<AI & Equality> African Toolbox | Case study

Empowering African Languages through NLP: KenCorpus Project

Watch the video

This case study is part of the **African <AI & Equality> Toolbox**, which builds upon the methodology of the global <AI & Equality> Human Rights Toolbox—an initiative of Women At The Table in collaboration with the United Nations Office of the High Commissioner for Human Rights (OHCHR). The African Toolbox is a collaboration between the <AI & Equality> initiative and the African Centre for Technology Studies (ACTS). To learn more visit aiequalitytoolbox.com

The Silent Crisis

In 1992, linguist Parcel Hill made a chilling prediction: by the year 2100, the world's linguistic diversity would largely disappear, with most languages becoming obsolete as people gravitated toward English and other dominant tongues. What seemed like a distant academic concern has become a pressing reality, particularly visible in the digital realm where artificial intelligence is reshaping how we communicate, learn, and preserve knowledge.

Dr. Lilian Wanzare, a researcher at Maseno University in Kenya, witnessed this crisis firsthand. Despite Africa being home to over 2,000 languages and Kenya alone hosting more than 50 distinct languages across Nilotic, Bantu, and Cushitic families, the digital world remained largely silent in these tongues. The statistics were stark and sobering: while 77% of natural language processing tools supported English and other "global north" languages, only 6% supported low-resource languages. Yet this 6% represented the linguistic reality of 3 billion people – nearly half the world's population.

The irony was profound. The very technologies designed to bridge communication gaps were actually widening them, creating a digital apartheid where the world's linguistic diversity was being systematically erased, one algorithm at a time.

The Awakening: Understanding the roots of exclusion

Dr. Wanzare and her team began to understand why African languages were disappearing from the digital landscape. The problem wasn't just technological – it was fundamentally about data. Every AI system, every translation tool, every speech recognition service needed vast amounts of digital text and audio to learn from. But African languages existed primarily in the oral tradition, in the stories told by elders, in the daily conversations of rural communities, in the songs sung during harvest seasons.

The educated African population, ironically, had become part of the problem. Colonial legacies meant that English, French, or Portuguese served as official languages in most African countries. Educated Africans often couldn't write fluently in their native tongues. They didn't blog in Dholuo, didn't tweet in Kalenjin, didn't write academic papers in Kikuyu. The natural generators of digital content – the educated, urban, connected populations – were creating content in colonial languages, not indigenous ones.

This created a vicious cycle: no digital content meant no data, no data meant no AI tools, no AI tools meant these languages remained excluded from the digital future, making them appear less valuable and further accelerating their decline.

The decision: Community at the center

Faced with this reality, Dr. Wanzare made a radical decision. Instead of accepting that African languages were "low-resource," she would mobilize entire communities to become active participants in creating the digital future of their own languages. This wasn't going to be a top-down technological solution imposed by researchers in university labs. It would be a grassroots movement, with communities as partners, not subjects.

The KenCorpus project was born from this philosophy. Over what would become a five-year journey, Dr. Wanzare and her team would need to go beyond traditional academic research. They would need to become community organizers, cultural ambassadors, and bridge-builders between oral traditions and digital futures.

Building the Foundation: Stories become data

The first phase of KenCorpus was deceptively simple yet profoundly challenging. The team began traveling to rural communities, sitting with elders, talking with families, and asking them to do something that had never been systematically done before: tell their traditional stories and have them recorded and transcribed into digital form.

This wasn't just data collection – it was cultural preservation in action. Each story captured wasn't just text for training algorithms; it was a piece of living heritage being transferred from the oral realm into the digital one. Grandmothers who had never seen a computer became, unknowingly, the first contributors to Kenya's digital language infrastructure.

The team faced immediate challenges. How do you capture the tonal variations of different languages? How do you account for the fact that the same language might be spoken differently in coastal areas versus highland regions? How do you respect cultural protocols around storytelling while creating standardized digital formats?

The solution emerged through deep community engagement. Local chiefs provided credibility and mobilization support. Primary school teachers helped with transcription and verification. Church leaders opened their congregations as venues for recording sessions. The project became a community affair, with everyone understanding they were participating in preserving their linguistic heritage for future generations.

