EACL 2026 - Birds-of-a-Feather Session

OpenData Collection Playbook and Annotation for African Languages

A Practical, Community-Driven Guide for Building High-Quality Datasets for African Languages

Shamsuddeen Hassan Muhammad (Imperial College London)
Seid Muhie Yimam (University of Hamburg)

26/03/2026

State of NLP in African Languages

Challenges in Data Curation in African Languages

Annotation Playbook and Tool

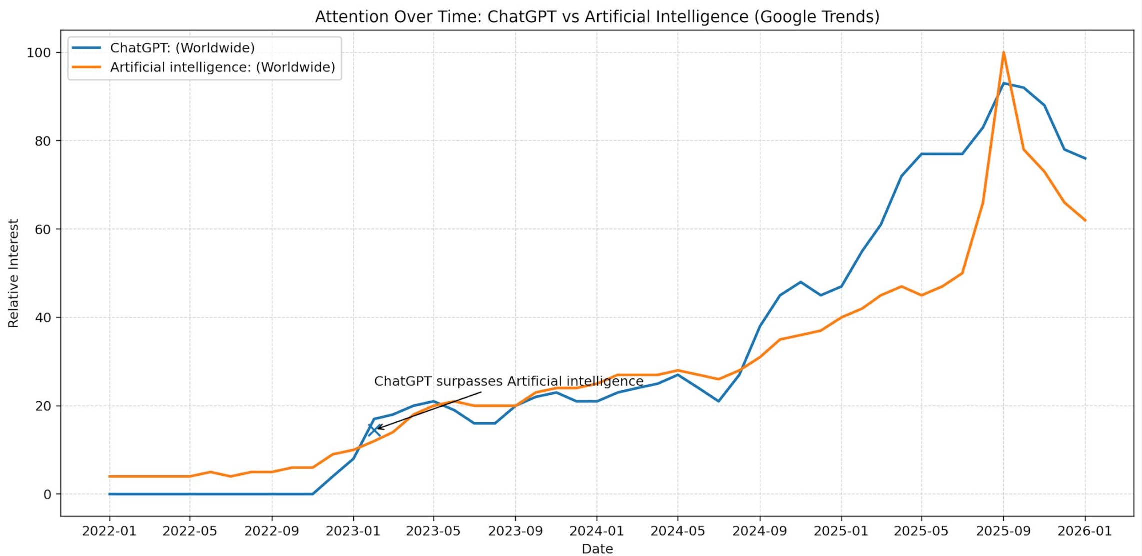
State of NLP in African Languages

AI Research in the Global South

Language Technology in Low-Resource Languages

LLMs are redefining AI research and decision making

Africa’s > 2,000 languages → **minimal inclusion**



Significant improvements in language models have primarily benefited English texts

98% of African languages are still unsupported by today’s LLMs—leaving most of Africa outside the AI revolution.

‘*The limits of my language mean the limits of my world.’….* Wittgenstein

The State of Large Language Models for African Languages: Progress and Challenges. (Hussen, K.Y., et.al 2025) – Best Paper Deep Learning Indaba 2025

State of NLP in Africa

Community Driven Participatory research

Compared to other regions where the AI ecosystem is shaped by Universities, big corporations or strong policies and regulation frameworks:

Africa’s AI ecosystem is dominated by grassroots movements, such as ‘Deep Learning Indaba’ and ‘Data Science Africa’.

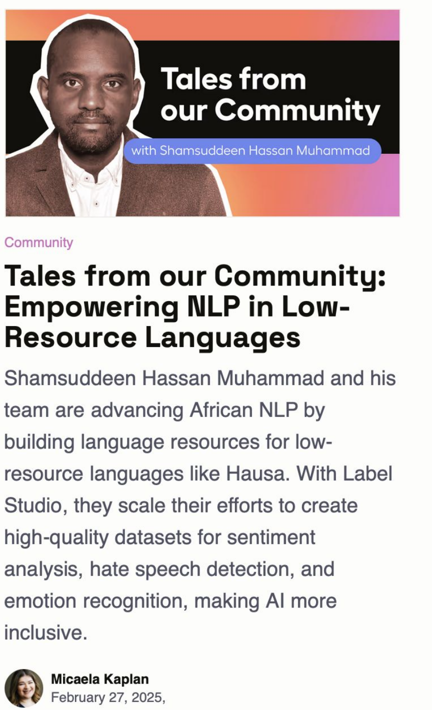
AI and the Future of Work in Africa White Paper O'Neill, J., et al. (2024, June).

