AfricaNLP 2026

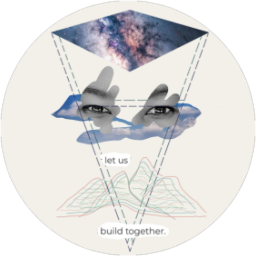
7th Workshop on African Natural Language Processing (AfricaNLP 2026)

Proceedings of the Workshop

March 28, 2026

The AfricaNLP organizers gratefully acknowledge the support from the following sponsors.

Niger-Congo Tier



Nilo-Saharan Tier

Microsoft

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-364-7

Introduction

We are pleased to present the proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), held in Rabat, Morocco on March 28, 2026, as part of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026).

This year's theme, "Multilingual Multimodal LLMs," reflects the growing importance of developing language models that can process and understand African languages across multiple modalities. The workshop continues our mission to advance natural language processing research for African languages, bringing together researchers, practitioners, and stakeholders from across the continent and beyond.

We received 56 submissions spanning topics such as machine translation, speech recognition, language modeling, multimodal AI, and culturally grounded NLP for African languages. Of these, 30 papers were accepted as archival contributions and appear in these proceedings, while 14 papers were presented as non-archival contributions at the workshop. Together, the accepted papers reflect research across diverse African languages and highlight the continued growth and vibrancy of the African NLP research community.

We thank all authors for their contributions, our program committee for their thorough reviews, and our sponsors, the Masakhane Research Foundation and Microsoft, for their generous support. We also thank the EACL 2026 organizers for hosting our workshop.

With gratitude,
The AfricaNLP 2026 Organizing Committee

Organizing Committee

General Chair

Shamsuddeen Hassan Muhammad, Imperial College London

Program Chairs

Constantine Lignos, Brandeis University
Shamsuddeen Hassan Muhammad, Imperial College London
David Ifeoluwa Adelani, McGill University and Mila

Publication Chairs

Everlyn Asiko Chimoto, University of Cape Town and Lelapa AI
Idris Abdulmumin, University of Pretoria
Clemencia Siro, Centrum Wiskunde & Informatica

Sponsorship Chair

Clemencia Siro, Centrum Wiskunde & Informatica

Mentoring Chair

Sang Yun Kwon, The University of British Columbia

Communications Chairs

Millicent Ochieng, Microsoft Research Lab - Africa
David Ifeoluwa Adelani, McGill University and Mila

Publicity Chair

Jessica Ojo, Mila & McGill University

Program Committee

Reviewers

Idris Abdulmumin, Henok Biadglign Ademtew, Ibrahim Said Ahmad, Felermino D. M. A. Ali, Lukman Jibril Aliyu, Anietie Andy, Berk Atıl, Abinew Ali Ayele

Tadesse Destaw Belay, Meriem Beloucif, Emmanuel Bolarinwa, Jan Buys, Happy Buzaaba

Rendi Chevi, Everlyn Asiko Chimoto, Chiamaka Ijeoma Chukwuneke

Daryna Dementieva, Emmanuel Dorley, Bonaventure F. P. Dossou

Chris Chinenye Emezue

David Guzmán, Tajuddeen Gwadabe

Khaldi Hadjer

Oana Ignat, Amina Abubakar Imam, Sukairaj Hafiz Imam, Ahmad Ibrahim Ismail, Sheriff Issaka

Adejumobi Monjolaoluwa Joshua

Salomon Kabongo Kabenamualu, Sulaiman Kagumire, Andrew Kiprop Kipkebut, Alfred Malengo Kondoro, Sujay S Kumar, Sang Yun Kwon

Falalu Ibrahim Lawan, Eric Le Ferrand, En-Shiun Annie Lee, Senyu Li, Weiran Lin

Rahmad Mahendra, Marek Masiak, Dunstan Matekenya, Francois Meyer, Kausar Yetunde Moshood, Anjishnu Mukherjee

Quang Phuoc Nguyen, Gebregziabihier Nigusie

Jacki O'Neill, Millicent Ochieng, Perez Ogayo, Odunayo Ogundepo, Jessica Ojo, Ugochi Okafor, Chibuzor Okocha, Akintunde Oladipo, Flora Oladipupo, Abigail Oppong, Salomey Osei, Abraham Toluwase Owodunni

Chester Palen-Michel, Ted Pedersen, Van-Thuy Phi

Samuel Rutunda

Elizabeth Salesky, Avinash Kumar Sharma, Kathleen Siminyu, Clemencia Siro, Rui Sousa-Silva, Nirmal Surange, Jonne Sälevä

Allahsera Auguste Tapo, Atnafu Lambebo Tonja

Kosei Uemura

Eric Peter Wairagala

Debela Desalegn Yadeta, Kweku Andoh Yamoah, Seid Muhie Yimam

Miaoran Zhang

Tolúlopé Ògúnrèmí

Invited Talk

Data-Efficient Language Modelling for Low-Resource Languages

Francois Meyer
University of Cape Town

Abstract: Progress in language modelling has been driven by scaling data and model size, but this approach is infeasible for most African languages. In this talk, I will present our work on developing data-efficient language models – architectures and training algorithms that improve performance on limited training data. I will present examples of how linguistically informed modelling, which targets and leverages the linguistic properties of specific languages, can improve sample efficiency. Finally, I will discuss the emerging intersection between low-resource NLP and developmentally inspired NLP, exploring how insights from human language learning can help us build more efficient models.

