data.org

Digitisation of Oral Data for NLP of Low-Resource Languages

PRACTICAL METHODS AND PROCESSES FOR SCALABLE AND SUSTAINABLE ECOSYSTEM DEVELOPMENT

AUTHORS:

Tsosheletso Chidi
DSFSI at University of Pretoria

Nontokozo Manukuza
DSFSI at University of Pretoria

Vukosi Marivate
DSFSI at University of Pretoria

Shaimaa Lazem
ArabHCI

Tajuddeen Gwadabe
Masakhane

Victor Odumuyiwa
Nithub at University of Lagos

Ayomide Fagoroye
Nithub at University of Lagos

Similoluwa Ola-obaado
Nithub at University of Lagos

Kingsley Ogbonna
Nithub at University of Lagos

Olubayo Adekanmbi
Data Science Nigeria

Oluwaseun Nifemi
Data Science Nigeria

Gideon George
Data Science Nigeria

Emmanuel Davis
Data Science Nigeria

Paul Korir
data.org

Danil Mikhailov
data.org

Uyi Stewart
Mastercard Center for Inclusive Growth

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Table of Content

Acronyms

- NLP – Natural Language Processing: A field of AI focused on enabling computers to understand, interpret, and generate human language.
- AI – Artificial Intelligence: The simulation of human intelligence in machines, allowing them to perform tasks such as reasoning, learning, and decision-making.
- NLMs – Neural Language Models: AI models that use neural networks to understand and generate human language.
- SLMs – Small Language Models: Lightweight language models designed for efficiency, often used in low-resource or specialized applications.

Overview

Introduction

Africa is home to over 2,000 languages[1], making it one of the most linguistically diverse regions in the world. Yet, despite being spoken by millions, most African languages remain underrepresented in AI and Natural Language Processing (NLP) systems[2]. Only a handful of languages, such as English, Mandarin, and French, have significant digital representation. This leaves low-resource languages (LRLs) in a state of digital silence, creating barriers to knowledge, connection, and inclusion.

The challenges for African Low Resource Languages (LRLs) are multi-layered:

- **Data scarcity:** Many languages lack large text or speech datasets, limiting the ability of AI systems to learn effectively.
- **Oral tradition:** Some languages are primarily spoken with little written material, making corpus creation difficult.
- **Linguistic complexity:** Features such as tonal shifts (e.g., in Igbo and Yorùbá) or critical diacritics (e.g., ṣ vs. s in Yorùbá) are often lost in preprocessing, which reduces model accuracy.
- **Framework limitations:** Most mainstream NLP tools assume structures and rules that do not apply to many African languages which result in poor performance.

These challenges are not just technical, they impact access and equity[3]. For instance, many African languages are unsupported by tools like Google Translate, Siri, or Alexa, making digital resources inaccessible to non-English or non-French speakers.

[1] Ethnologue - Africa https://www.ethnologue.com/region/Africa/

[2] Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. "The state and fate of linguistic diversity and inclusion in the NLP world." arXiv preprint arXiv:2004.09095 (2020).

[3] Adebara, Ife. AI and Language Data Flaring in Africa: Addressing the Low-Resource Challenge. Centre for International Governance Innovation, https://www.cigionline.org/publications/ai-and-language-data-flaring-in-africa-addressing-the-low-resource-challenge/. Policy Brief No. 216.

Understanding the Landscape of African Languages

Africa has one of the most diverse linguistic landscapes in the world. The continent's languages differ not only in their sounds and grammar but also in how they reflect culture, values, oral traditions, and history. To understand this diversity, it is important to pay attention to both the cultural depth that shapes African languages and the technical issues that come with building digital tools that can handle them properly.

Many African languages are described as low-resource even though they are spoken by millions of people. This description often arises from several factors such as the lack of standard writing systems, the limited availability of written or recorded materials, and the low presence of these languages in schools and on the internet. However, this situation is not just a matter of missing data. It is tied to deeper issues such as colonial language policies, unequal access to technology, and the way most existing language tools fail to recognise the unique nature of African languages.

This playbook treats the idea of low-resource not as a technical label but as the result of years of historical, infrastructural, and institutional neglect. To bridge the gap between modern technology and African languages, it is necessary to first understand their structures, patterns, and the social contexts in which they are used.

Several features make African languages complex to represent and process digitally:

1. **Complex Word Formation**
   Many African languages form words by joining smaller units together, allowing a single word to carry a lot of meaning. This process, known as agglutination, affects how words relate to one another and how meaning is built in sentences. It also reflects the way many African cultures view relationships and communication as connected and harmonious.

2. **Tonal Variation**
   In many African languages, the tone of a word, how the pitch rises or falls, can change its meaning completely. A single word can mean different things depending on how it is said. This makes the language rich in expression but also harder to handle when creating written or spoken systems that can capture these variations.

