i> Maria Sukhareva and Christian Chiarcos	11
<i>Pos-tagging different varieties of Occitan with single-dialect resources</i> Marianne Vergez-Couret and Assaf Urieli	21
<i>Unsupervised adaptation of supervised part-of-speech taggers for closely related languages</i> Yves Scherrer	30
<i>Morphological Disambiguation and Text Normalization for Southern Quechua Varieties</i> Annette Rios Gonzales and Richard Alexander Castro Mamani	39
<i>The Varitext platform and the Corpus des variétés nationales du français (CoVaNa-FR) as resources for the study of French from a pluricentric perspective</i> Sascha Diwersy	48
<i>A Report on the DSL Shared Task 2014</i> Marcos Zampieri, Liling Tan, Nikola Ljubešić and Jörg Tiedemann	58
<i>Employing Phonetic Speech Recognition for Language and Dialect Specific Search</i> Corey Miller, Rachel Strong, Evan Jones and Mark Vinson	68
<i>Part-of-Speech Tag Disambiguation by Cross-Linguistic Majority Vote</i> Noëmi Aepli, Ruprecht von Waldenfels and Tanja Samardžić	76
<i>Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging</i> Nora Hollenstein and Noëmi Aepli	85
<i>Automatically building a Tunisian Lexicon for Deverbal Nouns</i> Ahmed Hamdi, Nuria Gala and Alexis Nasr	95
<i>Statistical Morph Analyzer (SMA++) for Indian Languages</i> Saikrishna Srirampur, Ravi Chandibhamar and Radhika Mamidi	103
<i>Improved Sentence-Level Arabic Dialect Classification</i> Christoph Tillmann, Saab Mansour and Yaser Al-Onaizan	110
<i>Using Maximum Entropy Models to Discriminate between Similar Languages and Varieties</i> Jordi Porta and José-Luis Sancho	120
<i>Exploring Methods and Resources for Discriminating Similar Languages</i> Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook and Timothy Baldwin	129
<i>The NRC System for Discriminating Similar Languages</i> Cyril Goutte, Serge Léger and Marine Carpuat	139
<i>Experiments in Sentence Language Identification with Groups of Similar Languages</i> Ben King, Dragomir Radev and Steven Abney	146

<i>A Simple Baseline for Discriminating Similar Languages</i>	
Matthew Purver	155