Blossoming of Local Communities

Addressing the challenges of NLP for African languages through participatory approach

Masakhane

A grassroots NLP community for Africa, by Africans



Community

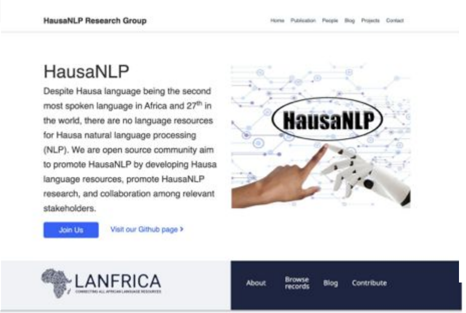
Tales from our Community: Empowering NLP in Low-Resource Languages

Shamsuddeen Hassan Muhammad and his team are advancing African NLP by building language resources for low-resource languages like Hausa. With Label Studio, they scale their efforts to create high-quality datasets for sentiment analysis, hate speech detection, and emotion recognition, making AI more inclusive.

Micaela Kaplan
February 27, 2025,

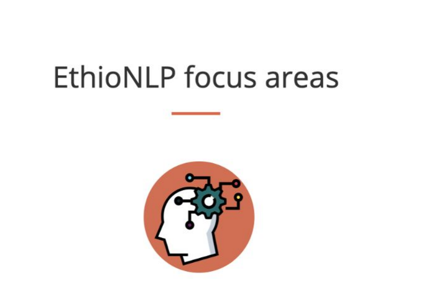
Ghana Natural Language Processing

Processing (NLP) Of Ghanaian Languages & It's Applications To Local Problems



HausaNLP

Despite Hausa language being the second most spoken language in Africa and 27th in the world, there are no language resources for Hausa natural language processing (NLP). We are open source community aim to promote HausaNLP by developing Hausa language resources, promote HausaNLP research, and collaboration among relevant stakeholders.

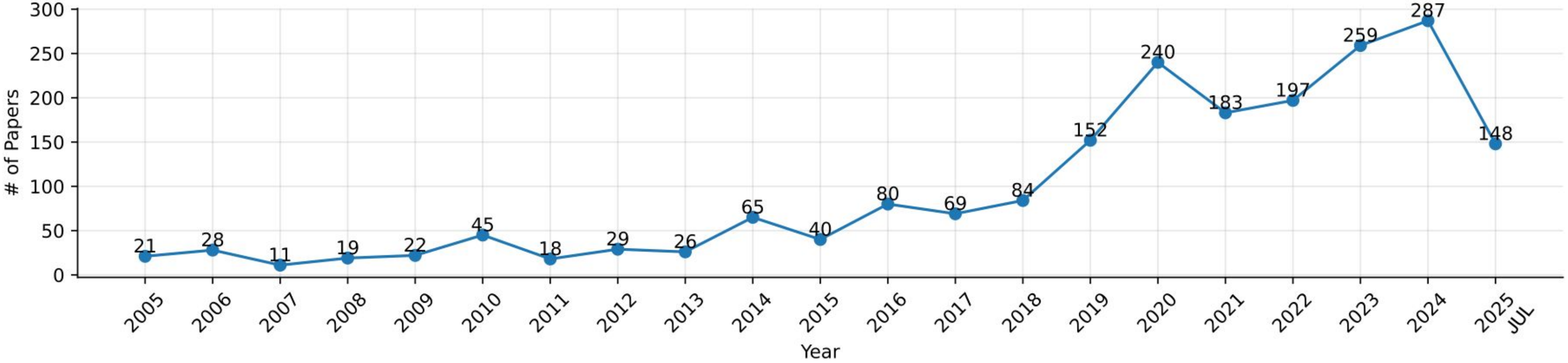


EthioNLP focus areas

The Rise of African NLP (2005-2025)