3. **Different Writing Systems**
   Languages across Africa use a variety of writing systems including Latin, Arabic, and Ge'ez scripts. Some are still mainly spoken, with strong oral traditions passed down over generations. Finding ways to preserve these oral forms while meeting the needs of modern technology remains an important task.

Apart from these features, another factor that shapes the label low-resource is the small amount of online text and recorded speech available for many African languages. The next section discusses this issue in more detail and explores how these languages are represented in digital spaces.

Categorization of African Languages by the Amount of Available Digital Data

African languages can be grouped by the amount of digital text and recorded speech available for each language. This includes unlabelled material such as web text and Wikipedia articles, and labelled material such as curated datasets used for research. These groupings show how much technological and digital support a language currently has, based on the amount of available data and research resources. Based on the work of Joshi et al. (2020) and Adelani (2025), a six-level classification of resource availability in African languages is presented in Figure 1 indicates the description of each level alongside its data size range based on the CC100-XL (Common Crawl 100 – Extra Large) dataset.

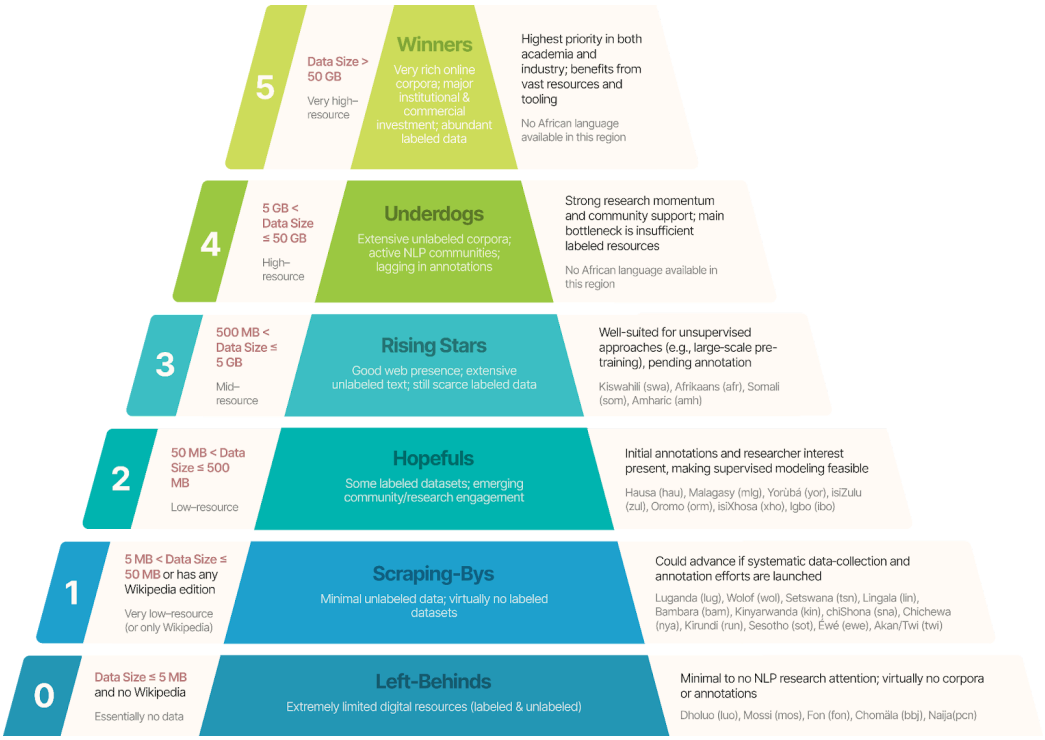

Figure 1: Taxonomy based on African languages resource availability and CC100-XL corpus size

The CC100-XL dataset is a high quality monolingual dataset of at least 100 languages crawled from the Web, representing publicly available digitized content in these languages (Lin et. al., 2022). From the diagram above, according to the CC100-XL dataset, no African language has a corpus size of 5GB and above. This shows that many of these languages remain low-resource and need more investment in data collection as well as research and development. To do this, the next step is to look at the main NLP tasks discussed in existing studies and the progress made so far in Africa. The goal is to help readers understand which areas have been dealt with and which still need attention. Identifying these gaps can inspire further work and practical engagement. The following section focuses on the categorisation of NLP tasks.

Categorisation of NLP Tasks

Natural Language Processing (NLP) covers a wide range of tasks, but most can be grouped into five main categories. Each category reflects a specific way computers process or generate human language. Below is a concise summary:

1. **Classification Tasks**
   These involve sorting text into predefined groups. For example, identifying whether a tweet is positive or negative (sentiment analysis) or detecting if a post contains hate speech. Classification tasks help systems organise and understand language at a general level. Several efforts have been made in Sentiment classification in multiple African languages. This is one of the most explored NLP tasks in LRLs. However tasks like multi-label classification, fine-grained classification tasks such as detecting stance, sarcasm, humor vs. non-humor (especially code-switched or multilingual settings), and classification tasks for domain-specific content (e.g. legal, medical, agricultural) in LRL have been under explored.

