RAIL 2023

Fourth workshop on Resources for African Indigenous Languages (RAIL)

Proceedings of the Workshop

May 6, 2023

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 978-1-959429-58-6

Preface

Africa is a multilingual continent with an estimation of 1500 to 2000 indigenous languages. Many of the languages currently have no or very limited language resources available and are often structurally quite different from more well-resourced languages, therefore requiring the development and use of specialized techniques. To bring together and emphasize research in these areas, the Resources for African Indigenous Languages (RAIL) workshop series aims to provide an interdisciplinary platform for researchers working on resources (data collections, tools, etc.) specifically targeted towards African indigenous languages. These events provide an overview of the current state-of-the-art and emphasize the availability of African indigenous language resources, including both data and tools.

With the UNESCO-supported Decade of Indigenous Languages, there is currently much interest in indigenous languages. The Permanent Forum on Indigenous Issues mentioned that "40 percent of the estimated 6,700 languages spoken around the world were in danger of disappearing" and the "languages represent complex systems of knowledge and communication and should be recognized as a strategic national resource for development, peace building and reconciliation."

This year's RAIL workshop is the fourth in a series. The first workshop was co-located with the Language Resources and Evaluation Conference (LREC) in 2020, whereas the second RAIL workshop, in 2021, was co-located with the Digital Humanities Association of Southern Africa (DHASA) conference. Both of these events were virtual. The third RAIL workshop was co-located with the tenth Southern African Microlinguistics Workshop and took place in person in Potchefstroom, South Africa.

Previous RAIL workshops showed that the presented problems (and solutions) are typically not only applicable to African languages. Many issues are also relevant to other low-resource languages, such as different scripts and properties like tone. As such, these languages share similar challenges. This allows for researchers working on these languages with such properties (including non-African languages) to learn from each other, especially on issues pertaining to language resource development.

For the fourth RAIL workshop, in total, nineteen very high-quality submissions were received. Out of these, fourteen submissions were selected for presentation in the workshop using double blind review. Additionally, one presentation that is published in EACL's Findings proceedings is incorporated in the programme as well. The RAIL workshop took place as a full day workshop in Dubrovnik, Croatia. It was co-located with the EACL 2023 conference, the seventeenth Conference of the European Chapter of the Association for Computational Linguistics. Each presentation consisted of 25 minutes (including time for discussion).

This publication adheres to South Africa's DHET's 60% rule, authors in the proceedings come from a wide range of institutions.

The workshop has "Impact of impairments on language resources" as its theme, but submissions on any topic related to properties of African indigenous languages were considered. In fact, several suggested topics for the workshop were mentioned in the call for papers:

- Digital representations of linguistic structures
- Descriptions of corpora or other data sets of African indigenous languages
- Building resources for (under resourced) African indigenous languages
- Developing and using African indigenous languages in the digital age
- Effectiveness of digital technologies for the development of African indigenous languages

- Revealing unknown or unpublished existing resources for African indigenous languages
- Developing desired resources for African indigenous languages
- Improving quality, availability and accessibility of African indigenous language resources

The goals for the workshop are:

- to bring together researchers who are interested in showcasing their research and thereby boosting the field of African indigenous languages,
- to create the conditions for the emergence of a scientific community of practice that focuses on data, as well as tools, specifically designed for or applied to indigenous languages found in Africa,
- to create conversations between academics and researchers in different fields such as African indigenous languages, computational linguistics, sociolinguistics, and language technology, and
- to provide an opportunity for the African indigenous languages community to identify, describe and share their Language Resources.

We would like to mention explicitly that the term "indigenous languages" used in the RAIL workshop is intended to refer to non-colonial languages (in this case those used in Africa). In no way is this term used to cause any harm or discomfort to anyone. Many of these languages were or still are marginalized and the aim of the workshop is to bring attention to the creation, curation, and development of resources for these languages in Africa.

The organizers would like to thank the authors who submitted publications and the programme committee who provided feedback on the quality and the content of the submissions.

