

Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-5)

Dublin, Ireland
26-27 May 2022

ISBN: 978-1-7138-6730-2

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© (2022) 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>Development of the Siberian Ingrian Finnish Speech Corpus</i> Ivan Ubaleht and Taisto-Kalevi Raudalainen	1
<i>New syntactic insights for automated Wolof Universal Dependency parsing</i> Bill Dyer	5
<i>Corpus Development of Kiswahili Speech Recognition Test and Evaluation sets, Preemptively Mitigating Demographic Bias Through Collaboration with Linguists</i> Kathleen Siminyu, Kibibi Mohamed Amran, Abdulrahman Ndegwa Karatu, Mnata Resani, Mwimbi Makobo Junior, Rebecca Ryakitimbo and Britone Mwasaru	13
<i>CLD² Language Documentation Meets Natural Language Processing for Revitalising Endangered Languages</i> Roberto Zariquiey, Arturo Oncevay and Javier Vera	20
<i>One Wug, Two Wug+s Transformer Inflection Models Hallucinate Affixes</i> Farhan Samir and Miikka Silfverberg	31
<i>Automated speech tools for helping communities process restricted-access corpora for language revival efforts</i> Nay San, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Helen Simpson and Dan Jurafsky	41
<i>G_i2P_i Rule-based, index-preserving grapheme-to-phoneme transformations</i> Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher D Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo and Sabrina Yu	52
<i>Shallow Parsing for Nepal Bhasa Complement Clauses</i> Borui Zhang, Abe Kazemzadeh and Brian Reese	61
<i>Using LARA to create image-based and phonetically annotated multimodal texts for endangered languages</i> Branislav Bédi, Hakeem Beedar, Belinda Chiera, Nedelina Ivanova, Christèle Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, John Sloan and Ghil'ad Zuckermann	68
<i>Recovering Text from Endangered Languages Corrupted PDF documents</i> Nicolas Stefanovitch	78
<i>Learning Through Transcription</i> Mat Bettinson and Steven Bird	83
<i>Developing a Part-Of-Speech tagger for te reo Māori</i> Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan and Gianna Leoni	93
<i>Challenges and Perspectives for Innu-Aimun within Indigenous Language Technologies</i> Antoine Cadotte, Tan Le Ngoc, Mathieu Boivin and Fatiha Sadat	99
<i>Using Speech and NLP Resources to build an iCALL platform for a minority language, the story of An Scéalaí, the Irish experience to date</i> Neasa Ní Chiaráin, Oisín Nolan, Madeleine Comtois, Neimhin Robinson Gunning, Harald Berthelsen and Ailbhe Ni Chasaide	109
<i>Closing the NLP Gap Documentary Linguistics and NLP Need a Shared Software Infrastructure</i> Luke Gessler	119

<i>Can We Use Word Embeddings for Enhancing Guarani-Spanish Machine Translation?</i>	
Santiago Góngora, Nicolás Giossa and Luis Chiruzzo	127
<i>Faoi Gheasa an adaptive game for Irish language learning</i>	
Liang Xu, Elaine Uí Dhonnchadha and Monica Ward	133
<i>Using Graph-Based Methods to Augment Online Dictionaries of Endangered Languages</i>	
Khalid Alnajjar, Mika Hämmäläinen, Niko Tapio Partanen and Jack Rueter	139
<i>Reusing a Multi-lingual Setup to Bootstrap a Grammar Checker for a Very Low Resource Language without Data</i>	
Inga Lill Sigga Mikkelsen, Linda Wiechetek and Flammie A Pirinen	149
<i>A Word-and-Paradigm Workflow for Fieldwork Annotation</i>	
Maria Copot, Sara Court, Noah Diewald, Stephanie Antetomaso and Micha Elsner	159
<i>Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family)</i>	
Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyễn and Maxime Fily	170
<i>Morphologically annotated corpora of Pomak</i>	
Ritván Jusúf Karahóĝa, Panagiotis G. Krimpas, Vivian Stamou, Vasileios Arampatzakis, Dimitrios Karamatskos, Vasileios Sevetlidis, Nikolaos Constantinides, NIKOLAOS KOKKAS, George Pavlidis and Stella Markantonatou	179
<i>Enhancing Documentation of Hupa with Automatic Speech Recognition</i>	
Zoey Liu, Justin Spence and Emily Tucker Prud'hommeaux	187