Expanding the Vision:
From Stories to Systems

As the initial collections grew, Dr. Wanzare and her team began to understand what communities actually needed from these digital language tools. Three clear priorities emerged from their conversations with language speakers:

Translation became the first critical need. People wanted to communicate across language barriers – not just from English to local languages, but between local languages themselves. A Dholuo speaker needed to communicate with a Kalenjin speaker. Government information in English needed to be accessible in local languages. This meant creating parallel corpora – the same sentences translated across multiple languages and carefully aligned. The team made a strategic decision to use Kiswahili as an anchor language. Rather than making English the central hub, they recognized Kiswahili as a widely understood African language that could serve as a bridge between different Kenyan languages. This wasn't just technically sound; it was culturally appropriate and politically significant.

Speech recognition emerged as the second priority. Communities envisioned a future where they could speak to their phones in their native languages, where meetings could be automatically transcribed in Kikuyu, where oral traditions could be instantly converted to written form. This required building massive speech corpora – targeting 3,000 hours of recorded speech across five languages.

Language infrastructure became the third need. Behind every grammar checker, every spell-check system, every language learning app are fundamental NLP tasks like part-of-speech tagging. These might seem mundane to technologists, but they're the backbone of language technology. Without them, no advanced language tools can function properly.

The Technical Challenge:
Building AI for the unconnected

Creating AI systems for languages with no existing digital infrastructure required innovative approaches. Traditional machine learning assumes you can scrape vast amounts of text from the internet. For Kenyan languages, the internet was essentially empty. Dr. Wanzare's team had to become experts not just in AI, but in linguistics, anthropology, and community organizing. They needed to understand how code-switching worked – the way speakers naturally mixed their native languages with Kiswahili or English within single conversations. They needed to capture not just formal language, but the way people actually spoke in their daily lives.

The technical architecture they developed was multilingual by design, with Kiswahili serving as the anchor. This meant a Dholuo speaker could ask a question to an AI system like ChatGPT by speaking in Dholuo. The system would translate to English, process the query, generate a response in English, then translate back to Dholuo. For the first time, global AI systems could become accessible to speakers of indigenous African languages.

Confronting Deeper Questions: Who owns language?

As the project grew, deeper questions emerged. Who owns the data being collected? What happens when global tech companies want to use these datasets? How do you ensure that communities benefit from the AI systems built on their linguistic contributions?

Working with Mozilla Common Voice, the team began developing community-based licensing frameworks. These weren't just legal documents; they were attempts to encode indigenous concepts of collective ownership and community sovereignty into the digital age. Traditional open-source licenses assumed individual ownership and global access. But languages belong to communities, not individuals. The stories being recorded were part of cultural heritage, not just data points.

This innovation had implications far beyond Kenya. Indigenous communities worldwide were grappling with similar questions as AI systems began to incorporate their languages and cultural knowledge. The KenCorpus approach offered a model for how communities could maintain sovereignty over their linguistic heritage while still participating in global technological development.

The Human Network: Beyond technology

Five years into the project, it became clear that KenCorpus's greatest innovation wasn't technological – it was social. The project had created a network of thousands of people across Kenya who understood themselves as active participants in shaping their languages' digital future. Local research assistants were working in Somaliland, in rural Kalenjin communities, in urban Nairobi neighborhoods. University linguists were collaborating with primary school teachers. County governments were providing resources. Media houses were contributing their archives. Traditional chiefs were endorsing the work in community meetings.

This network solved the fundamental challenge of scaling data collection for low-resource languages. You can't build linguistic infrastructure without massive community participation. But you can't get community participation without trust, cultural sensitivity, and genuine partnership.

Dr. Wanzare learned that incentivization was about more than payment. People participated because they understood the long-term vision: their children would grow up in a world where their native languages weren't barriers to accessing education, healthcare, government services, or economic opportunities. Their languages wouldn't just survive; they would thrive in the digital age.

Scaling the vision: Small models, big impact

The project also pioneered a different approach to AI development. Instead of pursuing ever-larger language models, the KenCorpus team focused on small, domain-specific models tailored to community needs. These models could run on modest hardware, could be customized for specific dialects, and were more accurate for their intended use cases than generic large models.

