

1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)

Marseille, France
11-16 May 2020

Editors:

**Dorothee Beermann
Laurent Besacier
Sakriani Sakti
Claudia Soria**

ISBN: 978-1-7138-1277-7

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© (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-NC. (<https://creativecommons.org/licenses/by-nc/4.0/>)

Printed with permission by Curran Associates, Inc. (2020)

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>Neural Models for Predicting Celtic Mutations</i> Kevin Scannell	1
<i>Eidos: An Open-Source Auditory Periphery Modeling Toolkit and Evaluation of Cross-Lingual Phonemic Contrasts</i> Alexander Gutkin	9
<i>Open-Source High Quality Speech Datasets for Basque, Catalan and Galician</i> Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin and Clara Rivera	21
<i>Two LRL & Distractor Corpora from Web Information Retrieval and a Small Case Study in Language Identification without Training Corpora</i> Armin Hoenen, Cemre Koc and Marc Rahn	28
<i>Morphological Disambiguation of South Sámi with FSTs and Neural Networks</i> Mika Härmäläinen and Linda Wiecheteck	36
<i>Effects of Language Relatedness for Cross-lingual Transfer Learning in Character-Based Language Models</i> Mittul Singh, Peter Smit, Sami Virpioja and Mikko Kurimo	41
<i>Multilingual Graphemic Hybrid ASR with Massive Data Augmentation</i> Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf and Geoffrey Zweig	46
<i>Neural Text-to-Speech Synthesis for an Under-Resourced Language in a Diglossic Environment: the Case of Gascon Occitan</i> Ander Corral, Igor Leturia, Aure Séguier, Michäel Barret, Benaset Dazéas, Philippe Boula de Mareüil and Nicolas Quint	53
<i>Transfer Learning for Less-Resourced Semitic Languages Speech Recognition: the Case of Amharic</i> Yonas Woldemariam	61
<i>Semi-supervised Acoustic Modelling for Five-lingual Code-switched ASR using Automatically-segmented Soap Opera Speech</i> Nick Wilkinson, Astik Biswas, Emre Yilmaz, Febe De Wet, Ewald Van der westhuizen and Thomas Niesler	70
<i>Investigating Language Impact in Bilingual Approaches for Computational Language Documentation</i> Marcely Zanon Boito, Aline Villavicencio and Laurent Besacier	79
<i>Design and evaluation of a smartphone keyboard for Plains Cree syllabics</i> Eddie Santos and Atticus Harrigan	88
<i>MultiSeg: Parallel Data and Subword Information for Learning Bilingual Embeddings in Low Resource Scenarios</i> Efsun Sarioglu Kayi, Vishal Anand and Smaranda Muresan	97
<i>Poio Text Prediction: Lessons on the Development and Sustainability of LTs for Endangered Languages</i> Gema Zamora Fernández, Vera Ferreira and Pedro Manha	106
<i>Text Corpora and the Challenge of Newly Written Languages</i> Alice Millour and Karën Fort	111

<i>Scaling Language Data Import/Export with a Data Transformer Interface</i> Nicholas Buckeridge and Ben Foley	121
<i>Fully Convolutional ASR for Less-Resourced Endangered Languages</i> Bao Thai, Robert Jimerson, Raymond Ptucha and Emily Prud'hommeaux	126
<i>Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis</i> Sashi Novitasari, Andros Tjandra, Sakriani Sakti and Satoshi Nakamura	131
<i>Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model</i> San Pa Pa Aung, Win Pa Pa and Tin Lay Nwe	139
<i>Phoneme Boundary Analysis using Multiway Geometric Properties of Waveform Trajectories</i> BHAGATH PARABATTINA and Pradip K. Das	144
<i>Natural Language Processing Chains Inside a Cross-lingual Event-Centric Knowledge Pipeline for European Union Under-resourced Languages</i> Diego Alves, Gaurish Thakkar and Marko Tadić	153
<i>Component Analysis of Adjectives in Luxembourgish for Detecting Sentiments</i> Joshgun Sirajzade, Daniela Gierschek and Christoph Schommer	159
<i>Acoustic-Phonetic Approach for ASR of Less Resourced Languages Using Monolingual and Cross-Lingual Information</i> shweta bansal	167
<i>An Annotation Framework for Luxembourgish Sentiment Analysis</i> Joshgun Sirajzade, Daniela Gierschek and Christoph Schommer	172
<i>A Sentiment Analysis Dataset for Code-Mixed Malayalam-English</i> Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly and John Philip McCrae	177
<i>Speech-Emotion Detection in an Indonesian Movie</i> Fahmi Fahmi, Meganingrum Arista Jiwanggi and Mirna Adriani	185
<i>Macsen: A Voice Assistant for Speakers of a Lesser Resourced Language</i> Dewi Jones	194
<i>Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text</i> Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini and John Philip McCrae	202
<i>Gender Detection from Human Voice Using Tensor Analysis</i> Prasanta Roy, Parabattina Bhagath and Pradip Das	211
<i>Data-Driven Parametric Text Normalization: Rapidly Scaling Finite-State Transduction Verbalizers to New Languages</i> Sandy Ritchie, Eoin Mahon, Kim Heiligenstein, Nikos Bampounis, Daan van Esch, Christian Schallhart, Jonas Mortensen and Benoit Brard	218
<i>Lenition and Fortition of Stop Codas in Romanian</i> Mathilde Hutin, Oana Niculescu, Ioana Vasilescu, Lori Lamel and Martine Adda-Decker	226

