

COMPUTEL-5 2022

**Fifth Workshop on the Use of Computational Methods in the  
Study of Endangered Languages**

**Proceedings of the Workshop**

May 26-27, 2022

©2022 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
209 N. Eighth Street  
Stroudsburg, PA 18360  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 978-1-955917-30-8

## Introduction

These proceedings contain the papers presented at the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages, held as a hybrid event May 25-26, 2022 in Dublin, Ireland, and co-located with the 60th Association of Computational Linguistics (ACL) conference. As the name implies, this is the fifth workshop held on the topic—the first meeting was co-located with the ACL main conference in Baltimore, Maryland in 2014 and the second, third, and fourth ones in 2017, 2019, and 2021 were co-located with the 5th, 6th, and 7th editions of the International Conference on Language Documentation and Conservation (ICLDC) at the University of Hawai‘i at Mānoa. This is the second time this workshop has been co-located with the ACL main conference and it enhances ACL 2022’s Theme Track: “Language Diversity: from Low-Resource to Endangered Languages”.

The workshop covers a wide range of topics relevant to the study and documentation of endangered languages, ranging from technical papers on working systems and applications, to reports on community activities with supporting computational components.

The purpose of the workshop is to bring together computational researchers, documentary linguists, and people involved with community efforts of language documentation and revitalization to take part in both formal and informal exchanges on how to integrate rapidly evolving language processing methods and tools into efforts of language description, documentation, and revitalization. The organizers are pleased with the range of papers, many of which highlight the importance of interdisciplinary work and interaction between the various communities that the workshop is aimed towards.

We received 36 submissions as papers or extended abstracts. After a thorough review process, 23 submissions were selected to be published in the ACL Anthology. Twelve submissions were accepted as posters and twelve for oral presentations.

The Organizing Committee would like to thank the Program Committee for their thoughtful review of the submissions. We are also grateful to the Social Sciences and Humanities Research Council (SSHRC) of Canada for supporting the workshop through their Partnership Grant #895-2019-1012. We would moreover want to acknowledge the support of the organizers of ACL 2022.

# Organizing Committee

## Organizers

Sarah Moeller, University of Florida, USA  
Antonios Anastasopoulos, George Mason University, USA  
Antti Arppe, University of Alberta, Canada  
Aditi Chaudhary, Carnegie Mellon University, USA  
Atticus Harrigan, University of Alberta, Canada  
Josh Holden, University of Alberta, Canada  
Jordan Lachler, University of Alberta, Canada  
Alexis Palmer, University of Colorado Boulder, USA  
Shruti Rijhwani, Carnegie Mellon University, USA  
Lane Schwartz, University of Illinois, USA

# Program Committee

## Program Committee

Alexandre Arkhipov, Universität Hamburg  
Alexis Michaud, CNRS  
Alexis Palmer, University of Colorado, Boulder  
Anna Kazantseva, National Research Council Canada  
Antti Arppe, University of Alberta  
Borini Lahiri, Indian Institute of Technology Kharagpur  
Borui Zhang, University of Florida  
Christopher D Cox, Carleton University  
Claire Bower, Yale University  
Daan van Esch, Leiden University  
Daisy Rosenblum, University of British Columbia  
Dorothee Beermann, Norwegian University of Science and Technology  
Elizabeth Salesky, Johns Hopkins University  
Emily M. Bender, University of Washington  
Emily Prud'hommeaux, Boston College  
Emmanuel Schang, Université d'Orléans  
Francis M. Tyers, Indiana University, Bloomington  
František Kratochvíl, Palacky University  
Gary F Simons, SIL International  
Jean Maillard, Facebook AI  
Jeffrey Good, State University of New York at Buffalo  
Jordan Lachler, University of Alberta  
Jörg Tiedemann, University of Helsinki  
Josh Holden, University of Alberta  
Judith Lynn Klavans, University of Maryland, College Park  
Lane Schwartz, University of Alaska Fairbanks  
Lori Levin, School of Computer Science, Carnegie Mellon University  
Luke Gessler, Georgetown University  
Martin Benjamin, Kamusi Project International  
Meladel Mistica, The University of Melbourne  
Menzo Windhouwer, University of Amsterdam  
Olga Lovick, University of Saskatchewan  
Olivia Sammons, First Nations University of Canada  
Paul Trilsbeek, Max Planck Institute for Psycholinguistics  
Rebecca Knowles, National Research Council Canada  
Richard Sproat, Massachusetts Institute of Technology  
Ritesh Kumar, Dr. Bhimrao Ambedkar University  
Robert Forkel, Max-Planck Institute for Evolutionary Anthropology  
Roland Kuhn, National Research Council of Canada  
Sakriani Sakti, Japan Advanced Institute of Science and Technology  
Sonal Sinha, Google  
Steven Bird, Charles Darwin University  
Worthy Martin, University of Virginia, Charlottesville  
Yves Scherrer, University of Helsinki  
Zahra Azin, Carleton University

Zoey Liu, Boston College

## 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>CLD<sup>2</sup> 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<sub>i</sub>2P<sub>i</sub> 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 Hämäläinen, 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 Karahóga, 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