This approach challenged the prevailing Silicon Valley wisdom that bigger is always better. For communities with limited technological infrastructure, smaller, specialized models were actually more appropriate and more empowering.

The team also established critical research questions: What's the minimum viable amount of data needed to create functional language models? How do you balance model accuracy with cultural appropriateness? How do you ensure AI systems respect the way languages are actually spoken in communities rather than imposing academic standards?

The ripple effect: Beyond Kenya

By its fifth year, KenCorpus had become more than a Kenyan project. Researchers from across Africa were adapting its methodologies. International organizations were funding similar initiatives. The approach was being discussed in academic conferences, policy forums, and community meetings across the Global South.

The project demonstrated that technological marginalization wasn't inevitable. Communities could become active agents in their own digital empowerment. Languages that had been written off as “low-resource” could become fully functional in the digital ecosystem through systematic community engagement and culturally appropriate technical approaches.

More importantly, KenCorpus showed that AI development could be genuinely participatory. Instead of technology being developed for communities, it could be developed with communities as equal partners and primary beneficiaries.

Lessons from the Field: What KenCorpus taught us

After five years of intensive work, several critical insights emerged:

- **Community engagement must be continuous and authentic.** You can't extract linguistic data and disappear. Building language technology requires ongoing relationships and genuine partnership.
- **Cultural context is as important as technical accuracy.** AI systems that don't respect how languages are actually used in communities will fail, no matter how technically sophisticated they are.
- **Incentivization is complex.** People contribute not just for immediate payment but for long-term community benefit. The most sustainable models align technological development with community empowerment.
- **Diversity within languages matters.** Even small languages have dialects, regional variations, and social differences. Effective language technology must account for this internal diversity rather than assuming homogeneity.
- **Innovation happens at the margins.** Some of the most important breakthroughs came from constraints. Limited resources forced creative solutions. Community needs drove technical innovation. Working with "low-resource" languages revealed possibilities that weren't visible when working with well-resourced languages.

The future: What comes next

As KenCorpus enters its next phase, the vision is expanding. Speech recognition systems are being deployed in local schools. Translation tools are being integrated into government services. Community members are being trained as data collectors and language technology specialists.

But perhaps most importantly, a new generation of young Kenyans is growing up understanding that their native languages are not barriers to technological participation – they are pathways to it. Children are learning that speaking Dholuo or Kalenjin isn't a limitation; it's a superpower that makes them uniquely valuable in an increasingly multilingual digital world.

The project has also inspired similar initiatives across Africa and beyond. In Nigeria, researchers are applying KenCorpus methodologies to Yoruba and Igbo. In South Africa, similar work is beginning with Xhosa and Zulu. Indigenous communities in the Americas are adapting the community engagement strategies for their own language preservation efforts.

The broader transformation:
From extraction to partnership

KenCorpus represents something larger than a single research project. It embodies a fundamental shift in how technology development can work. Instead of Silicon Valley companies extracting data from global communities to build products sold back to them, KenCorpus demonstrates true technological partnership.

Communities aren't just data sources; they're co-designers, co-owners, and primary beneficiaries. Technology isn't imposed from outside; it emerges from community needs and community participation. Linguistic diversity isn't a problem to be solved; it's a resource to be celebrated and empowered.

This model has implications far beyond language technology. As AI systems become more central to education, healthcare, governance, and economic life, the KenCorpus approach offers a template for ensuring that technological advancement serves community empowerment rather than community marginalization.

Mapping the AI Lifecycle HRIA Framework for the KenCorpus case

1 Stage 1: Objective and Team Composition

Problem Definition: KenCorpus began with a clear understanding that the digital marginalization of African languages wasn't just a technical problem – it was a human rights issue. The objective emerged directly from community needs rather than technological possibilities. Dr. Wanzare and her team recognized that less than 0.01% of the world's languages were supported by NLP tools, leaving 3 billion speakers without access to digital language technologies.