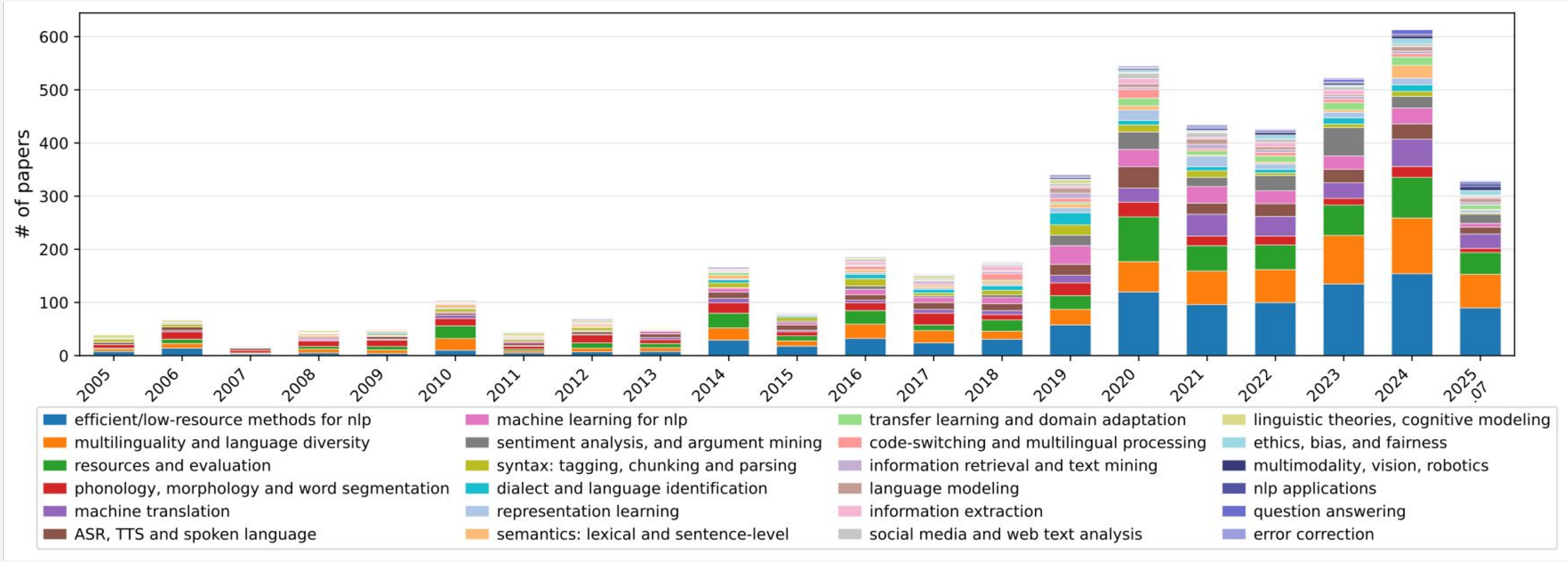
Two-Decade Analysis of African NLP

Papers from 1.9K AfricaNLP papers spanning 2005–2025.



Sharp increase from 2019 onward, reflecting the rise of community-driven initiatives (e.g., *Masakhane*, and *AfricaNLP Workshops*).

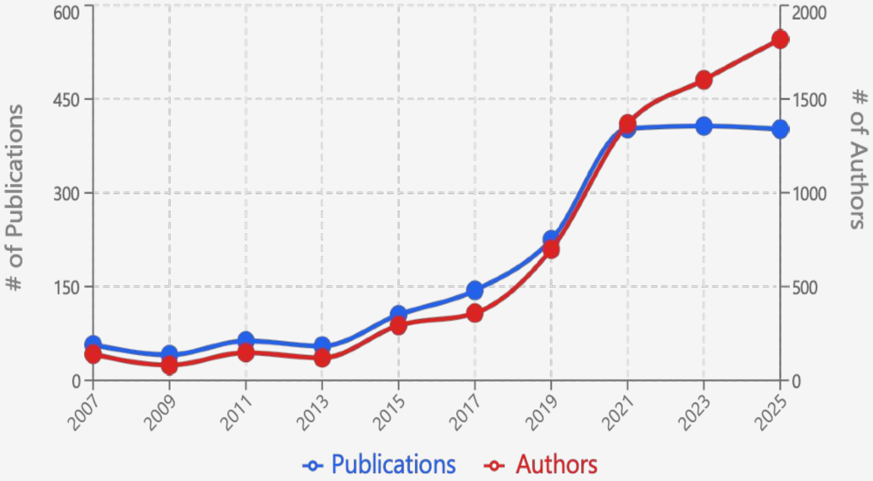
Evolving Research Themes Across Two Decades (2005–2025)



The Paradox of Progress: AfricaNLP in the LLM Era

LLMs still under-represent African **languages** and **cultures**

Progress at the periphery, **exclusion at the core** of global AI (only 42 supported African languages).