2. **Sequence Labelling Tasks**
   Here, each word or phrase in a sentence is given a label. For instance, identifying names of people or places in a news article (Named Entity Recognition) or tagging each word by its grammatical role, like noun or verb. While some exploration of Named Entity Recognition (NER) and Part-of-Speech (POS) tagging for some LRLs have been recorded in AfricaNLP, however tasks like coreference resolution, discourse parsing and information extraction in dialogue settings are largely missing.

3. **Generation and Transformation Tasks**
   This category covers creating or reformulating text. Examples include translating between languages, summarising long reports, or rephrasing sentences while keeping the same meaning. Machine Translation (MT) between African languages and English have recorded some active research and development. However, tasks like abstractive summarization and dialogue generation are still very under-explored.

4. **Retrieval and Question-Answering (Q&A) Tasks**
   These tasks help systems find and deliver information that answers a question. Search engines and digital assistants rely on this ability, for instance, retrieving the date when Nigeria gained independence. Though some Q&A datasets exist in African languages to support Q&A tasks, multilingual Q&A and fact-checking/claim verification tasks in LRLs are underexplored.

5. **Speech Tasks**
   Speech-based systems convert spoken words into text (speech recognition) or generate speech from text (text-to-speech). They are behind voice assistants, automated captions, and accessibility tools for the visually impaired. Existing works in AfricaNLP include speech representation models like AfriHuBERT and datasets like NaijaVoices (Yorùbá, Igbo, Hausa). However, spoken dialogue systems in LRL are largely underexplored.

Figure 2 presents the current landscape of Africa NLP highlighting the categories discussed and some key datasets (more information on existing datasets is provided in Appendix A). Several NLP initiatives for African languages have emerged in recent years across the five categories of NLP tasks. These efforts combine careful data collection, community involvement, and modern techniques for building AI models. Initiatives like Masakhane have been driving research, dataset creation, and advocacy. Progress has also been made in speech processing and adaptation of large language models for African languages. However, existing research and resources have only focused on a small portion of the continent's linguistic diversity.

