



LREC 2026

**Resources for African Indigenous Languages (RAIL)
2026 @ LREC 2026**

Workshop Proceedings

Editors

**Muzi Matfunjwa, Mmasibidi Setaka,
Rooweither Mabuya, Menno van Zaanen**

12 May 2026

©ELRA Language Resources Association (ELRA), 2026

These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-74-6

Preface

The seventh workshop on Resources for African Indigenous Languages (RAIL) was held in Palau de Congressos de Palma, Palma de Mallorca, Spain, on 12 May 2026. It was co-located with the 15th edition of the Language Resources and Evaluation Conference (LREC 2026), which took place from 11 to 16 May 2026.

The RAIL workshop series is an interdisciplinary platform for researchers working on resources such as Natural Language Processing tools, Human Language Technologies, data collections, and annotations, specifically targeted towards African indigenous languages. It aims 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. Many African indigenous languages currently have no or very limited resources available, and they are often structurally different from well-resourced languages, requiring the development and use of specialised techniques. By bringing together researchers from different fields, such as computational linguistics, sociolinguistics, and language technology, to discuss the development of language resources for African indigenous languages, we hope to boost research in these fields.

The theme of this seventh RAIL workshop was “creating resources for less-resourced African languages”. It aimed to bring together researchers interested in showcasing their research and thereby boosting the field of African indigenous languages. This provided an overview of the current state-of-the-art and emphasised the availability of African indigenous language resources, including both data and tools. The workshop also facilitated information sharing among researchers interested in African indigenous languages and started discussions on improving the quality and availability of the resources.

For the seventh RAIL workshop, in total, 20 high-quality submissions were received. Out of these, 13 submissions were selected for presentation in the workshop. All submissions received at least three reviews using a double-blind review process. As organisers, we would like to thank the authors of all submissions for submitting high quality material, and the programme committee for their in-depth review of the submissions and for providing useful recommendations to the authors.

This RAIL workshop took place as a half-day workshop. Each presentation consisted of 15 minutes for papers (including time for discussion). This publication adheres to South Africa’s DHET’s 60% rule, as authors in the proceedings come from a wide range of institutions.

Programme Committee

- Ronny Mabokela, University of Johannesburg, South Africa
- Audrey Mbogho, United States International University - Africa, Kenya
- Alfred Kondoro, Hanyang University, South Korea
- Anuj Tiwari, ML Collective
- Fatiha Sadat, University of Quebec at Montreal, Canada
- Mmanape Hlungwane, University of the Witwatersrand, South Africa
- Pericles Adjovi, Carnegie Mellon University Africa, Rwanda

- Cindy McKellar North-West University, South Africa
- Sibonelo Dlamini, University of KwaZulu Natal, South Africa
- Hussein Suleman, University of Cape Town, South Africa
- Elias Malete, University of the Free State, South Africa
- Ayodele James Akinola, Michigan Technological University, United States of America
- Roald Eiselen, North-West University, South Africa
- Tanja Gaustad, North-West University, South Africa
- Marissa Griesel, South African Centre for Digital Language Resources, South Africa
- Deon du Plessis, South African Centre for Digital Language Resources, South Africa
- Papi Lemeko, Central University of Technology, Free State
- Michelle White, South African Centre for Digital Language Resources, South Africa
- Malefu Mahloane, University of the Free State, South Africa
- Muzi Matfunjwa, South African Centre for Digital Language Resources, South Africa
- Martin Puttkammer, North-West University, South Africa
- Mmasibidi Setaka, South African Centre for Digital Language Resources, South Africa
- Sibeko Johannes, Nelson Mandela University, South Africa
- Nomsa Skosana, South African Centre for Digital Language Resources, South Africa
- Benito Trollip, South African Centre for Digital Language Resources, South Africa
- Menno van Zaanen, South African Centre for Digital Language Resources, South Africa
- Ilana Wilken, Council for Scientific and Industrial Research, South Africa
- Friedel Wolff, South African Centre for Digital Language Resources, South Africa