<i>Adapting a Welsh Terminology Tool to Develop a Cornish Dictionary</i> Delyth Prys	235
<i>Multiple Segmentations of Thai Sentences for Neural Machine Translation</i> Alberto Poncelas, Wichaya Pidchamook, Chao-Hong Liu, James Hadley and Andy Way	240
<i>Automatic Extraction of Verb Paradigms in Regional Languages: the case of the Linguistic Crescent varieties</i> elena knyazeva, Gilles Adda, Philippe Boula de Mareüil, Maximilien Guérin and Nicolas Quint	245
<i>FST Morphology for the Endangered Skolt Sami Language</i> Jack Rueter and Mika Hämäläinen	250
<i>Voted-Perceptron Approach for Kazakh Morphological Disambiguation</i> Gulmira Tolegen, Alymzhan Toleu and Rustam Mussabayev	258
<i>DNN-Based Multilingual Automatic Speech Recognition for Wolaytta using Oromo Speech</i> Martha Yifiru Tachbelie, Solomon Teferra Abate and Tanja Schultz	265
<i>Building Language Models for Morphological Rich Low-Resource Languages using Data from Related Donor Languages: the Case of Uyghur</i> Ayimunishagu Abulimiti and Tanja Schultz	271
<i>Basic Language Resources for 31 Languages (Plus English): The LORELEI Representative and Incident Language Packs</i> Jennifer Tracey and Stephanie Strassel	277
<i>On the Exploration of English to Urdu Machine Translation</i> Sadaf Abdul Rauf, Syeda Abida, Noor-e- Hira, Syeda Zahra, Dania Parvez, Javeria Bashir and Qurat-ul-ain Majid	285
<i>Developing a Twi (Asante) Dictionary from Akan Interlinear Glossed Texts</i> Dorothee Beermann, Lars Hellan, Pavel Mihaylov and Anna Struck	294
<i>Adapting Language Specific Components of Cross-Media Analysis Frameworks to Less-Resourced Languages: the Case of Amharic</i> Yonas Woldemariam and Adam Dahlgren	298
<i>Phonemic Transcription of Low-Resource Languages: To What Extent can Preprocessing be Automated?</i> Guillaume Wisniewski, Séverine Guillaume and Alexis Michaud	306
<i>Manual Speech Synthesis Data Acquisition - From Script Design to Recording Speech</i> Atli Sigurgeirsson, Gunnar Örnólfsson and Jón Guðnason	316
<i>Owóksape - An Online Language Learning Platform for Lakota</i> Jan Ullrich, Elliot Thornton, Peter Vieira, Logan Swango and Marek Kupiec	321
<i>A Corpus of the Sorani Kurdish Folkloric Lyrics</i> Sina Ahmadi, Hossein Hassani and Kamaladdin Abedi	330
<i>Improving the Language Model for Low-Resource ASR with Online Text Corpora</i> Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler and Francis Tyers ...	336

<i>A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization</i>	
Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan S Sharma and Patrick Littell	342
<i>"A Passage to India": Pre-trained Word Embeddings for Indian Languages</i>	
Saurav Kumar, Saunack Kumar, Diptesh Kanojia and Pushpak Bhattacharyya	352
<i>A Counselling Corpus in Cantonese</i>	
John Lee, Tianyuan Cai, Wenxiu Xie and Lam Xing	358
<i>Speech Transcription Challenges for Resource Constrained Indigenous Language Cree</i>	
Vishwa Gupta and Gilles Boulianne	362
<i>Turkish Emotion Voice Database (TurEV-DB)</i>	
Salih Firat Canpolat, Zuhul Ormanoğlu and Deniz Zeyrek	368