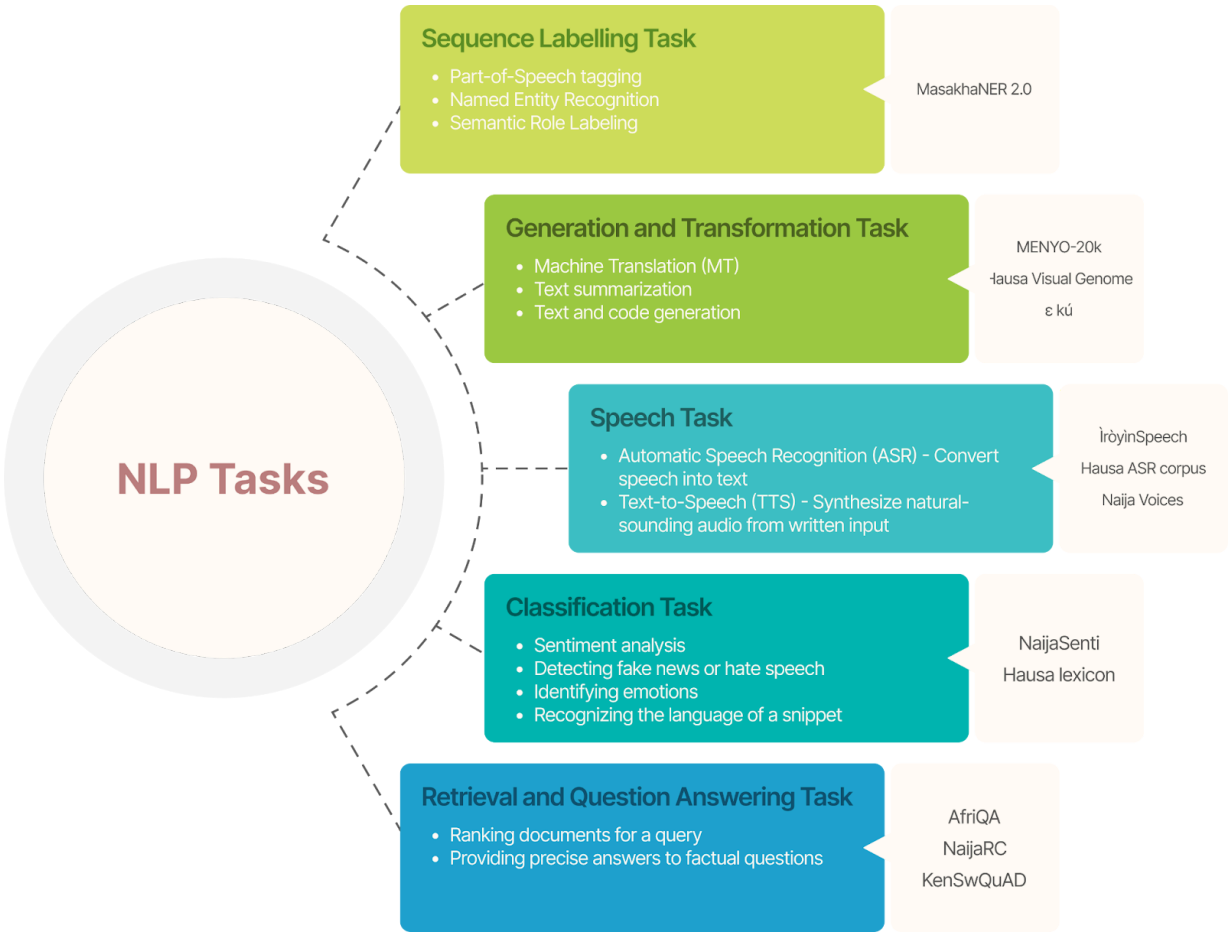

Figure 2: Landscape of NLP for African Languages: Categories and Core Datasets

Improving African NLP requires large-scale and ethical data collection involving native speakers, linguists, and communities. It also involves the creation of benchmarks for fair evaluation and a stronger collaboration between universities, tech hubs, and research centers. This requires training of native speakers in data collection and cleaning, which will be the focus of the next section of this chapter.

Summary of Chapters

The playbook opens by setting out a deliberately ambitious and human-centred vision for the digitisation of African low-resource languages. Chapter One positions this work not as a narrow technical exercise but as a collective undertaking that sits within a broader ecosystem of human collaboration, policy reform, and technological innovation. It reframes Africa's immense linguistic diversity—so often treated as an obstacle—as a reservoir of cultural wealth, arguing that language itself is a form of digital equity and a cultural catalyst. The chapter insists that the disappearance of a language from digital spaces is never a trivial loss of vocabulary; it is the disappearance of a worldview. The playbook is presented as a direct response to this crisis of digital exclusion.

To avoid the naïveté that often plagues language-technology projects, the chapter introduces an ecosystemic approach rooted in systems thinking. Drawing on the Quadruple Helix Model, it positions communities, academia, industry, and government not as isolated stakeholders but as interconnected actors in a single developmental loop. The chapter outlines a methodology that has already been tested across Africa: beginning with deep engagement of language communities, proceeding to ontology design, followed by structured participant recruitment, and then moving into rigorous data collection through standardised and ethical recording protocols. It closes with an emphasis on validation—multi-layered transcription, annotation, and peer review by both linguists and native speakers—to ensure technical accuracy and cultural fidelity.

The chapter then widens its scope, showing how this ecosystemic framework enables bridges between local and global actors. Governments emerge as enablers that can institutionalise ethical AI policies, national repositories, and funding pipelines. Academic institutions contribute research standards and methodological rigour. Private-sector innovators help to ensure that language technologies reach real markets while safeguarding against exploitative commercialisation. International partners and funders are cast not as patrons but as collaborators in a shared pursuit of digital justice. These commitments are organised into five strategic pillars: knowledge sharing, capacity building, innovation and research, policy advocacy, and international collaboration. The chapter concludes with concrete case studies from ongoing African language initiatives, demonstrating that although progress is unevenly distributed, it is undeniably underway.

Chapter Two shifts the focus from vision to the technical realities of digitising African languages for Natural Language Processing tasks especially as it relates to audio (oral) data collection and processing. It argues for a community-driven, ethical workflow for data collection and cleaning.

This workflow unfolds through five stages: Foundational Readiness and Ethical Grounding, Ontology Development and Prompt Design, Participant Recruitment and Distribution Logic, Data Collection and Technical Quality Assurance, and Data Processing and Validation. The chapter makes a pointed argument for large-scale, community-led data collection efforts and contends that only through collaboration among linguists, native speakers, researchers, and technologists can African languages move from digital silence to meaningful digital presence.

Chapter Three confronts the question that haunts most initiatives in African NLP: how can digitisation scale sustainably, rather than appearing in isolated bursts of activity? Drawing on a combined methodology of literature review and a continent-wide survey of ninety practitioners, the chapter maps current capacities, gaps, and opportunities across the ecosystem. It reveals persistent fragmentation, with many projects concentrating on single languages and regions receiving uneven attention. Anchored in a Theory of Change perspective, the chapter argues that sustainable scaling requires more than additional data and better tools. It demands participatory models, ethical frameworks, and accessible infrastructure that communities can use autonomously—without reliance on external gatekeepers.