The State of Large Language Models for African Languages: Progress and Challenges. (Hussen, K.Y., et.al 2025) – Best Paper Deep Learning Indaba 2025

Pretraining Data — Biased and Noisy

- Most "African" datasets are **machine-translated**, **noisy**, or ***mislabeled***
- African languages become **statistically invisible** during LLM pretraining.

At least 15 corpora have no usable text, and a significant fraction contains less than 50% sentences of acceptable quality

Kreutzer et al. (2022). ***Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets.*** **TACL, 10: 50–72.**

Evaluation Data — Inaccurate and Unreliable

- FLORES is widely used to evaluate MT for African languages.
- We audited — *Hausa, Xitsonga, Sesotho, and isiZulu*, we found translation errors, inconsistencies, and of machine-generated text.
- This means model performance has been misreported for years.

If the evaluation set is broken, then claims of model performance are also broken.

***Correcting FLORES** evaluation dataset for four African languages — Abdulmumin et al. (2024)*

Challenges (1/2)

- Access to native speakers and domain experts.
- Orthographic and dialectal variation.
- Code-switching and multilingualism.
- Cultural and contextual sensitivity
- Translation Quality

Challenges (1/2)

- Infrastructure and funding constraints.
- Ethical considerations around data collection
- Access to crowdsourcing
- Digital divide- Offline first
- Mobile-friendly annotation tool

Solving Africa's NLP challenges requires community at every stage—sourcing, annotation, and quality assurance.

The people who speak the languages must drive how their data is collected, labeled, and validated.

AFRICAN LANGUAGES SHAPING THE FUTURE OF AI

MASAKHANE
AFRICAN LANGUAGES HUB

WE ARE THE HUB

African led AI. Built for impact.

MASAKHANE
AFRICAN LANGUAGES HUB

Anchored by the Masakhane Research Foundation (MRF), the Hub addresses the critical underrepresentation of African languages in AI development and the barriers to equitable access to AI-driven innovations.

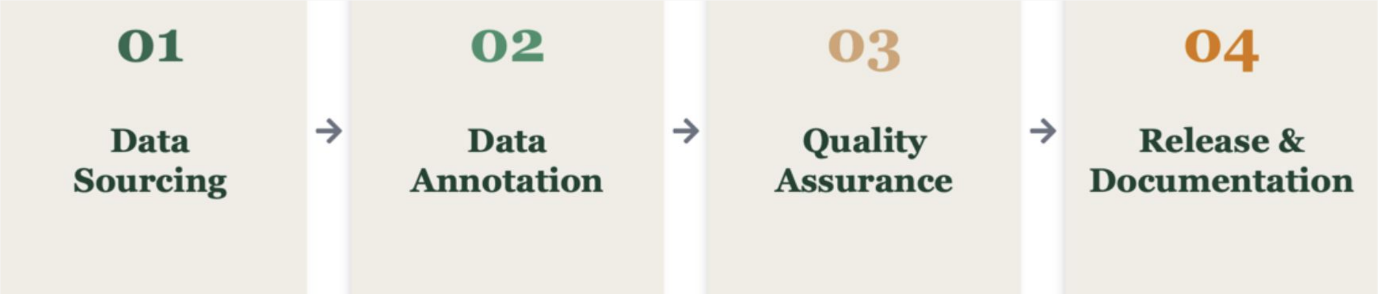
Open Data Collection Playbook & Collection Platform

Masakhane Playbook

A participatory guide that puts communities in the driver's seat across the entire data pipeline: sourcing, annotation, and quality assurance."

