ComputEL 2024

The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages

Proceedings of the Workshop

March 21-22, 2024

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

ISBN 979-8-89176-086-8

Introduction

These proceedings contain the papers presented at the 7th Workshop on the Use of Computational Methods in the Study of Endangered Languages, held as a hybrid event March 21-22, 2024 in St. Julians, Malta, and co-located with the 18th Conference of the European Chapter of the Association for Computational Linguistics. As the name implies, this is the seventh 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, fourth and sixth ones in 2017, 2019, 2021 and 2023 were co-located with the 5th, 6th, 7th, and 8th editions of the International Conference on Language Documentation and Conservation (ICLDC) at the University of Hawaiʻi at Mānoa. The fifth iteration of the workshop was held in 2022 alongside the 60th Association of Computational Linguistics (ACL) conference in Dublin, Ireland. This is the third time this workshop has been co-located with the ACL main conference.

The primary aim of the workshop is to continue narrowing the gap between computational linguists interested in methods for endangered languages, field linguists documenting these languages, and the language communities who are striving to maintain their languages. The intention of the workshop is not merely to allow for the presentation of research, but also to build a network of computational linguists, documentary linguists, and community language activists who are able to effectively join together and serve their common interests. 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.

In addition to the regular program, we hosted a special theme session discussion at the workshop. The theme for this Special Session is “Partnerships in Practice”. The goal of this Special Session is to increase our shared understanding of how best to work together across disciplinary and cultural boundaries to support community goals for language revitalization.

We received 34 submissions as papers or extended abstracts or submissions to the Special Session. After a thorough review process, 19 submissions were accepted of which 13 were selected to be published in the ACL Anthology excluding the extended abstracts.

The Organizing Committee would like to thank the Program Committee for their thoughtful review of the submissions. We would moreover want to acknowledge the support of the organizers of EACL 2024.

Program Committee

Chairs

Sarah Moeller, University of Florida
Godfred Agyapong, University of Florida
Christopher Cox, Carleton University
Aditi Chaudhary, Google Research
Shruti Rijhwani, Google DeepMind
Ryan Henke, University of Wisconsin–Madison
Alexis Palmer, University of Colorado Boulder
Daisy Rosenblum, University of British Columbia, Canada
Lane Schwartz, University of Alaska-Fairbanks, USA
Antti Arppe, University of Alberta

Program Committee

Steven Abney, Univ of Michigan
Antonios Anastasopoulos, George Mason University
Alexandre Arkhipov, Universität Hamburg
Tara Azin, Carleton University
Dorothee Beermann, Norwegian University of Science and Technology
Martin Benjamin, Kamusi Project International
Claire Bowern, Yale University
Rolando Coto-Solano, Dartmouth College
Vera Ferreira, CIDLeS - Interdisciplinary Centre for Social and Language Documentation
Luke Gessler, University of Colorado, Boulder
Michael Ginn, University of Colorado
Jeff Good, University at Buffalo
Michael Goodman, LivePerson, Inc.
Atticus Harrigan, University of Alberta
Gary Holton, University of Hawaii
Raphael Iyamu, University of Florida
Marie-Odile Junker, Carleton University
Anna Kazantseva, National Research Council Canada
Frantisek Kratochvil, Palacky University Olomouc
Roland Kuhn, National Research Council of Canada
Ritesh Kumar, Dept. of Linguistics, Dr. Bhimrao Ambedkar University, Agra
Ngoc Tan Le, Universite du Quebec a Montreal
Éric Le Ferrand, Boston College
Gina-Anne Levow, University of Washington
Zoey Liu, Department of Linguistics, University of Florida
Olga Lovick, University of Saskatchewan
Jean Maillard, Meta AI
Ali Marashian, University of Colorado at Boulder
Bradley McDonnell, University of Hawai'i at Mānoa
Alexis Michaud, CNRS - LACITO
Steven Moran, Université de Neuchâtel
Saliha Muradoglu, The Australian National University
Claire Post, University of Colorado Boulder

Emily Prud'hommeaux, Boston College
Karthick Narayanan Ramakrishnan, Krea University
Enora Rice, University of Colorado Boulder
Daisy Rosenblum, UBC
Elizabeth Salesky, Johns Hopkins University
Olivia Sammons, First Nations University of Canada
Nay San, Stanford University
Emmanuel Schang, Université d'Orléans
Yves Scherrer, University of Oslo
Miikka Silfverberg, University of British Columbia
Gary Simons, SIL International
Sonal Sinha, K.M.Institute of Hindi and Linguistics, Dr. B. R Ambedkar University
Nick Thieberger, University of Melbourne
Paul Trilsbeek, Max Planck Institute for Psycholinguistics
Francis Tyers, Indiana University
Daan Van Esch, Google Research
Borui Zhang, University of Florida

Table of Contents

Cloud-based Platform for Indigenous Language Sound Education
Min Chen, Chris Lee, Naatosi Fish, Mizuki Miyashita and James Randall 1

Technology and Language Revitalization: A Roadmap for the Mvskoke Language
Julia Mainzinger 7

Investigating the productivity of Passamaquoddy medials: A computational approach
James Roberts 13

T is for Treu, but how do you pronounce that? Using C-LARA to create phonetic texts for Kanak languages
Pauline Welby, Fabrice Wacalie, Manny Rayner and Chatgpt-4 C-Lara-Instance 21

Machine-in-the-Loop with Documentary and Descriptive Linguists
Sarah Moeller and Antti Arppe 27