The survey findings illuminate how practitioner roles shape which languages receive attention, reinforcing inequalities that are rarely acknowledged. The chapter also scrutinises dynamics of inclusion and exclusion in African NLP. It raises a central ethical concern: participation itself is shaped by language accessibility. Inclusive digitisation, it argues, must accommodate plain-language communication, multilingual participation channels, and culturally grounded narrative forms—not merely written, standardised, or academic conventions—as valid expressions of expertise.

The chapter concludes with a set of recommendations that reject technocratic shortcuts. It calls for the democratisation of training, the fostering of interdisciplinary collaboration, stronger mentorship structures, and the development of open, accessible tools. It encourages the adoption of evaluation metrics that privilege inclusion and local relevance rather than scale alone. Ultimately, the chapter asserts that African NLP will be sustainable only if it values social impact alongside technical achievement, and only if it nurtures an ecosystem in which contributors can express themselves clearly, engage confidently, and succeed on their own terms.

Chapter 1: A Holistic Ecosystem for African Language Technology

Introduction and Ecosystem Overview

With its rich tapestry of over two thousand living languages, Africa has embarked on a decisive voyage of digital integration. Fueled by foundational efforts in language digitization, the continent is now charting an innovative course that aligns with global benchmarks. This proactive stance is crucial to steer clear of the perils of digital silence, economic marginalization, and the erosion of cultural memory and instead, to move firmly into a future of reliable inclusion and ongoing opportunity. Language technologies like Automatic Speech Recognition (ASR) and AI-driven translation serve as the essential tools creating the necessary on-ramps for every community.

This chapter offers a blueprint to support the African leaders, researchers, and communities who are choosing this hopeful future of cultural preservation and shared opportunity. It provides a comprehensive framework for them to construct a sustainable and inclusive language technology ecosystem on their own terms

The chapter is structured to guide the reader from theory to practice. Section 1 begins by mapping the current ecosystem of community, academic, commercial, and governmental actors. Section 2 presents a detailed, five-stage methodology for creating high-quality audio datasets, forming the technical core of the guide. Section 3 expands on the key actors and the collaborative trends that are accelerating progress. Finally, Section 4 offers five pillars of actionable strategic recommendations for achieving long-term, sustainable impact, followed by a concluding summary. The chapter concludes with Appendices including a glossary, an ecosystem actor directory, a resource toolkit, and a full bibliography.

However, successfully navigating this challenging transformation at the necessary scale for low-resource languages requires a concerted, strategic effort, demanding a holistic ecosystem of interconnected actors, resources, and activities, as illustrated in Figure 1 below:

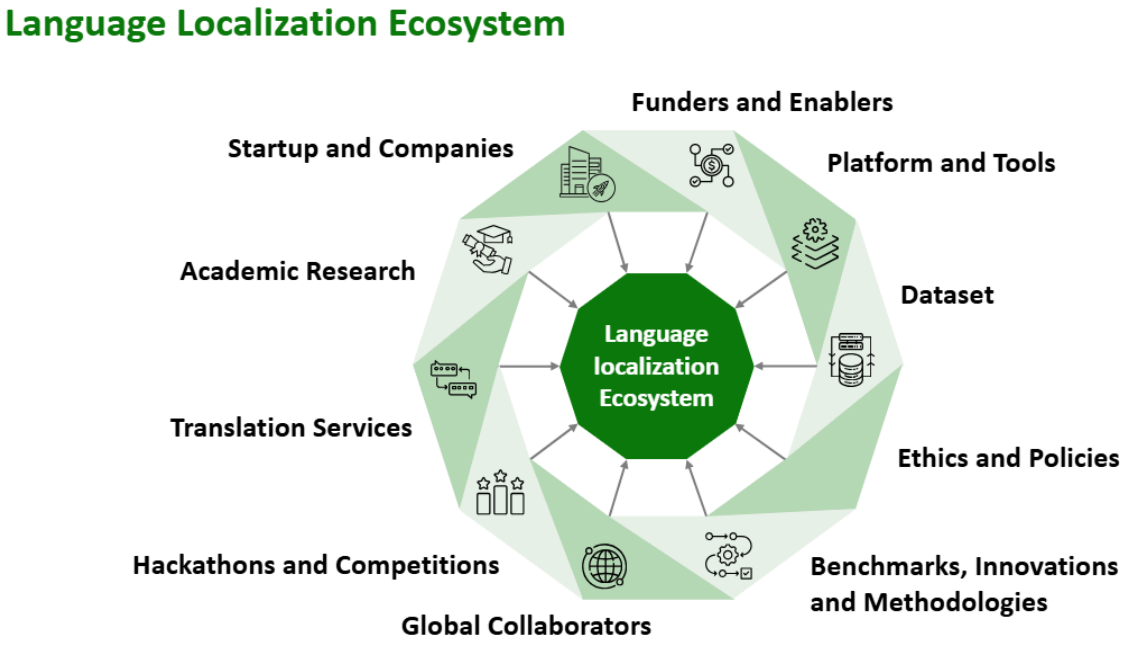

Figure 1: The Ecosystem: From Literacy to Cultural Preservation

The reward for choosing to pave this path is the series of remarkable destinations it is unlocking—from enhanced literacy and information access to ensuring cultural reservation—as illustrated in Figure 2 below:

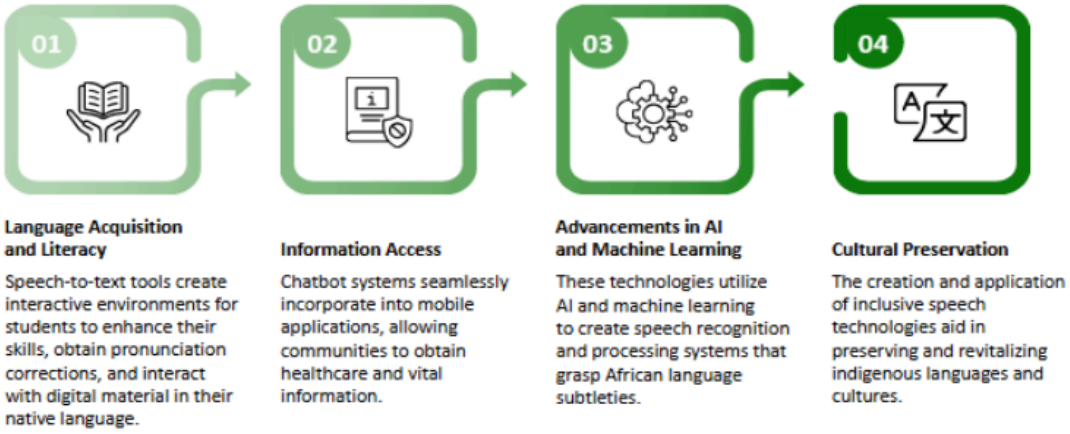

Figure 2: Impact Cascade: From Literacy to Cultural Preservation

The journey to digitize Africa's languages is not powered by a single source; instead, several distinct forces fuel the movement. First, pioneering startups like Lelapa AI (South Africa), Vambo AI (South Africa), and EqualyzAI (Nigeria) forge new commercial pathways, crafting enterprise-grade tools that prove the viability of the route [1-3]. Second, grassroots, community-led movements provide the core momentum. Pan-African initiatives like Masakhane and the African Languages Lab (All Lab) mobilize volunteers to build the shared infrastructure—the open-source maps and engines—that enable the entire ecosystem [4, 5]. Finally, strategic industry-academic collaborations act as powerful accelerators. Partnerships between global tech companies, telecom operators like Orange, and university hubs like NITHub (University of Lagos) connect global resources with local expertise to widen the digital on-ramps for more communities [6, 7]. As illustrated in Figure 3, these three currents of innovation converge, creating a synergy that powers a robust and self-sustaining ecosystem.

Figure 3: The Key Actors of the African Language Technology Ecosystem

Ecosystem Actors and Collaborative Momentum

While the journey to digital inclusion requires the map of accurate data and a compass of ethical sourcing, the network of actors itself is a dynamic, living ecosystem. The sustainable digitization of African languages requires more than the work of any single entity. Such a complex achievement will ultimately necessitate a collective effort driven by a diverse and interconnected network of actors, each playing a vital role. Understanding how these actors—from the grassroots organisms to the canopy-level institutions—collaborate is essential for fostering the health and growth of this ecosystem. This section maps these key players and highlights the symbiotic relationships that are accelerating success.

Key Actors in the Ecosystem

The following table categorizes the primary actors in the ecosystem, providing examples for each. A more detailed directory is available in Appendix B.

