LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

First workshop on
Resources for African Indigenous Languages
(RAIL)

PROCEEDINGS

Editors:

Rooweither Mabuya, Phathutshedzo Ramukhadi, Mmasibidi Setaka, Valencia Wagner, Menno van Zaanen

Proceedings of the LREC 2020 first workshop on Resources for African Indigenous Languages (RAIL)

Edited by:
Rooweither Mabuya, Phathutshedzo Ramukhadi, Mmasibidi Setaka, Valencia Wagner, and Menno van Zaanen

ISBN: 979-10-95546-60-3
EAN: 9791095546603

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Introduction

Africa is a multilingual continent with an estimation of 1500 to 2000 indigenous languages. Many of these 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.

The Resources for African Indigenous Languages (RAIL) workshop is an interdisciplinary platform for researchers working on resources (data collections, tools, etc.) specifically targeted towards African indigenous languages to 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 International Year 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.” As such, the workshop falls within one of the hot topic areas of this year’s conference: “Less Resourced and Endangered Languages”.

In total, 24, in general, very high quality submissions were received. Out of these 9 submissions were selected using double blind review for presentation in the workshop. Unfortunately, due to the Covid-19 pandemic, the physical workshop had to be cancelled, however, it is replaced by a virtual workshop.

The topics on which the call for papers was issued are the following:

- Computational linguistics for African indigenous languages
- 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.

Organizers:

Rooweither Mabuya
Phathutshedzo Ramukhadi
Mmasibidi Setaka
Valencia Wagner
Menno van Zaanen
South African centre for Digital Language Resources (SADiLaR), South Africa

Program Committee:

Richard Ajah, University of Uyo, Nigeria
Ayodele James Akinola, Chrisland University, Nigeria
Felix Ameka, Leiden University, the Netherlands
Sonja Bosch, University of South Africa, South Africa
Ibrahima Cissé, University of Humanities, Mali
Roald Eiselen, Eiselen software consulting, South Africa
Tanja Gaustad, Centre for Text Technology, South Africa
Elias Malete, University of the Free State, South Africa
Dimakatso Mathe, South African centre for Digital Language Resources, South Africa
Elias Mathipa, University of South Africa, South Africa
Fekede Menuta, Hawassa University, Ethiopia
Innocentia Mhlambi, Wits University, South Africa
Emmanuel Ngue Um, University of Yaoundé I, Cameroon
Guy de Pauw, Antwerp University and Textgain, Belgium
Sara Petrollino, Leiden University, the Netherlands
Pule Phindane, Central University of Technology, South Africa
Danie Prinsloo, University of Pretoria, South Africa
Martin Puttkammer, Centre for Text Technology, South Africa
Justus Roux, Stellenbosch University, South Africa
Msindisi Sam, Rhodes University, South Africa
Gilles-Maurice de Schryver, Ghent University, Belgium
Lorraine Shabangu, Bangula Lingo Centre, South Africa
Elsabé Taljard, University of Pretoria, South Africa

Table of Contents

Conference Program

9:00–9:10 ***Opening/Introduction***

09:10–09:30 *Endangered African Languages Featured in a Digital Collection: The Case of the Khomani San, Hugh Brody Collection*
Kerry Jones and Sanjin Muftic

09:30–09:50 *Usability and Accessibility of Bantu Language Dictionaries in the Digital Age: Mobile Access in an Open Environment*
Thomas Eckart, Sonja Bosch, Uwe Quasthoff, Erik Körner, Dirk Goldhahn and Simon Kaleschke

09:50–10:10 *Investigating an Approach for Low Resource Language Dataset Creation, Curation and Classification: Setswana and Sepedi*
Vukosi Marivate, Tshephisho Sefara and Abiodun Modupe

10:10–10:30 *Complex Setswana Parts of Speech Tagging*
Gabofetswe Malema, Boago Okgetheng, Bopaki Tebalo, Moffat Motlhanka and Goaletsa Rammidi

10:30–10:50 *Comparing Neural Network Parsers for a Less-resourced and Morphologically-rich Language: Amharic Dependency Parser*
Binyam Ephrem Seyoum, Yusuke Miyao and Baye Yimam Mekonnen

10:50–11:10 *Mobilizing Metadata: Open Data Kit (ODK) for Language Resource Development in East Africa*
Richard Griscom

11:10–11:40 ***Coffee break***

11:40–12:00 *A Computational Grammar of Ga*
Lars Hellan

12:00–12:20 *Navigating Challenges of Multilingual Resource Development for Under-Resourced Languages: The Case of the African Wordnet Project*
Marissa Griesel and Sonja Bosch

12:20–12:40 *Building Collaboration-based Resources in Endowed African Languages: Case of NTeALan Dictionaries Platform*
Elvis Mboning Tchiaze, Jean Marc Bassahak, Daniel Baleba, Ornella Wandji and Jules Assoumou

12:40–13:00 ***Closing***