ACL 2014

Comput*EL* 2014

2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages

Proceedings of the Workshop

June 26, 2014
Baltimore, Maryland, USA

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 978-1-941643-07-5

Preface

Contemporary efforts to document the world's endangered languages—often going under the rubric of *documentary linguistics*—are dependent on the widespread availability of modern recording technologies, in particular digital audio and video recording devices and software to annotate the recordings that such devices produce. However, despite well over a decade of dedicated funding efforts aimed at the documentation of endangered languages, the technological landscape that supports the work of those involved in this research remains fragmented, and the promises of new technology remain largely unfulfilled. Moreover, the efforts of computer scientists, on the whole, are mostly disconnected from the day-to-day work of documentary linguists, making it difficult for the knowledge of each group to inform the other. On the one hand, this deprives documentary linguists of tools making use of the latest research results to speed up the time-consuming task of describing an underdocumented language. On the other hand, it severely limits the ability of computational linguists to test their methods on the full range of world's linguistic diversity.

Despite the concerns listed above, recent efforts do indicate that there is significant potential in collaboration between computational linguists (and other computer scientists) and linguists working on endangered languages. For instance, machine labeling and active learning can make the process of textual analysis for low-resource languages more efficient, and state-of-the-art tools in grammar engineering can be applied at a relatively low cost to new languages that are typologically divergent from those that primarily informed their design. Moreover, new models of data collection based on the ubiquity of low-cost, networkable devices with recording capabilities, such as smartphones, show the extent to which the barriers to collecting significant amounts of primary data have fallen in recent years, and it has similarly been found that the pairing of crowdsourcing and machine translation techniques can yield useful results for low resource languages in a short time frame. Research along these latter lines, in particular, indicates that computationally-driven advances in the documentation of the world's languages may need to rely as much on clever engineering and user interface solutions as on methods for processing language data developed within computational linguistics proper, in a manner parallel to efforts in other domains that have considered how new online services can be used to facilitate computational linguistic research.

A different set of activities within the documentary linguistics community involving the increasing use of open standards for encoding language data is also significant in this regard. For instance, in the last decade, standardized XML formats have become more widely used to encode text annotations and lexical data. This facilitates the reuse of documentary materials. Even in the absence of the use of such standards, significant results have been achieved in gathering structured data from materials placed on the web. As more data becomes available in standardized forms, there will only be increased potential for building new kinds of language resources.

The papers in these proceedings cover the full range of work at the intersection of computational and endangered language linguistics. Some contributions come from scholars primarily identifying as computer scientists who are exploring how tools developed in their areas of expertise can be applied to endangered language research. Others derive from the work of individuals primarily identifying as descriptive linguists who are reporting on the results of the application of new computational methods to traditional language work. There is also a division among contributions which have more practical orientations versus programmatic ones, with topics ranging from discussion of software under development to high-level considerations of where our research priorities should lie.

We would like to thank those who made this workshop possible: the ACL staff, 2014 annual meeting organizers, the program committee, workshop participants, and research assistant Daniel Fox. Further support came from National Science Foundation Award Nos. BCS-1404352 and IIS-1027289.

Jeff Good, Julia Hirschberg, and Owen Rambow

Organizers:

Jeff Good, University at Buffalo, USA
Julia Hirschberg, Columbia University, USA
Owen Rambow, Columbia University, USA

Program Committee:

Steven Abney, University of Michigan, USA
Helen Aristar-Dry, University of Texas at Austin, USA
Alexandre Arkhipov, Moscow State University, Russia
Timothy Baldwin, The University of Melbourne, Australia
Dorothee Beermann, Norwegian University of Science and Technology, Norway
Emily M. Bender, University of Washington, USA
Andrea Berez, University of Hawai'i, USA
Steven Bird, The University of Melbourne, Australia
Guy De Pauw, University of Antwerp, Belgium
Harald Hammarström, Max Planck Institute for Psycholinguistics, The Netherlands
Judith Klavans, University of Maryland, USA
Terry Langendoen, University of Arizona, USA
Lori Levin, Carnegie Mellon University, USA
William D. Lewis, Microsoft Research, USA
Worthy Martin, University of Virginia, USA
Mike Maxwell, Center for the Advanced Study of Language, USA
Steven Moran, University of Zurich, Switzerland
Alexander Nakhimovsky, Colgate University, USA
Alexis Palmer, Saarland University, Germany
Kevin Scannell, Saint Louis University, USA
Gary Simons, SIL International, USA
Nick Thieberger, The University of Melbourne, Australia
Paul Trilsbeek, Max Planck Institute for Psycholinguistics, The Netherlands
Doug Whalen, CUNY Graduate Center, USA
Menzo Windhouwer, Max Planck Institute for Psycholinguistics, The Netherlands
Fei Xia, University of Washington, USA

Sponsor:

US National Science Foundation (award nos. BCS-1404352 and IIS-1027289)

Table of Contents

Conference Program

Thursday, June 26, 2014

9:00–9:10 Introduction

Paper Session 1: Computational Tools for Endangered Languages Research

9:10–9:30 *Aikuma: A Mobile App for Collaborative Language Documentation*
Steven Bird, Florian R. Hanke, Oliver Adams and Haejoong Lee

9:30–9:50 *Documenting Endangered Languages with the WordsEye Linguistics Tool*
Morgan Ulinski, Anusha Balakrishnan, Daniel Bauer, Bob Coyne, Julia Hirschberg and Owen Rambow

9:50–10:10 *Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages*
Martin Benjamin and Paula Radetzky

10:10–10:30 *LingSync & the Online Linguistic Database: New Models for the Collection and Management of Data for Language Communities, Linguists and Language Learners*
Joel Dunham, Gina Cook and Joshua Horner

10:30–11:00 Coffee Break

Paper Session 2: Applying Computational Methods to Endangered Languages

11:00–11:30 *Modeling the Noun Morphology of Plains Cree*
Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen and Trond Trosterud

11:30–12:00 *Learning Grammar Specifications from IGT: A Case Study of Chintang*
Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman and Fei Xia

12:00–12:30 *Creating Lexical Resources for Endangered Languages*
Khang Nhut Lam, Feras Al Tarouti and Jugal Kalita

12:30–14:00 Lunch

14:00–15:00 Posters and Demonstrations of Tools Presented in Paper Session 1

Estimating Native Vocabulary Size in an Endangered Language
Timofey Arkhangelskiy

Thursday, June 26, 2014 (continued)

InterlinguaPlus Machine Translation Approach for Local Languages: Ekegusii & Swahili
Edward Ombui, Peter Wagacha and Wanjiku Ng'ang'a

Building and Evaluating Somali Language Corpora
Abdillahi Nimaan

Paper Session 3: Infrastructure and Community Development for Computational Research on Endangered Languages

15:00–15:30 *SeedLing: Building and Using a Seed corpus for the Human Language Project*
Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer and Michaela Regneri

15:30–16:00 Coffee Break

16:00–16:20 *Short-Term Projects, Long-Term Benefits: Four Student NLP Projects for Low-Resource Languages*
Alexis Palmer and Michaela Regneri

16:20–16:50 *Data Warehouse, Bronze, Gold, STEC, Software*
Doug Cooper

16:50–17:20 *Time to Change the "D" in "DEL"*
Stephen Beale

17:20–17:30 Concluding Remarks