Team Composition & Community Partnership: The project exemplified participatory development from the outset. The team composition evolved to include:

- Academic researchers (Dr. Wanzare and university partners).
- Community leaders (chiefs, elders, religious leaders).
- Educational partners (teachers, school administrators).
- Linguistic experts (native speakers, cultural specialists).
- Government representatives (county officials).
- Media partners (local broadcasters, content creators).
- Technical specialists (ML engineers, linguists).

Human Rights integration

The project directly addressed multiple human rights principles:

- **Cultural rights:** Preserving and promoting linguistic heritage.
- **Participation rights:** Communities as co-designers, not data subjects.
- **Non-discrimination:** Ensuring technological access regardless of language.
- **Self-determination:** Communities controlling their linguistic data.

Key decisions made

- Kiswahili chosen as anchor language rather than English (cultural appropriateness).
- Community needs prioritized over technical convenience.
- Long-term sustainability valued over short-term data extraction.
- Traditional knowledge systems respected alongside academic expertise.

2 Stage 2: Defining System Requirements

Value Ecosystem Navigation

KenCorpus navigated complex trade-offs between different values:

- **Accuracy vs. Cultural appropriateness:** Choosing community-validated translations over technically optimized ones.
- **Efficiency vs. Inclusivity:** Including multiple dialects despite increased complexity.
- **Speed vs. Sustainability:** Building long-term community relationships over rapid data collection.
- **Standardization vs. Authenticity:** Preserving natural language variation while creating usable datasets.

Community-Driven Requirements

System requirements emerged through extensive community consultation:

1. **Translation systems**: Cross-language communication (local-to-local, not just English-centric).
2. **Speech recognition:** Automatic transcription in native languages.
3. **Fundamental NLP infrastructure:** Grammar checking, spell checking, part-of-speech tagging.
4. **Cultural preservation:** Maintaining oral traditions in digital form.
5. **Educational support:** Tools for language learning and literacy.

Explainability & Transparency

The project prioritized community understanding over technical sophistication:

- Explanations provided in culturally appropriate formats.
- Community members trained to understand system capabilities and limitations.
- Decision-making processes made transparent to all stakeholders.
- Clear documentation of why certain approaches were chosen.

Accountability Structures

- Community representatives included in all major decisions.
- Regular feedback sessions with language speakers.
- Cultural appropriateness reviews by elders and traditional authorities.
- Academic oversight balanced with community sovereignty.

3 Stage 3: Data Discovery

Ethical Data Collection.

KenCorpus revolutionized data collection by prioritizing community ownership:

- **Consent processes:** Developed in consultation with traditional authorities.
- **Cultural protocols:** Respected storytelling traditions and sacred knowledge boundaries.
- **Community licensing:** Pioneered community-based data ownership models.
- **Benefit sharing:** Ensured communities retained control over their linguistic data.

Addressing Historical Bias

The project confronted multiple forms of bias:

- **Colonial bias:** Rejecting English-centric approaches in favor of indigenous frameworks.
- **Urban bias:** Actively seeking rural and traditional speakers.
- **Educational bias:** Including non-literate speakers as valuable contributors.
- **Gender bias:** Ensuring women's voices and perspectives were included.
- **Generational bias:** Capturing both traditional and contemporary language use.

Data Diversity & Representation

- **Geographic diversity:** Coastal, highland, and urban dialect variations.
- **Social diversity:** Different educational backgrounds, age groups, professions.
- **Linguistic diversity:** Formal and informal registers, code-switching patterns.
- **Cultural diversity:** Different storytelling traditions, ceremonial language use.

Documentation & Preservation

- Raw audio preserved alongside processed datasets.
- Cultural context documented for each collection session.
- Metadata included information about speakers, contexts, and cultural significance.
- Version control maintained to track changes and improvements.

4 Stage 4: Selecting and Developing a Model

Model Architecture Decisions

KenCorpus made strategic choices that prioritized community needs:

- **Multilingual architecture:** Kiswahili as anchor rather than English-centric design.
- **Small, specialized models:** Domain-specific rather than general-purpose systems.
- **Explainable approaches:** Interpretable models over black-box systems.
- **Modular design:** Components could be updated independently as communities evolved.

