ComputEL-2

Proceedings of the

2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages

March 6–7, 2017
Honolulu, Hawai‘i, USA

Support:

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

UNIVERSITY of HAWAI'I® MĀNOA

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

Preface

These proceedings contain the papers presented at the 2nd Workshop on the Use of Computational Methods in the Study of Endangered languages held in Honolulu, March 6–7, 2017. The workshop itself was co-located and took place after the 5th International Conference on Language Documentation and Conservation (ICLDC) at the University of Hawai'i at Mānoa. As the name implies, this is the second workshop held on the topic—the previous meeting was co-located with the ACL main conference in Baltimore, Maryland in 2014.

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 39 submissions as long papers, short papers, or extended abstracts, of which 23 were selected for this volume (59%). In the proceedings, all papers are either short ($\leq$5 pages) or long ($\leq$9 pages). In addition, the workshop also features presentations from representatives of the National Science Foundation (NSF). Two panel dicussions on the topic of interaction between computational linguistics and the documentation and revitalization community as well as future planning of ComputEL underlined the demand and necessity of a workshop of this nature.

The organizing committee would like to thank the program committee for their thoughtful input on the submissions, as well as the organizers of ICLDC. We are also grateful to the NSF for funding part of the workshop (awards #1404352 and #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

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)

Program Committee:

Steve Abney (University of Michigan)
Dorothee Beermann (Norwegian University of Science and Technology)
Emily Bender (University of Washington)
Martin Benjamin (Kamusi Project International)
Damir Cavar (Indiana University)
Chris Cox (Carleton University)
Joel Dunham (The University of British Columbia)
Judith Klavans (University of Maryland)
Terry Langendoen (National Science Foundation)
Lori Levin (Carnegie Mellon University)
Will Lewis (Microsoft Research)
Worthy Martin (University of Virginia)
Michael Maxwell (Center for Advanced Study of Language)
Detmar Meurers (University of Tübingen)
Steve Moran (University of Zurich)
Sebastian Nordhoff (Glottotopia)
Kevin Scannell (Saint Louis University)
Gary Simons (SIL International)
Richard Sproat (Google, Inc.)
Trond Trosterud (University of Tromsø—The Arctic University of Norway)

Table of Contents

Conference Program

Monday, March 6th, 2017

09:00–09:15 University of Hawai‘i/ICLDC5 Welcome and opening remarks

09:15–09:30 Colleen Fitzgerald (Program Director for Documenting Endangered Languages (DEL)): Special greetings from the National Science Foundation

09:30–10:00 *A Morphological Parser for Odawa*
Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen and Trond Trosterud

10:00–10:30 *Creating lexical resources for polysynthetic languages—the case of Arapaho*
Ghazaleh Kazeminejad, Andrew Cowell and Mans Hulden

10:30–11:00 *From Small to Big Data: paper manuscripts to RDF triples of Australian Indigenous Vocabularies*
Nick Thieberger and Conal Tuohy

11:00–11:30 Break: coffee and tea available

11:30–12:00 *Issues in digital text representation, on-line dissemination, sharing and re-use for African minority languages*
Emmanuel Ngué Um

12:00–12:30 *Developing collection management tools to create more robust and reliable linguistic data*
Gary Holton, Kavon Hooshiar and Nick Thieberger

12:30–13:00 *STREAMLInED Challenges: Aligning Research Interests with Shared Tasks*
Gina-Anne Levow, Emily M. Bender, Patrick Littell, Kristen Howell, Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, David Inman, Michael Maxwell, Michael Tjalve and Fei Xia

13:00–14:15 Lunch

14:15–14:45 *Work With What You've Got*
Lucy Bell and Lawrence Bell

14:45–15:45 Panel Discussion

15:45–16:15 Break: coffee and tea available

16:15–16:45 Poster boasters: 2–3 minutes for each poster presenter

Monday, March 6th, 2017 (continued)

16:45–18:30: Poster session and late afternoon buffet

Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection
Antti Arppe, Marie-Odile Junker and Delasie Torkornoo

Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region
Ciprian Gerstenberger, Niko Partanen and Michael Rießler

Inferring Case Systems from IGT: Enriching the Enrichment
Kristen Howell, Emily M. Bender, Michel Lockwood, Fei Xia and Olga Zamaraeva

Case Studies in the Automatic Characterization of Grammars from Small Wordlists
Jordan Kodner, Spencer Kaplan, Hongzhi Xu, Mitchell P. Marcus and Charles Yang

Endangered Data for Endangered Languages: Digitizing Print dictionaries
Michael Maxwell and Aric Bills

A computationally-assisted procedure for discovering poetic organization within oral tradition
David Meyer

Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network
Jeffrey Micher

Click reduction in fluent speech: a semi-automated analysis of Mangetti Dune !Xung
Amanda Miller and Micha Elsner

DECCA Repurposed: Detecting transcription inconsistencies without an orthographic standard
C. Anton Rytting and Julie Yelle

Jejueo talking dictionary: A collaborative online database for language revitalization
Moira Saltzman

Computational Support for Finding Word Classes: A Case Study of Abui
Olga Zamaraeva, František Kratochvíl, Emily M. Bender, Fei Xia and Kristen Howell

Tuesday, March 7th, 2017

08:45–09:00 Arrival, breakfast & chat

09:00–09:30 Q&A on National Science Foundation funding and application process (Colleen Fitzgerald and Terry Langendoen)

09:30–10:00 *Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages*
Patrick Littell, Aidan Pine and Henry Davis

10:00–10:30 *Connecting Documentation and Revitalization: A New Approach to Language Apps*
Alexa N. Little

10:30–11:00 Break: coffee and tea available

11:00–11:30 *Developing a Suite of Mobile Applications for Collaborative Language Documentation*
Mat Bettinson and Steven Bird

11:30–12:00 *Cross-language forced alignment to assist community-based linguistics for low resource languages*
Timothy Kempton

12:00–12:30 *A case study on using speech-to-translation alignments for language documentation*
Antonios Anastasopoulos and David Chiang

12:30–14:00 Lunch

14:00–15:00 ComputEL-3: Planning for the future