ComputEL-4

Proceedings of the

4th
Workshop on the
Use of Computational Methods in the Study of Endangered Languages

Volume 1 (Papers)

March 2–3, 2021
Online

Support:

NSF

Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada

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 978-1-954085-01-5

Preface

These proceedings contain the papers presented at the 4th Workshop on the Use of Computational Methods in the Study of Endangered languages, held virtually March 2–3, 2021, and co-timed to immediately precede the 7th International Conference on Language Documentation and Conservation (ICLDC7). As the name implies, this is the fourth workshop held on the topic—the first meeting was co-located with the ACL main conference in Baltimore, Maryland in 2014 and the second and third ones in 2017 and 2019 were co-located with the 5th and 6th editions of the International Conference on Language Documentation and Conservation (ICLDC) at the University of Hawaiʻi at Mānoa.

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 24 submissions as papers or extended abstracts. After a thorough review process, 8 of the submissions were selected for Volume 1 as long papers and 4 as as short papers, to be published in the ACL Anthology, representing a 50% acceptance rate. In addition, an additional 2 long papers and 1 short paper were accepted as resource oriented papers, alongside 5 extended abstracts (1 as a resource presentation), for Volume 2, to be published in CU Scholar.

The Organizing Committee would like to thank the Program Committee for their thoughtful input on 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 ICLDC7, in particular Brad McDonnell.

ANTTI ARPPE
JEFF GOOD
ATTICUS HARRIGAN
MANS HULDEN
JORDAN LACHLER
SARAH MOELLER
ALEXIS PALMER
LANE SCHWARTZ
MIIKKA SILFVERBERG

Organizing Committee:

Antti Arppe, University of Alberta
Jeff Good, Buffalo University
Atticus Harrigan, University of Alberta
Mans Hulden, University of Colorado, Boulder
Jordan Lachler, University of Alberta
Sarah Moeller, University of Colorado, Boulder
Alexis Palmer, University of North Texas
Miikka Silfverberg, University of British Columbia
Lane Schwartz, University of Illinois, Urbana-Champaign

Program Committee:

Andrea Berez-Kroeker, U Hawaii at Manoa
Antti Arppe, University of Alberta
Christopher Cox, Carleton University
Conor Snoek, University of Lethbridge
Dorothee Beermann, Norges teknisk-naturvitenskaplige universitet
František Kratochvíl, Palacký University Olomouc, Czech Republic
Harald Hammarström, Uppsala University
Inge Genee, University of Lethbridge
Jeff Good, University at Buffalo
Kevin Scannell, Saint Louis University
Lane Schwartz, University of Illinois at Urbana-Champaign
Laura Welcher, The Long Now Foundation
Lori Levin, Carnegie Mellon University
Martin Benjamin, Kamusi Project International
Nick Thieberger, University of Melbourne
Olga Lovick, University of Saskatchewan
Steven Abney, University of Michigan
Steven Bird, Charles Darwin University
Steven Moran, University of Zurich
W.N. Martin, University of Virginia
Andrew Cowell, University of Colorado
Jordan Lachler, University of Alberta
Olivia Sammons, Carleton University
Vera Ferreira, CIDLeS – Interdisciplinary Centre for Social and Language Documentation
Alexandre Arkhipov, Hamburg University, Lomonosov Moscow State University
Alexis Palmer, University of North Texas
Antonios Anastasopoulos, Carnegie Mellon University
Damir Cavar, Indiana University
Fei Xia, University of Washington

Graham Neubig, Carnegie Mellon University
Michael Wayne Goodman, Nanyang Technological University
Miikka Silfverberg, University of British Columbia
Mike Maxwell, University of Maryland
Paul Trilsbeek, The Language Archive, Max Planck Institute for Psycholinguistics
Richard Sproat, Google, Japan
Robert Forkel, Max Planck Institute for the Science of Human History
Roland Kuhn, National Research Council of Canada
Tommi A Pirinen, UiT Norgga árktalaš universitehta
Yves Scherrer, University of Helsinki
Helen Aristar-Dry, University Of Texas at Austin (Research Affiliate)
Lane Schwartz, University of Illinois
Sebastian Drude, Museu Paraense Emílio Goeldi
Felix K. Ameka, Leiden University Centre for Linguistics