Bio: Francois Meyer is a Lecturer in the Computer Science Department at the University of Cape Town and co-investigator in the UCT NLP research group. His research is on data-efficient language modelling and linguistically informed subword tokenisation. He completed his PhD at the University of Cape Town and previously obtained a masters in AI at the University of Amsterdam.

Invited Talk
The Emergence of Multilingual Representations: Tracing Linguistic Capabilities During Language Model Pretraining

Barbara Plank
LMU Munich

Abstract: Multilingual large language models exhibit remarkable zero-shot and cross-lingual transfer capabilities. However, most analyses focus on fully trained models, leaving limited understanding of how and when different types of linguistic information emerge, interact, and align within multilingual representation spaces during training.

In this talk, I present a series of studies investigating the training dynamics of linguistic knowledge in language models, tracing how linguistic structure and cross-lingual alignment develop over time. Studying these dynamics requires access to intermediate checkpoints, which are only available to a limited extent. Nevertheless, analyzing emerging representations opens up new avenues for diagnosing and improving multilingual LLMs. Understanding how alignment forms during pretraining is particularly important for models intended to support underrepresented and low-resource languages, where effective transfer and shared representations are crucial for performance.

Bio: Barbara Plank is Full Professor and Chair for AI and Computational Linguistics at LMU Munich, Co-director of the Center for Information and Language Processing and Head of the MaiNLP (Munich AI and NLP) lab at LMU. Barbara Plank is an ELLIS Fellow (European Laboratory for Learning and Intelligent Systems) and regularly serves in international organizations and on scientific advisory committees.

Invited Talk
Towards Multimodal AI for African Languages and Cultures: Lessons from Afri-MCQA

Atnafu Lambebo Tonja
MBZUAI

Abstract: What will it take to develop multimodal AI that truly comprehends African languages and cultures? In this talk, I explore this question through lessons from Afri-MCQA, a benchmark covering 15 African languages across 12 countries. Our evaluation highlighted that current models face major challenges, such as 1) they are unable to process speech in African languages, 2) they lack cultural context, and 3) they struggle to generate culturally relevant responses, rather than merely recognizing them. I will share these insights and outline a pathway forward, emphasizing the importance of speech-first development, culturally grounded training, and cross-lingual knowledge transfer as critical steps in creating effective multimodal AI for Africa.

Bio: Atnafu Lambebo Tonja is a Postdoctoral Researcher at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) in the UAE, where he leads projects on culturally-diverse multilingual visual question answering and multimodal machine translation. He earned his PhD in Computer Science from Instituto Politécnico Nacional in Mexico City, focusing on neural machine translation for low-resource languages. His research focuses on advancing NLP for underrepresented languages, particularly African and Ethiopian languages, through the development of multilingual language models and sustainable data curation frameworks. His work on low-resource languages, especially for African and Ethiopian languages, has been published in top NLP venues including ACL, EMNLP, and NAACL.

Invited Talk

Beyond Parallel Data: Harnessing External Knowledge for Low-Resource MT

Felermino Ali
Porto University

Abstract: Translating from high-resource languages into Mozambican languages remains a pressing challenge in African NLP. The scarcity of parallel corpora, orthographic variation across dialects, and the frequent presence of loanwords and code-switching complicate the task of building robust translation systems. In this talk, I will share how we address these barriers through lexicon-guided neural machine translation. By integrating bilingual dictionaries and systematic loanword mappings directly into the training process, we move beyond data scarcity toward structured lexical enrichment. Our approach leverages dictionary entries and loanword mappings to construct sentence-specific glossaries, dynamically incorporated via input augmentation. On FLORES benchmarks, this method demonstrates clear gains: stronger lexical coverage, fewer inconsistencies, and translations that better capture contextual nuance. Beyond the technical improvements, this work points to a broader vision: advancing low-resource machine translation not only by scaling data but by intelligently bridging vocabulary gaps with structured linguistic knowledge. For Mozambican languages, this means opening pathways to more inclusive digital communication, empowering communities, and ensuring that the linguistic richness of African languages is represented in the global NLP landscape.

Bio: Felermino Ali is a researcher at MSR Africa and a PhD Candidate at the University of Porto in Portugal, focused on natural language processing (NLP) with a specialization in low-resource African languages. His work centers on building neural machine translation systems for low-resource languages and advancing methods to more effectively evaluate MT performance in low-resource settings.

Invited Talk

The knowns and unknowns of multilingual data augmentation

Julia Kreutzer
Cohere Labs

Abstract: In this talk I will present recipes for multilingual fine-tuning data augmentation that have been developed to overcome data scarcity in languages beyond English. We will then discuss what the limitations of these approaches are, and what directions are relevant for future research.

Bio: Julia Kreutzer is a Senior Research Scientist at Cohere Labs, where she focuses on research around multilingual large language models. She has a background in machine translation, with a PhD from Heidelberg University and prior work experience at Google Translate. She's passionate about advancing NLP technologies for underrepresented languages and has been part of multiple open science initiatives to work towards this goal collaboratively.

Table of Contents