Masakhane Playbook

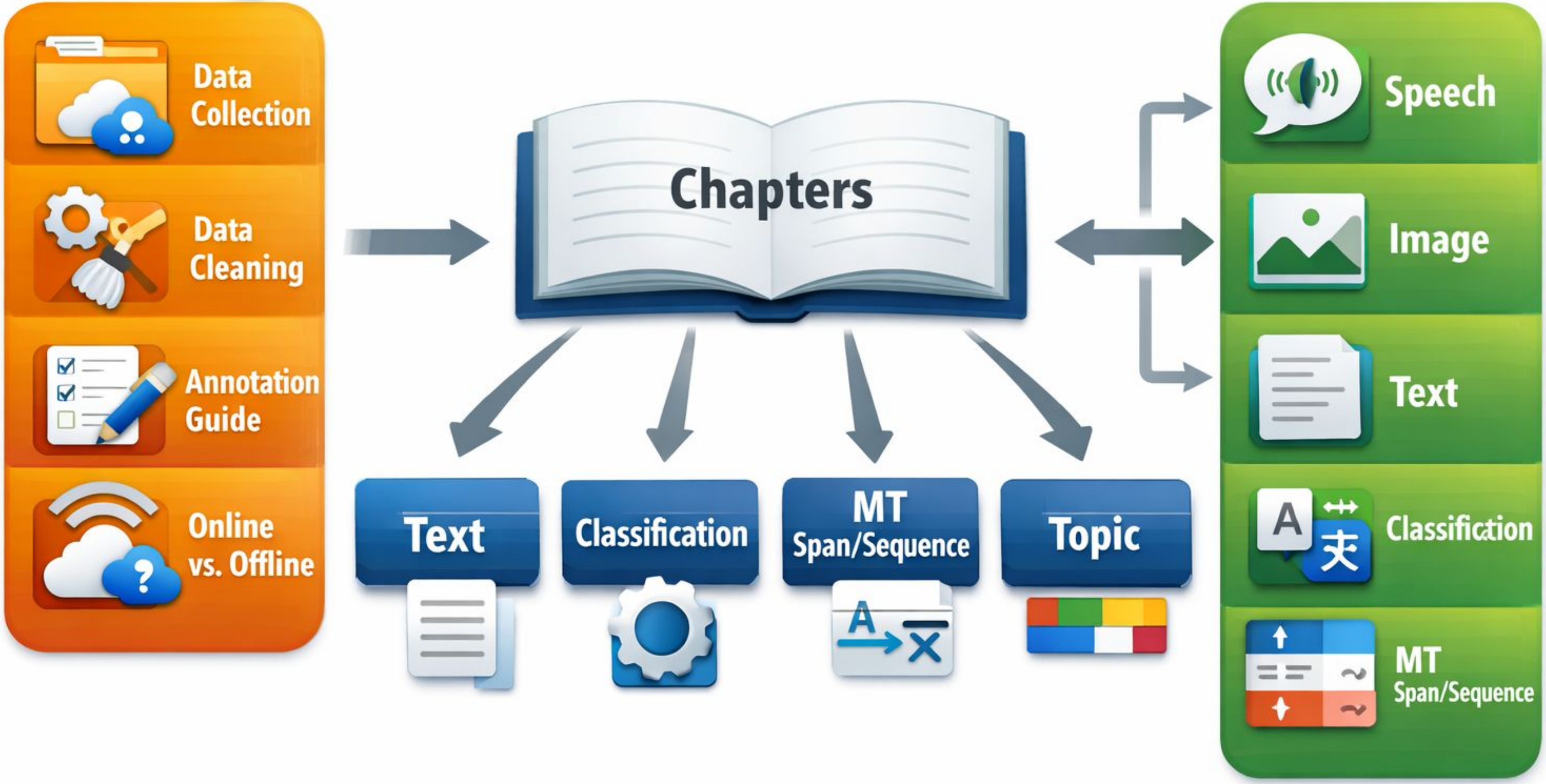
Community involvement at every stage



01 Data Sourcing → 02 Data Annotation → 03 Quality Assurance → 04 Release & Documentation

Each stage is led by native language speakers and local community members.