Table of Contents

Conference Program

Tuesday, March 2nd, 2021

10:00–10:15 ***Opening remarks***

10:15–10:30 **Translating Fieldwork into Datasets: the development of a corpus for the quantitative investigation of grammatical phenomena in Eibela. Grant Aiton**

10:30–10:45 **Theoretical and methodological considerations on building a corpus of Tundra Nenets. Nikolett Mus and Réka Metzger**

10:45–11:00 *Expanding the JHU Bible Corpus for Machine Translation of the Indigenous Languages of North America*
Garrett Nicolai, Edith Coates, Ming Zhang and Miikka Silfverberg

11:00–11:15 *The language documentation quartet*
Simon Musgrave and Nick Thieberger

11:15–11:30 *LARA in the Service of Revivalistics and Documentary Linguistics: Community Engagement and Endangered Languages*
Ghil'Ad Zuckermann, Sigurður Vigfússon, Manny Rayner, Neasa Ní Chiaráin, Nedelina Ivanova, Hanieh Habibi and Branislav Bédi

11:30–11:45 *Fossicking in dominant language teaching: Javanese and Indonesian 'low' varieties in language teaching resources*
Zara Maxwell-Smith

11:45–12:00 *Computational Analysis versus Human Intuition: A Critical Comparison of Vector Semantics with Manual Semantic Classification in the Context of Plains Cree*
Daniel Dacanay, Atticus Harrigan and Antti Arppe

12:00–15:00 ***Break***

15:00–15:15 *The Usefulness of Bibles in Low-Resource Machine Translation*
Ling Liu, Zach Ryan and Mans Hulden

15:15–15:30 *User-friendly Automatic Transcription of Low-resource Languages: Plugging ESPnet into Elpis*
Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques and Nathan Hill

15:30–15:45 *The Relevance of the Source Language in Transfer Learning for ASR*
Nils Hjortnaes, Niko Partanen, Michael Rießler and Francis M. Tyers

Tuesday, March 2nd, 2021 (continued)

15:45–16:00 *Shared Digital Resource Application within Insular Scandinavian*
Hinrik Hafsteinsson and Anton Karl Ingason

16:00–16:15 mScrabble: A Multilingual Scrabble and Lexicon Collector. Ritesh Kumar, Bornini Lahiri and Siddharth Singh

16:15–16:30 "Mansi corPŌS": Archiving Mansi written materials published between 1937 and 2020. Csilla Horvath

16:30–16:45 Lexeme: the Concept of System and the Creation of Speech Corpora for Two Endangered Languages. Ivan Ubaleht

Wednesday, March 3rd, 2021

10:00–10:15 *Towards an Open Source Finite-State Morphological Analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl*
Robert Pugh and Francis Tyers

10:15–10:30 *Integrating Automated Segmentation and Glossing into Documentary and Descriptive Linguistics*
Sarah Moeller and Mans Hulden

10:30–10:45 Advances in Low Resource and Endangered Languages. Evelyne Tzoukermann and Jason D. Duncan

10:45–11:00 *Developing a Shared Task for Speech Processing on Endangered Languages*
Gina-Anne Levow, Emily Ahn and Emily M. Bender

Wednesday, March 3rd, 2021 (continued)

11:00–11:10 **A Digital Corpus of St. Lawrence Island Yupik. Lane Schwartz, Emily Chen, Hyunji Hayley Park, Edward Jahn and Sylvia L.R. Schreiner**

11:10–11:20 **Migration of Small and Endangered Languages into the Wikipedia. Armin Hoenen and Marc D. Rahn**

11:20–12:20 ***Discussion (topic will be announced)***