Proposal

5th Workshop on African Natural Language Processing (AfricaNLP 2024)

2024 Theme: Adaptation of Generative AI Models for African Languages

Summary

Over 1 billion people live in Africa, and its residents speak more than 2,000 languages. But those languages are among the least represented in NLP research, and work on African languages is often sidelined at major venues. Over the past few years, a vibrant, collaborative community of researchers has formed around a sustained focus on NLP for the benefit of the African continent: national, regional, continental and even global collaborative efforts focused on African languages, African corpora, and tasks with importance in the African context. The AfricaNLP workshops have been a central venue in organizing, sustaining, and growing this focus, and we propose to continue this tradition with an AfricaNLP 2024 workshop in Vienna.

Starting in 2020, the AfricaNLP workshop has become a core event for the African NLP community and has drawn global attendance and interest. Many of the participants are active in the Masakhane grassroots NLP community, allowing the community to convene, showcase and share experiences with each other. Large scale collaborative works have been enabled by participants who joined from the AfricaNLP workshop such as MasakhaNER (61 authors), Quality assessment of Multilingual Datasets (51 authors), Corpora Building for Twi (25 authors), NLP for Ghanaian Languages (25 Authors). Many first-time authors, through the mentorship program, found collaborators and published their first paper. Those mentorship relationships built trust and coherence within the community that continues to this day. We aim to continue this.

In the contemporary AI landscape, generative AI has rapidly expanded with significant input and innovation from the global research community. This technology enables machines to generate novel content, showcases potential across a multitude of sectors. However, underrepresentation of African languages persists within this growth. Recognizing the urgency to address this gap has inspired the theme for the 2024 workshop: **Adaptation of Generative AI for African languages** which aspires to congregate experts, linguists, and AI enthusiasts to delve into solutions, collaborations, and strategies to amplify the presence of African languages in generative AI models.

Beyond the theme, we propose to:

- Continue to unite the African speech community with the African NLP community. Given the oral tradition of Africa, the support of speech tools is of utmost importance.
- Foster further relationships between the African linguistics and NLP communities. The linguistic structure of African languages is key in the evaluation and development of African models.
- Promote multidisciplinarity within the African NLP community, with the goal of creating a holistic participatory NLP community that will produce NLP research and technologies that value fairness, ethics, decolonial theory, and data sovereignty.
- Provide a platform for the groups involved with the various projects to meet, interact, share and forge closer collaboration.

- Provide a platform for junior researchers to present papers, solutions, and begin interacting with the wider NLP community.
- Present an opportunity for more experienced researchers to further publicize their work and inspire younger researchers through keynotes and invited talks.

Topics will include, but are not limited to:

- analyses of African languages by means of computational linguistics
- empirical studies reporting results from applying or adapting NLP developed for high-resource languages to African languages
- new model architectures tailored for African languages
- new corpora for African languages
- using NLP techniques on African datasets
- text generation for African languages
- methods addressing out-of-domain generalization for NLP tasks with training data in very limited domains
- transfer learning between African languages or from higher-resourced to lower-resourced languages
- challenges or solutions for resource gathering for African NLP tasks
- crowd-sourcing and open-sourcing software for African NLP
- multidisciplinary and participatory research in African NLP
- tutorials for African NLP for education or development purposes
- new resources for African NLP
- development of NLP systems for African languages for production
- socio-linguistic research for African languages and their decolonization
- ethical considerations for African NLP

We invite archival and non-archival submissions of long (up to 8 pages) and short (up to 4 pages) research papers, resource papers, position papers, tutorial/overview papers, and system design papers. We emphasize the reproducibility of empirical results and accessibility of resources in assessing the quality of submissions. We will grant unlimited space for the appendix and for author names (to encourage large-scale collaboration), as well as encourage short submissions.

In the interest of inspiring more scientific discourse in non-English languages, we invite researchers to include an abstract in another language. Reviewing will be double-blind. Non-archival submissions and previously published work can also be submitted if indicated upon submission.

Modality and Access

We are proposing a hybrid workshop, so participants will be distributed between on-site and virtual participation. Many members of the organizing committee have the experience of working in distributed and hybrid work environments across the globe, this brings onboard the experience of organizing hybrid gatherings that are inclusive across different time-zones and locations.

On-site participation: With the experience from the 2023 workshop in Kigali, there is no doubt that physical on-site participation presents endless opportunities. Participants are able to interact with a diverse community of researchers, share contacts, learn from not only what they have been working on but also what they are planning to work on and forge ways of collaborating. We are seeking sponsorships to help reduce the costs of African researchers to attend the workshop in-person as much as possible, and last year we were able to defray these costs for a significant number of presenters.

