ComputEL-3

Proceedings of the

3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages

Volume 1 (Papers)

February 26–27, 2019
Honolulu, Hawai‘i, USA

Support:

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

UNIVERSITY OF ALBERTA

University at Buffalo

University of Colorado Boulder

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UNT

UNIVERSITY OF NORTH TEXAS

ISBN XXX-X-XXXXXX-XX-X

Preface

These proceedings contain the papers presented at the 3rd Workshop on the Use of Computational Methods in the Study of Endangered languages held in Hawai'i at Mānoa, February 26–27, 2019. As the name implies, this is the third workshop held on the topic—the first meeting was co-located with the ACL main conference in Baltimore, Maryland in 2014 and the second one in 2017 was co-located with the 5th International Conference on Language Documentation and Conservation (ICLDC) at the University of Hawai'i 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 34 submissions as papers or extended abstracts. After a thorough review process, 12 of the submissions were selected for this volume as papers (35%) and an additional 7 were accepted as extended abstracts which appear in Volume 2 of the workshop proceedings. The organizing committee would like to thank the program committee for their thoughtful input on the submissions. We are also grateful to the NSF for funding part of the workshop (award #1550905), and the Social Sciences and Humanities Research Council (SSHRC) of Canada for supporting the workshop through their Connections Outreach Grant #611-2016-0207.

ANTTI ARPPE
JEFF GOOD
MANS HULDEN
JORDAN LACHLER
ALEXIS PALMER
LANE SCHWARTZ
MIIKKA SILFVERBERG

Organizing Committee:

Antti Arppe (University of Alberta)
Jeff Good (University at Buffalo)
Mans Hulden (University of Colorado)
Jordan Lachler (University of Alberta)
Alexis Palmer (University of North Texas)
Lane Schwartz (University of Illinois at Urbana-Champaign)
Miikka Silfverberg (University of Helsinki)

Program Committee:

Oliver Adams (Johns Hopkins University)
Antti Arppe (University of Alberta)
Dorothee Beermann (Norwegian University of Science and Technology)
Emily M. Bender (University of Washington)
Martin Benjamin (Kamusi Project International)
Steven Bird (University of Melbourne)
Emily Chen (University of Illinois at Urbana-Champaign)
Andrew Cowell (University of Colorado Boulder)
Christopher Cox (Carleton University)
Robert Forkel (Max Planck Institute for the Science of Human History)
Jeff Good (University at Buffalo)
Michael Wayne Goodman (University of Washington)
Harald Hammarström (Max Planck Institute for Psycholinguistics, Nijmegen)
Mans Hulden (University of Colorado Boulder)
Anna Kazantseva (National Research Council of Canada)
František Kratochvíl (Palacký University)
Jordan Lachler (University of Alberta)
Terry Langendoen (National Science Foundation)
Krister Lindén (University of Helsinki)
Worthy N Martin (University of Virginia)
Michael Maxwell (University of Maryland, CASL)
Steven Moran (University of Zurich)
Graham Neubig (Carnegie Mellon University)
Alexis Palmer (University of North Texas)
Taraka Rama (University of Oslo)
Kevin Scannell (Saint Louis University)
Lane Schwartz (University of Illinois at Urbana-Champaign)
Miikka Silfverberg (University of Helsinki)
Richard Sproat (Google)
Nick Thieberger (University of Melbourne / ARC Centre of Excellence for the Dynamics of Language)

Laura Welcher (The Long Now Foundation)
Menzo Windhouwer (KNAW Humanities Cluster — Digital Infrastructure)

Table of Contents

Conference Program

Tuesday, February 26th, 2019

08:30–09:00 ***Imin Conference Center, Asia Room/Arrival, coffee and chat***

09:00–09:15 ***Opening remarks***

First morning: Tools and processes for language documentation and description, Corpus creation

09:15–09:45 *An Online Platform for Community-Based Language Description and Documentation*
Rebecca Everson, Wolf Honoré and Scott Grimm

09:45–10:15 *Developing without developers: choosing labor-saving tools for language documentation apps*
Luke Gessler

10:15–10:45 *Future Directions in Technological Support for Language Documentation*
Daan van Esch, Ben Foley and Nay San

10:45–11:15 ***Break***

11:15–11:45 Towards a General-Purpose Linguistic Annotation Backend. Graham Neubig, Patrick Littell, Chian-Yu Chen, Jean Lee, Zirui Li, Yu-Hsiang Lin and Yuyan Zhang

11:45–12:15 *OCR evaluation tools for the 21st century*
Eddie Antonio Santos

Tuesday, February 26th, 2019 (continued)

12:15–12:45 **Building a Common Voice Corpus for Laiholh (Hakha Chin). Kelly Berkson, Samson Lotven, Peng Hlei Thang, Thomas Thawngza, Zai Sung, James Wamsley, Francis M. Tyers, Kenneth Van Bik, Sandra Kübler, Donald Williamson and Matthew Anderson**

12:45–14:15 ***Lunch***

First afternoon: Language technologies – lexical and syntactic

14:15–14:45 **Digital Dictionary Development for Torwali, a less-studied language: Process and Challenges. Inam Ullah**

14:45–15:15 *Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang*
Olga Zamaraeva, Kristen Howell and Emily M. Bender

15:15–15:45 **Applying Support Vector Machines to POS tagging of the Ainu Language. Karol Nowakowski, Michal Ptaszynski, Fumito Masui and Yoshio Momouchi**

15:45–16:15 ***Break***

16:15–16:45 *Finding Sami Cognates with a Character-Based NMT Approach*
Mika Hämäläinen and Jack Rueter

16:45–17:15 *Seeing more than whitespace — Tokenization and disambiguation of potential compounds in North Sámi grammar checking*
Linda Wiechetek, Sjur Nørstebø Moshagen and Kevin Brubeck Unhammer

17:15–20:00 ***Dinner***

Wednesday, February 27th, 2019

09:00–09:30 ***Arrival, coffee and chat***

Second morning: Use (and reuse) of corpora and collections, Language technologies – speech and morphology

09:30–10:00 **Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. Rogier Blokland, Niko Partanen, Michael Rießler and Joshua Wilbur**

10:00–10:30 **A software-driven workflow for the reuse of language documentation data in linguistic studies. Stephan Druskat and Kilu von Prince**

10:30–11:00 ***Break***

11:00–11:30 *Corpus of usage examples: What is it good for?*
Timofey Arkhangelskiy

11:30–12:00 *A Preliminary Plains Cree Speech Synthesizer*
Atticus Harrigan, Antti Arppe and Timothy Mills

12:00-12:30 *A biscriptual morphological transducer for Crimean Tatar*
Francis M. Tyers, Jonathan Washington, Darya Kavitskaya, Memduh Gökırmak, Nick Howell and Remziye Berberova

12:30–14:00 ***Lunch***

Wednesday, February 27th, 2019 (continued)

Second afternoon: Language technologies – speech and morphology

14:00-14:30 *Improving Low-Resource Morphological Learning with Intermediate Forms from Finite State Transducers*
Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell and Mans Hulden

14:30–15:00 Bootstrapping a Neural Morphological Generator from Morphological Analyzer Output for Inuktitut. Jeffrey Micher

15:00–15:30 *Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik from a Finite-State Transducer*
Lane Schwartz, Emily Chen, Benjamin Hunt and Sylvia L.R. Schreiner

15:30–16:00 ***Break***

16:00–17:00 ***Discussions, Looking ahead: CEL-4 and new ACL Special Interest Group***