LREC-COLING 2024

The Fifth Workshop on
Resources for African Indigenous Languages
@LREC-COLING-2024 (RAIL)

Workshop Proceedings

Editors
Rooweither Mabuya, Muzi Matfunjwa, Mmasibidi Setaka,
and Menno van Zaanen

25 May, 2024
Torino, Italia

Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL)

ISBN 978-2-493814-40-1
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

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 fifth in the series. The first RAIL 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 events were virtual. The third RAIL workshop was co-located with the tenth Southern African Microlinguistics Workshop (SAMWOP) and took place in person in 2022 in Potchefstroom, South Africa. The fourth RAIL workshop was co-located with the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL) in Dubrovnik, Croatia in 2023.

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 about language resource development.

For the fifth RAIL workshop, in total, 39 high-quality submissions were received. Out of these, 17 submissions (15 long papers and 2 short papers) were selected for presentation in the workshop. All submissions received three reviews using a double-blind review process. This RAIL workshop took place as a full day workshop in Lingotto Conference Centre, Torino, Italy on 25 May 2024. It was co-located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Each presentation consisted of 25 minutes for long papers (including time for discussion) and 10 minutes for short papers.

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

This RAIL workshop’s theme was “Creating resources for less-resourced languages”, but submissions on any topic related to properties of African indigenous languages were considered. 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 were:

- 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 workshop aims 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 manuscripts and the programme committee who provided feedback on the quality and content of the submissions.

The RAIL organizing committee and editors of the proceedings

- Rooweither Mabuya, South African Centre for Digital Language Resources
- Muzi Matfunjwa, South African Centre for Digital Language Resources
- Mmasibidi Setaka, South African Centre for Digital Language Resources
- Menno van Zaanen, South African Centre for Digital Language Resources

Organizing Committee

- Rooweither Mabuya, South African Centre for Digital Language Resources
- Muzi Matfunjwa, South African Centre for Digital Language Resources
- Mmasibidi Setaka, South African Centre for Digital Language Resources
- Menno van Zaanen, South African Centre for Digital Language Resources

Programme Committee

- Andiswa Bukula, South African Centre for Digital Language Resources
- Ayodele Akinola, Chrisland University
- Benito Trollip, South African Centre for Digital Language Resources
- Deon du Plessis, South African Centre for Digital Language Resources
- Elias Malete, University of the Free State
- Elsabe Taljard, University of Pretoria
- Emmanuel Ngue Um, University of Yaoundé I
- Febe De Wet, Stellenbosch University
- Friedel Wolff, South African Centre for Digital Language Resources
- Heather Brookes, University of Stellenbosch
- Hussein Suleman, University of Cape Town
- Innocentia Mhlambi, University of the Witwatersrand
- Johannes Sibeko, Nelson Mandela University
- Juan Steyn, South African Centre for Digital Language Resources
- Kaka Mokakale, North-West University
- Laurette Marais, Council for Scientific and Industrial Research
- Malefu Mahloane, University of the Free State
- Maria Keet, University of Cape Town
- Marissa Griesel, University of South Africa
- Martin Puttkammer, North-West University
- Menno van Zaanen, South African Centre for Digital Language Resources
- Mmasibidi Setaka, South African Centre for Digital Language Resources
- Mpho Raborife, University of Johannesburg
- Muzi Matfunjwa, South African Centre for Digital Language Resources

- Nomsebenzi Malele, University of South Africa
- Nuette Heyns, North-West University
- Papi Lemeko, Central University of Technology
- Pule Phindane, Central University of Technology
- Roald Eiselen, North-West University
- Rooweither Mabuya, South African Centre for Digital Language Resources
- Sibonelo Dlamini, University of KwaZulu-Natal
- Tanja Gaustad, North-West University
- Temitope Kekere, University of Pretoria
- Tunde Ope-Davies, University of Lagos

Table of Contents

Workshop Program

Saturday, May 25, 2024

09:00–09:05 ***Opening***

09:05–09:30 *Doing Phonetics in the Rift Valley: Sound Systems of Maasai, Iraqw and Hadza*
Alain Ghio, Didier Demolin, Michael Karani and Yohann Meynadier

09:30–09:55 *Kallaama: A Transcribed Speech Dataset about Agriculture in the Three Most Widely Spoken Languages in Senegal*
Elodie Gauthier, Aminata Ndiaye and Abdoulaye Guissé

09:55–10:20 *Long-Form Recordings to Study Children's Language Input and Output in Under-Resourced Contexts*
Joseph R. Coffey and Alejandrina Cristia

10:20–10:30 *Developing Bilingual English-Setswana Datasets for Space Domain*
Tebatso G. Moape, Sunday Olusegun Ojo and Oludayo O. Olugbara

10:30–11:00 ***Coffee break***

11:00–11:25 *Compiling a List of Frequently Used Setswana Words for Developing Readability Measures*
Johannes Sibeko

11:25–11:50 *A Qualitative Inquiry into the South African Language Identifier's Performance on YouTube Comments.*
Nkazimlo N. Ngcungca, Johannes Sibeko and Sharon Rudman

11:50–12:15 *The First Universal Dependency Treebank for Tswana: Tswana-Popapolelo*
Tanja Gaustad, Ansu Berg, Rigardt Pretorius and Roald Eiselen

12:15–12:40 *Adapting Nine Traditional Text Readability Measures into Sesotho*
Johannes Sibeko and Menno van Zaanen

12:40–13:05 *Bootstrapping Syntactic Resources from isiZulu to Siswati*
Laurette Marais, Laurette Pretorius and Lionel Clive Posthumus

13:05–14:20 ***Lunch break***

Saturday, May 25, 2024 (continued)

14:20–14:45 *Early Child Language Resources and Corpora Developed in Nine African Languages by the SADiLaR Child Language Development Node*
Michelle J. White, Frenette Southwood and Sefela Londiwe Yalala

14:45–15:10 *Morphological Synthesizer for Ge’ez Language: Addressing Morphological Complexity and Resource Limitations*
Gebrearegawi Gebremariam Gidey, Hailay Kidu Teklehaymanot and Gebregewergs Mezgebe Atsbha

15:10–15:35 *EthioMT: Parallel Corpus for Low-resource Ethiopian Languages*
Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh and Jugal Kalita

15:35–16:00 *Resources for Annotating Hate Speech in Social Media Platforms Used in Ethiopia: A Novel Lexicon and Labelling Scheme*
Nuhu Ibrahim, Felicity Mulford, Matt Lawrence and Riza Batista-Navarro

16:00–16:30 ***Coffee break***

16:30–16:55 *Low Resource Question Answering: An Amharic Benchmarking Dataset*
Tilahun Abedissa Taffa, Ricardo Usbeck and Yaregal Assabie

16:55–17:05 *The Annotators Agree to Not Agree on the Fine-grained Annotation of Hate-speech against Women in Algerian Dialect Comments*
Imane Guellil, Yousra Houichi, Sara Chennoufi, Mohamed Boubred, Anfal Yousra Boucetta and Faical Azouaou

17:05–17:30 *Advancing Language Diversity and Inclusion: Towards a Neural Network-based Spell Checker and Correction for Wolof*
Thierno Ibrahima Cissé and Fatiha Sadat

17:30–17:55 *Lateral Inversions, Word Form/Order, Unnamed Grammatical Entities and Ambiguities in the Constituency Parsing and Annotation of the Igala Syntax through the English Language*
Mahmud Mohammed Momoh

17:55–18:00 ***Closing***