Virtual participation: To accommodate different time zones, we are planning morning and afternoon sessions with a common panel in the late afternoon to allow participation from different time zones at a reasonable local time for them. During the sessions, workshop moderators will closely monitor online mediums such as RocketChat and echo the questions to the speakers.

Post-workshop participation: The list of accepted papers will be made available on OpenReview, and the links to recorded talks and panels will be made available on the ICLR website. Both will have direct links to them hosted on our workshop website which will stay live after the workshop day. From our experience in 2021, video material recorded on more accessible platforms such as YouTube gain more cumulative long term visibility, so additionally we will encourage authors of accepted papers to upload their presentations on YouTube and provide a compiled playlist on the workshop page.

Tentative Schedule

Morning Sessions	
9:00 - 9:05	Opening Remarks
9:05 - 9:45	Morning Keynote: Graham Neubig
9:45 - 10:25	Invited talk: Ife Adebara
10:25 - 10:40	Mini break
10:40 - 11:10	2 Spotlight Talks from Authors (15 minutes each)
11:10 - 11:50	Invited talk: Akintunde Oladipo
11:50 - 12:50	Hybrid Poster Session (online and in-person)
12:50 - 2:00	Social + Lunch Break
Afternoon Sessions	
2:00 - 2:40	Evening Keynote: Pelonomi Moiloa
2:40 - 3:10	2 Spotlight Talks from Authors (15 minutes each)
3:10 - 3:25	Break
3:25 - 4:05	Invited Talk: Claytone Sikasote
4:05 - 4:45	Invited Talk: LLMs for Speech (Josh Meyer)

4:45 - 5:45	Panel: Adaptation of Generative AI models for African languages (Graham Neubig, Ife Adebara, Josh Meyer, Akintunde Oladipo)
5:45 - 6:00	Closing Remarks & Best paper awards

Invited Speakers/Panelists

The following invited speakers have tentatively confirmed their presence in the workshop. As we strive for diversity, we invited speakers from various backgrounds, with various research interests and ranging from graduate students to experienced academic and industrial researchers. These speakers will either present an individual talk or be part of a panel. The confirmed speakers and panelists are as follows:

Invited Speaker / Affiliation	**Tentative Topic**	**Confirmed**
Graham Neubig Carnegie Mellon University	LLMs and African Languages	Confirmed
Claytone Sikasote University of Zambia	Building a Multimodal Dataset for African languages	Confirmed
Joshua Meyer Coqui AI	Generative Speech XTTS	Confirmed
Ife Adebara University of British Columbia	Scaling multilingual LMs to 500 African languages	Confirmed
Pelonomi Moiloa Lelapa AI	NLP Tools for low-resource African languages	Confirmed
Akintunde Oladipo University of Waterloo	Building LMs for African languages using small quality data	Confirmed

Anticipated Audience Size

The AfricaNLP workshop is the annual venue backed by members of several AfricaNLP communities such as Masakhane, GhanaNLP, Data Science Africa and Deep Learning Indaba. This year, we expect around 200 participants between virtual and in-person participation. Technical requirements are similar to those required by other hybrid workshops; we need to be able to have onsite and remote speakers and panelists and support a hybrid poster session.

Diversity Commitment

Diversity, equity and inclusion are core values of the African NLP community. Similar to previous events we are committed to those values in every aspect of the workshop.

Organizers and Invited Speakers Diversity

- **Organizers:** 11/12 organizers are African, 1/4 organizers are female, 11/12 organizers are people of color. Organizers have a diverse mix of academic and industry experience, including policy experience.
- **Invited Speakers**: The majority of invited speakers are African, people of color, and are multi-disciplinary between academia, non-profit research labs and industrial labs.
- **Programme Committee**: We ensured diversity of our program committee, where almost ½ are women, over ½ are African and over ½ are people of color.

Participants and Submissions Diversity

- **To promote diversity among participants**, we will use workshop funding to sponsor conference fees for participants, offer a non-archival submission option so that submitting to the workshop does not prevent further publication, and welcome a wide range of types of papers and encourage multidisciplinary work.
- **Mentorship Program:** In order to encourage participation and collaboration between different skills and seniority levels. We are going to organize a peer mentorship program. The Peer mentorship program philosophy is aligned with our belief that everyone has valuable knowledge that is worth listening to. Our Peer Mentorship program for AfricaNLP 2021 was a large success with 64 participants. Equipped with feedback from our previous mentorship program, we plan to run a further improved version for this workshop.
- **Accessibility:** We plan to use sponsorship money to provide Internet bandwidth grants to cover extra costs of internet access and reduce the barrier to participation in the workshop.