Category	Type of Actor	Examples
Community & Grassroots	Open-Source Movements & Crowdsourcing Platforms	Masakhane, GhanaNLP, TangaleNLP Project, African Languages Lab (All Lab) [4, 5, 8-10]
Academia	University Research Labs & Academic Centers	NITHub (University of Lagos), Africa Centre of Excellence on Technology Enhanced Learning (ACETEL) [6, 11]
Private Sector	AI-Native Startups & Language Service Providers (LSPs)	EqualyzAI, Lelapa AI, Vambo AI, African Translation, AfriLingual [1-3, 12-14]
Government & Policy	National AI Centers & Policy Advocates	Data Science Nigeria (in its policy role), Nigeria's NCAIR, South Africa's SADiLaR [15-17]
Global & Philanthropic	Global Tech Companies, Funders & Ecosystem Builders	Google, Meta, OpenAI, Lacuna Fund, BMGF, AfriLabs [7, 18-20]

Table 1: Key Actors in the Ecosystem

This diverse network of actors does not operate in isolation. Their interactions form several distinct models of global and local collaboration that are essential for building a sustainable digital language ecosystem, as outlined in Figure 10.

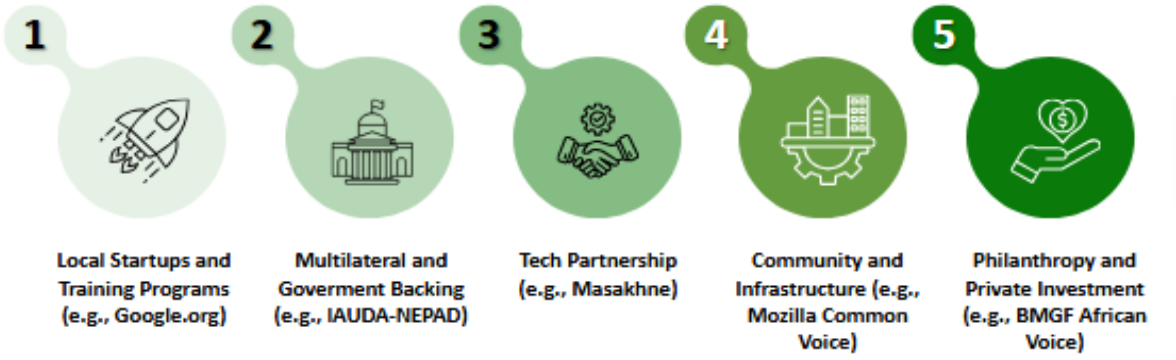

Figure 10: Models of Global Collaboration in Language Digitization

Emerging Collaborative Trends

Several key trends define the interaction between these actors:

- **Community-Industry-Academia-Government Helix:** The most successful projects emerge from a multi-stakeholder model of collaboration, often called a "quadruple helix," where academic research, community data, private sector application, and government policy reinforce one another in a virtuous cycle (see Figure 11). For example, a university research lab, supported by a government grant from an agency like Nigeria's NCAIR, might develop a new model. This model is then trained on data collected by a community group like Masakhane and finally productized by a startup like Lelapa AI to meet market needs fostered by national digital inclusion policies.

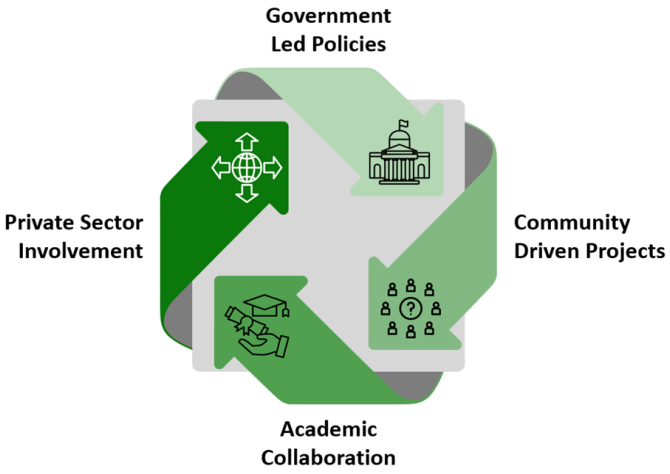

Figure 11: The 'Quadruple Helix' Model of Ecosystem Collaboration

- **Open-Source as a Default:** The ecosystem has a strong ethos of openness. Datasets like MasakhaNER and models like AfriBERTa are made publicly available, allowing researchers to build upon existing work. This creates a virtuous cycle, or "flywheel effect," where each contribution builds momentum for the next.
- **From Raw Data to Practical APIs:** There is a clear trend toward building practical, usable tools and platforms that other developers can easily integrate into their applications. This includes translation APIs, AI-powered keyboards, and voice-based services from telecom operators.
- **A Focus on Hyper-Local Context:** Both startups and community projects are emphasizing the need for "hyper-local" models that understand not just the language but also the cultural context, slang, and code-mixing prevalent in everyday communication.

Building the Infrastructure: Translation Services

If the journey crosses vast linguistic divides, then Language Service Providers (LSPs) are the master bridge-builders of the ecosystem. They construct the vital communication links that allow the convoy of actors to move forward as one. From community specialists building trusted footbridges to tech-enabled platforms deploying large-scale suspension bridges, their work is critical for connecting the collective efforts in business, government, and healthcare.