Fairness Considerations

- **Cross-dialectal fairness:** Ensuring systems worked across regional variations.
- **Intersectional analysis:** Considering gender, age, education, and regional factors.
- **Performance equity:** Avoiding accuracy disparities between different groups.
- **Cultural fairness:** Respecting different ways of expressing concepts.

Technical Innovation

- **Minimum viable data research:** Determining smallest datasets needed for functionality.
- **Code-switching capabilities:** Handling natural language mixing patterns.
- **Tonal language processing:** Accounting for tone markers and prosodic features.
- **Low-resource optimization:** Maximizing performance with limited training data.

Community Validation

- Native speakers involved in model testing and refinement.
- Cultural appropriateness evaluated by community authorities.
- Performance tested in real-world community contexts.
- Feedback loops established for continuous improvement.

5 Stage 5: Testing and Interpreting Outcomes

Multi-Stakeholder Testing

Testing involved diverse community members:

- **Native speakers:** Accuracy and naturalness evaluation.
- **Community leaders:** Cultural appropriateness assessment.
- **Educators:** Pedagogical effectiveness testing.
- **Technical users:** System reliability and performance evaluation.

Performance Metrics

Beyond technical accuracy, KenCorpus evaluated:

- **Cultural appropriateness:** Does the system respect traditional language use?
- **Community acceptance:** Do speakers feel their language is well-represented?
- **Practical utility:** Do the tools meet actual community needs?
- **Fairness across groups:** Do all community segments benefit equally?

Extreme Case Testing

- **Rare dialects:** Testing with less common regional variations.
- **Code-switching:** Evaluating mixed-language scenarios.
- **Cultural contexts:** Testing in ceremonial and formal contexts.
- **Technical edge cases:** Handling poor audio quality, background noise.

Documentation for Users

- **Community-friendly manuals:** Explanations in local languages and cultural contexts.
- **Training materials:** Building local capacity for system use and maintenance.
- **Limitation documentation:** Clear explanation of what systems can and cannot do.
- **Best practices:** Guidance for optimal use in different contexts.

6 Stage 6: Deployment & Post-Deployment Monitoring

Community-Controlled Deployment

- **Community consent:** Final deployment required explicit community approval.
- **Phased rollout:** Gradual implementation allowing for adjustment and feedback.
- **Local ownership:** Communities retained control over how systems were used.
- **Opt-out mechanisms:** Clear pathways for communities to withdraw participation.

Ongoing Monitoring Systems

- **Community feedback channels:** Regular mechanisms for reporting issues or suggestions.
- **Cultural evolution tracking:** Monitoring how language use changes over time.
- **Performance monitoring:** Continuous assessment of system accuracy and fairness.
- **Usage pattern analysis:** Understanding how communities actually use the tools.

Adaptive Management

- **Regular system updates:** Incorporating new community feedback and needs.
- **Dialect evolution:** Accounting for natural language change over time.
- **Technology evolution:** Updating systems as new approaches become available.
- **Community capacity building:** Training local experts for ongoing maintenance.

Impact Assessment

- **Language vitality metrics:** Measuring impact on language use and transmission.
- **Community empowerment:** Assessing changes in technological access and agency.
- **Educational outcomes:** Evaluating impact on literacy and learning.
- **Cultural preservation:** Measuring success in maintaining oral traditions.

Long-term Sustainability

- **Local expertise development:** Training community members as technical specialists.
- **Institutional partnerships:** Building sustainable relationships with schools, g overnment, media.
- **Financial sustainability:** Developing models that don’t depend on external funding.
- **Replication support:** Helping other communities adapt the methodology

Conclusion:
A New Paradigm for AI Development

KenCorpus demonstrates that AI development can be genuinely participatory, culturally appropriate, and community-empowering. By integrating human rights considerations throughout the AI lifecycle, the project shows how technology can serve linguistic diversity rather than undermining it.

The project's success lies not just in its technical achievements, but in its demonstration that communities can be equal partners in shaping their technological future. When AI development prioritizes human dignity, cultural preservation, and community empowerment, the resulting systems are not only more ethical – they're more effective, more sustainable, and more innovative.