Sponsorship

The AfricaNLP workshop follows a no one left behind spirit when it comes to allowing African students and researchers to attend the workshop. In the last two years the AfricaNLP workshop has been successful to raise funds to sponsor all applicants seeking financial assistance to attend the workshop.

This year, due to the hybrid nature of ICLR2024 allocated funds will go into sponsoring student registrations to ICLR, travel and accommodation funds. We are targeting sponsorship of approximately $60,000 amount, to cover 30-40 applicants between online and offline, we are leveraging our network to raise funds from the following sources:

- Cohere AI (TBC)
- DeepMind (previous year sponsor)
- GIZ Foundation (previous year sponsor)
- Google (previous year sponsor)
- Lacuna Fund (TBC)
- Meta (TBC)
- Naver Labs Europe (previous year sponsor)
- Lelapa (previous year sponsor)
- JP Morgan (TBC)
- EAMT (TBC)

Previous Related Workshops

1st AfricaNLP, ICLR 2020

- Number of submitted papers: 47
- Number of accepted papers: 27 (regular, no shared task)
- Number of invited speakers: 6
- Number of attendants: 1,148 unique page views of the ICLR workshop page and 743 unique viewers on the rocket chat system. Around 200 participants joined and actively participated.

2nd AfricaNLP, EACL 2021

- Number of submitted papers: 42 (largest number of submissions at EACL)
- Number of accepted papers: 40 (regular, no shared tasks)
- Number of invited speakers: 6
- Number of attendants: Maximum on zoom was 59 participants, with 2000 collective views on YouTube. Around 200 participants joined throughout the day
- Number of sponsored registrations: 60 (70% students)

3rd AfricaNLP, ICLR 2022

- Number of submitted papers: 21
- Number of accepted papers: 19
- Number of invited speakers: 8
- Number of sponsored registrations: 55 (36% female)

4th AfricaNLP, ICLR 2023

- Number of submitted papers: 48
- Number of accepted papers: 32
- Number of invited speakers: 7
- Number of sponsored registrations: 38 (71% students)
- Number of sponsored travel and accommodation: 24

Organizers and Biographies

David Adelani is a Research Fellow at the Computer Science department of University College London, United Kingdom. His research focuses on natural language processing for African languages. He obtained his PhD student from the department of Computer Science at Saarland University, Germany. He was an organizer for AfricaNLP 2022 and 2023.

Bonaventure F. P. Dossou holds a Bachelor of Science in Mathematics with honors and a Master of Science with honors in Computer Science and Data Engineering, with a Bioinformatics minor, from Jacobs University. Bonaventure is currently a Research Scientist at Lelapa AI, a CS PhD student at McGill University working on NLP & Health, and a co-chair of the WideningNLP (WiNLP) workshop. Previously, Bonaventure was a visiting student researcher at the Mila Quebec AI Institute working on drug discovery projects under the supervision of Professor Yoshua Bengio and Dr Dianbo Liu. Bonaventure previously also worked as a Researcher at Google Research in the language team and at Roche Canada, in the Pharmaceutical division. Bonaventure was a co-organizer of AfricaNLP 2021, 2022 and 2023. Read more about him and his work at https://bonaventuredossou.github.io/

Hady Elsahar (he/him) is a research scientist in Meta AI working on multilingual and multimodal machine translation. Prior to this he was a researcher in NAVERLABS Europe, and originally hails from Cairo, Egypt. He is particularly interested in Language Generation under constrained and controlled conditions.

He holds a PhD degree from L'Université de Lyon. He organized a workshop on Energy-Based Models in ICLR2021 and the AfricaNLP workshop at EACL2021, ICLR2022, and ICLR2023.

Constantine Lignos (he/him) is an assistant professor at Brandeis University (USA) where he directs the Broadening Linguistic Technologies Lab. His research focuses on natural language processing for less-resourced, indigenous, and historically marginalized languages. He is an action editor for ACL Rolling Review, associate editor of Language Resources and Evaluation, organized the 2021 Automatic Detection of Borrowings (ADoBo) workshop and shared task as part of the Iberian Languages Evaluation Forum, created the Dataset Creation for Lower-Resourced Languages workshop which was first held at LREC 2022, and was an organizer for AfricaNLP 2022 and 2023.

Atnafu Lambebo Tonja is research scholar at University of Colorado Colorado Springs, USA and a Ph.D. student at the Centro de Investigación en Computación, Instituto Politécnico Nacional (IPN), Mexico. His research focuses on machine translation for low-resource languages, large language models(LLMs) and low resource languages, Text Processing, Named-Entity Recognition, Sentiment Analysis, Offensive Language Identification, Author Profiling, Clinical NLP, and Code-Mixed Texts.