Automatic Transcription of Grammaticality Judgements for Language Documentation
Éric Le Ferrand and Emily Prud'hommeaux 33

Fitting a Square Peg into a Round Hole: Creating a UniMorph dataset of Kanien'kéha Verbs
Anna Kazantseva, Akwiratékha Martin, Karin Michelson and Jean-Pierre Koenig 39

Data-mining and Extraction: the gold rush of AI on Indigenous Languages
Marie-Odile Junker 52

Looking within the self: Investigating the Impact of Data Augmentation with Self-training on Automatic Speech Recognition for Hupa
Nitin Venkateswaran and Zoey Liu 58

Creating Digital Learning and Reference Resources for Southern Michif
Heather Souter, Olivia Sammons and David Huggins Daines 67

MunTTS: A Text-to-Speech System for Mundari
Varun Gumma, Rishav Hada, Aditya Yadavalli, Pamir Gogoi, Ishani Mondal, Vivek Seshadri and Kalika Bali 76

End-to-End Speech Recognition for Endangered Languages of Nepal
Marieke Meelen, Alexander O'neill and Rolando Coto-Solano 83

Akha, Dara-ang, Karen, Khamu, Mlabri and Urak Lawoi' language minorities' subjective perception of their languages and the outlook for development of digital tools
Joanna Dolinska, Shekhar Nayak and Sumittra Suraratdecha 94

Program

Thursday, March 21, 2024

09:30 - 10:00 *Day-1 Welcome + Opening Remarks*

10:00 - 10:30 *Day-1 Session A*

A Finite State Model for the Morphological Analysis of Eyak
Olivia Waring and Gary Holton

Akha, Dara-ang, Karen, Khamu, Mlabri and Urak Lawoi' language minorities' subjective perception of their languages and the outlook for development of digital tools
Joanna Dolinska, Shekhar Nayak and Sumittra Suraratdecha

10:30 - 11:00 *Day-1 Break*

11:00 - 12:30 *Day-1 Session B*

T is for Treu, but how do you pronounce that? Using C-LARA to create phonetic texts for Kanak languages
Pauline Welby, Fabrice Wacalie, Manny Rayner and Chatgpt-4 C-Lara-Instance

End-to-End Speech Recognition for Endangered Languages of Nepal
Marieke Meelen, Alexander O'neill and Rolando Coto-Solano

MunTTS: A Text-to-Speech System for Mundari
Varun Gumma, Rishav Hada, Aditya Yadavalli, Pamir Gogoi, Ishani Mondal, Vivek Seshadri and Kalika Bali

12:30 - 14:00 *Day-1 Lunch*

14:00 - 15:30 *Day-1 Session C*

Fitting a Square Peg into a Round Hole: Creating a UniMorph dataset of Kanien'kéha Verbs
Anna Kazantseva, Akwiratékha Martin, Karin Michelson and Jean-Pierre Koenig

Machine-in-the-Loop with Documentary and Descriptive Linguists
Sarah Moeller and Antti Arppe

Thursday, March 21, 2024 (continued)

Language Root Empowering Indigenous Communities through a Community-Centric Approach to Language Revitalization via an Innovative Mobile Application
Stephanie Witkowski

15:30 - 16:00 *Day-1 Break*

16:00 - 17:30 *Day-1 Special Session I Partnerships in North America*

Data-mining and Extraction: the gold rush of AI on Indigenous Languages
Marie-Odile Junker

Creating Digital Learning and Reference Resources for Southern Michif
Heather Souter, Olivia Sammons and David Huggins Daines

Cloud-based Platform for Indigenous Language Sound Education
Min Chen, Chris Lee, Naatosi Fish, Mizuki Miyashita and James Randall

Friday, March 22, 2024

09:00 - 10:30 *Day-2 Special Session II Partnerships in Europe and Australia*

Computel partnerships in practice
Flammie Pirinen and Tromsø Troms og Finnmark

How collaboration between Celtic language communities has improved
Leena Farhat and Preben Vangberg

Designing Indigenous PhD Projects
Steven Bird

10:30 - 11:00 *Day-2 Break*

11:00 - 12:30 *Day-2 Session D*

Automatic Transcription of Grammaticality Judgements for Language Documentation
Éric Le Ferrand and Emily Prud'hommeaux

Creating a Multimedia Online Dictionary for an Endangered Language
Yarjis Xueqing Zhong

Investigating the productivity of Passamaquoddy medials: A computational approach
James Roberts

DEVELOPING A NEPALBHĀSĀ E-CORPUS & CHALLENGES IN ENCODING ADJUSTMENT
Shahani Shrestha and Prajwal Shrestha

12:30 - 14:00 *Day-2 Lunch*

14:00 - 15:30 *Day-2 Session E*

The platform Open Text Collections as a provider of interoperable high-quality curated interlinear glossed text
Sebastian Nordhoff, Christian Döhler and Mandana Seyfeddinipur

Technology and Language Revitalization: A Roadmap for the Mvskoke Language
Julia Mainzinger

Looking within the self: Investigating the Impact of Data Augmentation with Self-training on Automatic Speech Recognition for Hupa
Nitin Venkateswaran and Zoey Liu

Phonetic Granularity Effects on Forced Alignment Across Panãra and English
Emily Ahn, Eleanor Chodroff, Myriam Lapierre and Gina-Anne Levow

15:30 - 15:45 *Day-2 Closing Remarks*