As AI systems become increasingly central to human life, the KenCorpus model offers a roadmap for development that enhances rather than diminishes human diversity. It proves that the choice between technological advancement and cultural preservation is a false one – with the right approach, technology can be the most powerful tool for cultural empowerment and human flourishing.

About the case study

This case study analyzes research conducted by Dr. Lilian Wanzare, Prof. Florence Indede, Dr. Owen McOnyango of Maseno University, Dr. Edward Ombui of USIU (then African Nazarene University), Dr. Lawrence Muchemi and Mr. Benard Wanjawa of University of Nairobi, and the KenCorpus language community, examining Languages spoken in Kenya in the lens of Natural Language Processing across several counties in Kenya between 2021 - 2022. This research was made possible by funding from Meridian Institute's Lacuna Fund under grant no. 0393-S-001 which is a funder collaboration between The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre.

Dr. Lilian Wanzare is a lecturer and chair of the Department of Computer Science at Maseno University. Her research interests are in Artificial Intelligence and Machine Learning, in particular Natural Language Processing (NLP), Sign Language research and building text processing tools for low-resource languages. She holds a PhD degree in Computational Linguistics and an Msc. in Language Science and Technology from Saarland University, Germany.

Other contributors to this case study are Caitlin Kraft-Buchman, Emma Kallina, and Sofia Kypraiou, authors of the original *Framework to AI Development: Integrating Human Rights Considerations Along the AI Lifecycle* upon which the Toolbox structure is based. Additional contributors are Amina Soulimani and Pilar Grant, from Women at the Table and the <AI & Equality> Human Rights Initiative.

Resources

Dataset Locations

- https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6N5V1K
- https://commonvoice.mozilla.org/luo
- https://commonvoice.mozilla.org/dav
- https://commonvoice.mozilla.org/kln
- https://zenodo.org/records/13355021

Publications

- Wanjawa, B., Wanzare, L., Indede, F., McOnyango, O., Ombui, E., & Muchemi, L. (2022). Kencorpus: A kenyan language corpus of swahili, dholuo and luhya for natural language processing tasks. arXiv preprint arXiv:2208.12081.
- Babirye, C., Nakatumba-Nabende, J., Katumba, A., Ogwang, R., Francis, J. T., Mukiibi, J., ... & David, D. (2022). Building text and speech datasets for low resourced languages: A case of languages in east africa. In 3rd Workshop on African Natural Language Processing.
- Wanjawa, B. W., Wanzare, L. D., Indede, F., McOnyango, O., Muchemi, L., & Ombui, E. (2023). KenSwQuAD—A Question Answering Dataset for Swahili Low-resource Language. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1-20.
- Awino, E., Wanzare, L., Muchemi, L., Wanjawa, B., Ombui, E., Indede, F., ... & Okal, B. (2022). Phonemic Representation and Transcription for Speech to Text Applications for Under-resourced Indigenous African Languages: The Case of Kiswahili. arXiv preprint arXiv:2210.16537.
- Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Adelani, D. I., Ahmad, I. S., Aliyu, S. M., ... & Ousidhoum, N. (2025). AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages. arXiv preprint arXiv:2501.08284.
- Amol, C. J., Chimoto, E. A., Gesicho, R. D., Gitau, A. M., Etori, N. A., Kinyanjui, C., ... & Tombe, R. (2024). State of NLP in Kenya: A Survey. arXiv preprint arXiv:2410.09948.
- Amol, C., Wanzare, L., & Obuhuma, J. (2023, December). Politikweli: A swahili-english code-switched twitter political misinformation classification dataset. In International Conference on Speech and Language Technologies for Low-resource Languages (pp. 3-17). Cham: Springer Nature Switzerland.
- Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Wahle, J. P., Ruas, T., Beloucif, M., ... & Mohammad, S. M. (2025). BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages. arXiv preprint arXiv:2502.11926.
- Amol, C., Wanzare, L., & Obuhuma, J. ()2025) Modelling Misinformation in Swahili-English Code-switched Texts. mecs-press.org
- Mbogho, A., Awuor, Q., Kipkebut, A., Wanzare, L., & Oloo, V. (2025). Building low-resource African language corpora: A case study of Kidawida,Kalenjin and Dholuo. arXiv preprint arXiv:2501.11003.