Salomey Osei is a research assistant at the Faculty of Engineering in DeustoTech, University of Deusto. Her research majorly focuses on data-driven decision making in transportation through Automated Machine Learning. She is also a researcher at Masakhane and the research lead of unsupervised methods for Ghana NLP. She is passionate about mentoring students, especially females in STEM.

Happy Buzaaba is a postdoctoral research associate at Princeton University. His research focuses on developing technologies for low resource African languages. Previously, he was a postdoctoral researcher at the Riken center for AIP where he spent time investigating the use of natural-gradient Bayesian methods to improve uncertainty estimation in PLMs. He received his PhD in computer science at the University of Tsukuba in Japan.

Aremu Anuoluwapo is a graduate student of the University of Lagos where he bagged a Bachelor's of Arts in Linguistics and African Studies. His research focuses on methods and techniques for linguistic data curation on low resourced languages for multi-domain NLP tasks. He previously worked at Translators Without Borders as a Terminologist.

Clemencia Siro is a PhD student at the University of Amsterdam where her research focuses on the evaluation of dialogue systems from user interactions. Apart from that she also has interest in research involving low resource languages where she has contributed to various projects conducted by Masakhane as a member.

Shamsuddeen Muhamamd is a faculty member at Bayero University, Kano. His research interests focus on text classification tasks such as sentiment analysis, hate speech and fake news detection. He is a founder of HausaNLP and active member of MasaKhane NLP. He was the organizer of AfricaNLP 2022 and 2023.

Tajuddeen Rabiu Gwadabe is a project manager with Masakhane Research Foundation. He obtained his PhD in Computer Science and Technology from the University of Chinese Academy of Sciences, China. His research interests focus on low resource NLP for African languages.

Kayode Olaleye is a Postdoctoral Fellow at the University of Pretoria in South Africa, affiliated with the Data Science for Social Impact (DSFSI) Lab led by Prof. Vukosi Marivate. His research explores approaches for processing and generating code-switched and code-mixed data in African languages. He obtained a PhD at the University of Stellenbosch South Africa.

Perez Ogayo is a software engineer at Oracle. She holds a Masters degree in Language Technologies from the Languages Technologies Institute at Carnegie Mellon University and a Bachelor's Degree (Hons) in Computer Science from African Leadership University. Her research focuses on machine translation and speech recognition for low resource languages

Israel Abebe He is a PhD student and research assistant at Saarland University with more than four years of Machine learning industry experience. His research interests include high-performance computing in AI, multimodal learning, natural language processing, and the application of deep learning.

Programme Committee

Based on previous years figures, we estimate receiving around 30-50 submitted papers this year. To provide a high-quality feedback process we target 3 reviews per submission and a light load of 3 submissions per reviewer. We carefully select our PC members to ensure diversity in terms of institutions, spoken languages and overall knowledge and experience of low-resourced NLP research. Below is a list of PC members from the past ICLR 2022 workshop that have agreed and confirmed to be part of our AfricaNLP workshop so far. We expect many to help review submissions for this workshop's edition as well, while looking forward to increasing the list to a set of 50 members.

PC Members

Ignatius Ezeani, Lancaster University, UK	Tosin Adewumi, Luleå University of Technology
Adewale Akinfaderin, Florida State University	Kosisochukwu Madukwe, Victoria University of Wellington
Orevaoghene Ahia, InstaDeep Ltd, Nigeria	Idris Abdulmumin, Ahmadu Bello University, Zaria - Nigeria
Layla El Asri, Borealis AI	Olamilekan Wahab, Independent Researcher
Wisdom d'Almeida, MILA	Surafel M. Lakew, Amazon
Ernie Chang, Saarland University	Wilhelmiena Nekoto, Independent Researcher, Namibia
Mathias Müller, University of Zurich	Amelia Tayor, University of Malawi
Alicia Tsai, Amazon	Rosanne Liu, ML Collective
Isaac Caswell, Google Research	Spandana Gella, Amazon
Musie Meressa, Sapienza University of Rome	Natalie Schluter, Google Brain & IT University Copenhagen
Nouha Dziri, University of Alberta	Michael A Hedderich, Saarland University
Colin Leong, University of Dayton	Chris Emezue, Technical University of Munich
Kathleen Siminyu, Mozilla Foundation	Salomon Kabongo, Leibniz Universität Hannover
Marius Mosbach, Saarland University	Yacine Jernite, HuggingFace