Organizing Committee

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

Table of Contents

<i>A Morpho-Syntactically Annotated Corpus of Ògè Folk Narratives with a Focus on Nominal Structure</i>	
Priscilla Adenuga	1
<i>Extension of Linguistic Resources for South African Languages: Part-of-Speech Annotated Domain-Specific Data</i>	
Tanja Gaustad, Roald Eiselen and Cindy Arlene McKellar	7
<i>Mining Large Language Models for Low-Resource Language Data: Comparing Elicitation Strategies for Hausa and Fongbe</i>	
Pericles Adjovi, Prasenjit Mitra and Roald Eiselen	20
<i>Comparing Source Language Selection Strategies for Multi-Source Cross-Lingual Transfer to African Languages</i>	
Tewodros Kederalah Idris, Roald Eiselen and Prasenjit Mitra	31
<i>Benchmarking Text Embedding Models for South African Languages</i>	
Ockert de Villiers and Roald Eiselen	41
<i>Improving Amharic Information Retrieval with Translative and Multi-Agent Debate Retrieval Augmented Generation</i>	
Abel Alemu Jotie and Prasenjit Mitra	52
<i>Less can be More: Towards a Parameter-Efficient Fine-Tuning of Wav2Vec2 XLSR for Low-Resource Cape Verdean Creole ASR</i>	
Mateus Neves Andrade, Mouhamadou Lamine Ba, Idy Diop and Arlindo Oliveira da Veiga	62
<i>From Script to Semantics: Prompting Strategies for African NLI</i>	
Anuj Tiwari, Terry Oko-odion and Hannah Nwokocha	72
<i>HaYo: Repurposing DiaSafety Dataset for Dialogue Safety Evaluation in Hausa and Yoruba</i>	
Tunde Oluwaseyi Ajayi, Bolade Deborah Ashaolu, Falalu Ibrahim Lawan, Daud Olamide Abolade, Amina Imam Abubakar, Oluwatosin Ayomide Akinrinde, Murja Sani Gadanya, Omodolapo Dorcas Ashaolu, Abubakar Khalid Auwal, Adewumi Awujoola, Shamsuddeen Umaru Adamu, Israel Olawole Ashaolu, Mihael Arcan and Paul Buitelaar	84
<i>Reclaiming African Voices: Surveying Indigenous Writing Systems for Inclusive NLP</i>	
Mamady Traore, Ngoc Tan Le and Fatiha Sadat	96
<i>Getting Close to Cloze: Investigating Language Model and Human Cloze-test Performance in Afrikaans</i>	
Susan Lotz, Rik van Noord and Gertjan van Noord	107
<i>The Hundzula Retreat-Based Infrastructure Model for African Natural Language Processing</i>	
Johannes Sibeko, Seani Rananga, Neo N. Putini and Dan Masethe	121
<i>Open but Unvetted: The Ethics of African Language Data</i>	
Ernst A.P. van Gassen	128

Workshop Program

Tuesday, May 12, 2026

- 14:00–18:00** **Session RAIL: The Seventh Workshop on Resources for African Indigenous Languages 2026**
Chairs: Muzi Matfunjwa, Mmasibidi Setaka and Menno van Zaanen
- 14:00–14:15 *Opening*
Menno van Zaanen
- 14:15–14:30 *A Morpho-Syntactically Annotated Corpus of Ògè Folk Narratives with a Focus on Nominal Structure*
Priscilla Adenuga
- 14:30–14:45 *Extension of Linguistic Resources for South African Languages: Part-of-Speech Annotated Domain-Specific Data*
Tanja Gaustad, Roald Eiselen and Cindy Arlene McKellar
- 14:45–15:00 *Mining Large Language Models for Low-Resource Language Data: Comparing Elicitation Strategies for Hausa and Fongbe*
Pericles Adjovi, Prasenjit Mitra and Roald Eiselen
- 15:00–15:15 *Comparing Source Language Selection Strategies for Multi-Source Cross-Lingual Transfer to African Languages*
Tewodros Kederalah Idris, Prasenjit Mitra and Roald Eiselen
- 15:15–15:30 *Benchmarking text embedding models for South African languages*
Ockert de Villiers and Roald Eiselen
- 15:30–15:45 *Improving Amharic Information Retrieval with Translative and Multi-Agent Debate Retrieval Augmented Generation*
Abel Alemu Jotie and Prasenjit Mitra
- 15:45–16:00 *Less can be More: Towards a Parameter-Efficient Fine-Tuning of Wav2Vec2 XLSR for Low-Resource Cape Verdean Creole ASR*
Mateus Neves Andrade, Mouhamadou Lamine BA, Idy Diop and Arlindo Oliveira da Veiga
- 16:00–16:30** ***Afternoon Coffee Break***
- 16:30–16:45 *From Script to Semantics: Prompting Strategies for African NLI*
Anuj Tiwari, Terry Oko-odion and Hannah Nwokocha

Tuesday, May 12, 2026 (continued)

- 16:45–17:00 *HaYo: Repurposing DiaSafety Dataset for Dialogue Safety Evaluation in Hausa and Yoruba*
Tunde Oluwaseyi Ajayi, Bolade Deborah Ashaolu, Falalu Ibrahim Lawan, Daud Olamide Abolade, Amina Imam Abubakar, Oluwatosin Ayomide Akinrinde, Murja Sani Gadanya, Omodolapo Dorcas Ashaolu, Abubakar Khalid Auwal, Adewumi Awujoola, Shamsuddeen Umaru Adamu, Israel Olawole Ashaolu, Mihael Arcan and Paul Buitelaar
- 17:00–17:15 *Reclaiming African Voices: Surveying Indigenous Writing Systems for Inclusive NLP*
Mamady Traore, Ngoc Tan Le and Fatiha Sadat
- 17:15–17:30 *Getting Close to Cloze: Towards Readability Resources for Afrikaans*
Susan Lotz, Rik van Noord and Gertjan van Noord
- 17:30–17:45 *The Hundzula Retreat-Based Infrastructure Model for African Natural Language Processing*
Johannes Sibeko, Seani Rananga, Neo N. Putini and Dan Masethe
- 17:45–18:00 *Open but Incompatible: A License Compatibility Analysis of Corpora for Low-Resource African Languages*
Ernst A.P. van Gassen