Annotation Playbook Development



Data Collection

Data Cleaning

Annotation Guide

Online vs. Offline

Chapters

Text

Classification

MT Span/Sequence

Topic

Speech

Image

Text

Classification

MT Span/Sequence

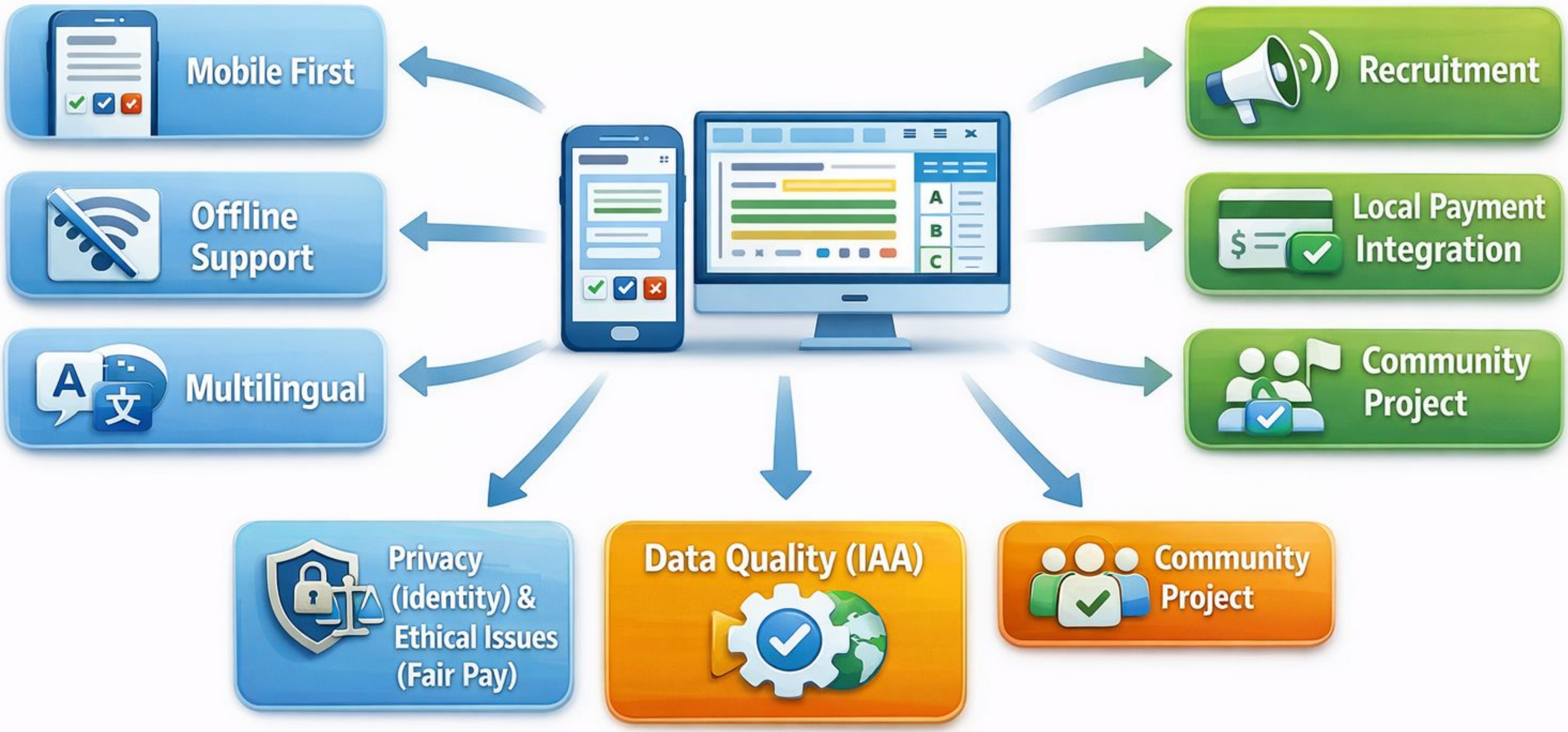
What's Inside the Playbook?

1. How to identify and select data sources
2. Designing culturally grounded annotation guidelines
3. Recruiting and training community annotators
4. Managing annotation workflows and tools
5. Quality assurance protocols and IAA metrics
6. Ethical considerations, consent, and data licensing
7. Documentation and datasheet templates
8. Lessons learned from Masakhane projects

Masakhane Annotation Tool

A tool with offline-support, mobile-first, crowdsourcing and community engagement for multimodal data annotation!

Annotation Tool Development



Mobile First

Offline Support

Multilingual

Recruitment

Local Payment Integration

Community Project

Privacy (Identity) & Ethical Issues (Fair Pay)

Data Quality (IAA)

Community Project

What's Inside the Tool?

1. Multilingual annotation support
2. Integrated with the Playbook
3. Local Payment and community project support
4. Local and crowdsourcing support
5. Predefined templates for multimodal annotation tasks
6. Integrated quality assurance
7. User and project management
8. Adjudication, error correction, and diagnostic support

Let's Discuss!

- What data collection challenges have you faced for your language?
- How can we better involve communities who are not yet digitally connected?
- What tools or platforms work best for collaborative annotation?
- How do we ensure data sovereignty and ethical use of community data?
- What should the Playbook include that we haven't thought of?

Call for Chapters

Coming soon

Thanks

s.muhammad@imperial.ac.uk,
seid.muhie.yimam@uni-hamburg.de