The RAIL organizing committee and editors of the proceedings:

- Rooweither Mabuya, South African Centre for Digital Language Resources
- Don Mthobela, Cam Foundation
- Mmasibidi Setaka, South African Centre for Digital Language Resources
- Menno van Zaanen, South African Centre for Digital Language Resources

Organizing Committee

Organizers

Rooweither Mabuya, South African Centre for Digital Language Resources
Don Mthobela, Cam Foundation
Mmasibidi Setaka, South African Centre for Digital Language Resources
Menno Van Zaanen, South African Centre for Digital Language Resources

Program Committee

Chairs

Rooweither Mabuya, South African Centre for Digital Language Resources
Don Mthobela, Cam Foundation
Mmasibidi Setaka, South African Centre for Digital Language Resources
Menno Van Zaanen, South African Centre for Digital Language Resources

Program Committee

Gilles-Maurice De Schryver, Ghent University
Febe De Wet, Stellenbosch University
Sibonelo Dlamini, University of KwaZulu-Natal
Roald Eiselen, Centre for Text Technology, North-West University
Tanja Gaustad, Centre for Text Technology, North-West University
Marissa Griesel, University of South Africa
Ayodele James Akinola, Chrisland University
C. Maria Keet, University of Cape Town
Papi Lemeko, Central University of Technology Free State
Vukosi Marivate, University of Pretoria, Lelapa AI
Muzi Matfunjwa, South African Centre for Digital Language Resources
Dimakatso Mathe, University of Limpopo
Innocentia Mhlambi, University of the Witwatersrand
Emmanuel Ngue Um, University of Yaoundé I
Makanjuola Ogunleye, Virginia Polytechnic Institute and State University
Tunde Ope-davies, Centre for Digital Humanities, University of Lagos
Sara Petrollino, Leiden University
Pule Phindane, Central University of Technology
Mpho Raborife, University of Johannesburg
Lorraine Shabangu, Wits University
Johannes Sibeko, Nelson Mandela University
Hussein Suleman, University of Cape Town
Elsabe Taljard, University of Pretoria
Valencia Wagner, Sol Plaatje University
Friedel Wolff, South African Centre for Digital Language Resources

Table of Contents

Program

Saturday, May 6, 2023

08:30 - 09:00 *Registration and opening remarks*

09:00 - 10:15 *Morning Session 1*

IsiXhosa Intellectual Traditions Digital Archive: Digitizing isiXhosa texts from 1870-1914
Jonathan Schoots, Amandla Ngwendu, Jacques De Wet and Sanjin Muftic

Preparing the Vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora
Richard Lastrucci, Jenalea Rajab, Matimba Shingange, Daniel Njini and Vukosi Marivate

Automatic Spell Checker and Correction for Under-represented Spoken Languages: Case Study on Wolof
Thierno Ibrahima Cissé and Fatiha Sadat

10:15 - 10:55 *Morning tea break*

10:55 - 12:30 *Morning Session 2*

SpeechReporting Corpus: annotated corpora of West African traditional narratives
Ekaterina Aplonova, Izabela Jordanoska, Timofey Arkhangelskiy and Tatiana Nikitina

Analyzing political formation through historical isiXhosa text analysis: Using frequency analysis to examine emerging African Nationalism in South Africa
Jonathan Schoots

Unsupervised Cross-lingual Word Embedding Representation for English-isiZulu
Derwin Ngomane, Rooweither Mabuya, Jade Abbott and Vukosi Marivate

Investigating Sentiment-Bearing Words- and Emoji-based Distant Supervision Approaches for Sentiment Analysis
Ronny Mabokela, Mpho Roborife and Turguy Celik

12:30 - 14:00 *Lunch break*

14:00 - 15:40 *Afternoon Session 1*

Saturday, May 6, 2023 (continued)

Towards a Swahili Universal Dependency Treebank: Leveraging the Annotations of the Helsinki Corpus of Swahili
Kenneth Steimel, Sandra Kübler and Daniel Dakota

Evaluating the Sesotho rule-based syllabification system on Sepedi and Setswana words
Johannes Sibeko and Mmasibidi Setaka

Deep learning and low-resource languages: How much data is enough? A case study of three linguistically distinct South African languages
Roald Eiselen and Tanja Gaustad

Comparing methods of orthographic conversion for Bàsàá, a language of Cameroon
Alexandra O'neil, Daniel Swanson, Robert Pugh, Francis Tyers and Emmanuel Ngue Um

15:40 - 16:20 *Afternoon tea break*

16:20 - 18:00 *Afternoon Session 2*

Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities
Tolúlọpẹ́ Ògúnrẹ̀mí, Dan Jurafsky and Christopher D. Manning

Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities
Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova and Seid Muhie Yimam

A Corpus-Based List of Frequently Used Words in Sesotho
Johannes Sibeko and Orphée De Clercq

Vowels and the Igala Language Resources
Mahmud Momoh

18:00 - 18:05 *Closing statements*