- Foundational LSPs: Companies like African Translation, which specializes in over 42 lesser-known languages, and Lagos Translation Service (Yoruba, Igbo, Hausa) provide certified, human-led translation and interpretation. They ensure that critical documents and communications are culturally and linguistically accurate.
- Tech-Enabled Solutions: The next wave of translation services is driven by technology. Startups are building powerful, AI-driven tools that offer scalable solutions. For instance, EqualyzAI's AfroSLM offers hybrid Machine Translation (MT) and NLP APIs for over 15 Nigerian languages, using culturally nuanced datasets to deliver context-aware translations for specific domains like healthcare and education [21]. This synergy of localized human expertise and scalable AI technology is vital for moving the entire ecosystem forward.

Catalyzing Innovation: Hackathons and Competitions

High-impact events like hackathons and innovation challenges act as powerful accelerators, injecting bursts of creative energy into the ecosystem. These short-term, intensive events catalyze rapid innovation, bringing together diverse talents to solve real-world problems and build prototypes under pressure [22, 23].

- Targeting Critical Issues: Events like the AfriLabs Llama 3.1 Impact Hackathon, supported by Meta and BMGF, directly target systemic challenges such as gender bias and linguistic inclusivity in AI, offering mentorship and funding to create tools for languages like Wolof and Swahili [24].
- Aligning with Global Goals: The AI4Good Impact Africa Summit aligns AI innovation with the UN Sustainable Development Goals (SDGs), creating a platform for solutions that address language accessibility and financial inclusion [25].
- Fostering Grassroots Innovation: National initiatives like Data Science Nigeria's AI Bootcamp & Hackathons combine training with practical problem-solving, empowering participants to build solutions for local needs in health, finance, and language inclusion [26].

By linking winners to grants and post-hackathon support, these events transform promising ideas into scalable solutions, fueling the ecosystem's momentum.

Measuring Progress: Benchmarks and Innovations

To navigate the complex terrain of language model development, the ecosystem relies on standardized benchmarks to measure progress and identify areas for improvement. These benchmarks are essential navigational tools that ensure models are not only technically proficient but also linguistically and culturally accurate.

- Sentiment Analysis Benchmarks: AfriSenti provides a critical benchmark for sentiment analysis, using over 110,000 annotated tweets across 14 African languages, including Amharic, Hausa, and Igbo. This allows developers to fine-tune models to understand the nuances of opinion and emotion in local contexts [27].
- Multilingual Evaluation Benchmarks: To expose performance gaps between high-resource and low-resource languages, benchmarks like IrokoBench offer human-translated evaluation datasets for 16 diverse African languages across tasks like reasoning and question-answering [28].
- Key Innovations: Beyond benchmarks, the ecosystem is defined by groundbreaking innovations. Community-led efforts by Masakhane establish standards for core NLP tasks, while research labs like AMMI curate vital text and speech datasets. A significant leap forward is represented by culturally grounded models like EqualyzAI's AfroSLM, a Small Language Model designed for real-time, domain-specific understanding of African languages, demonstrating strong performance in finance and health applications [29].

Together, these benchmarks and innovations provide the steering mechanisms and upgrades needed to accelerate the journey toward robust and equitable language technology.

Strategic Recommendations for Sustainable Impact

As Africa continues its decisive voyage of digital integration, intentional navigation must continue. The course has been charted away from the perilous waters of digital marginalization—and the eventual threat of extinction for many languages—and toward a safe harbor of prosperity, empowerment, and well-being for their speaker communities. To sustain this voyage, however, requires more than the existing momentum; it demands a deliberate, collective effort to power the movement forward across existing gaps. The following recommendations are therefore organized as the five core navigational principles for this mission. This section provides the strategic chart for the vital work needed to ensure the transformation is swift, scalable, and successful for all.

Pillar 1: Community-Driven Initiatives

Sustainable language technology must be rooted in the community. This requires a shift toward co-creation models, culturally embedded products, and fair compensation systems to ensure long-term participation and adoption.To put this principle into practice, stakeholders should prioritize the following key initiatives:

- Foster Community Co-Creation: Actively collaborate with grassroots organizations and native speakers to co-create and crowdsource culturally nuanced language data, following the example of Kenya's UlizaLlama project [63].
- Embed Localization and Cultural Depth: Design products that reflect deep cultural context, as demonstrated by Lelapa AI's partnership with Vodacom in South Africa to deploy a speech-to-text tool for isiZulu speakers [2, 30].
- Implement FAIR (Findable," "Accessible", "Interoperable," and "Reusable) Compensation Models: Implement fair and transparent compensation frameworks to reward local contributors, inspired by models like Malaysia's blockchain-based MaLLaM project [31].

Pillar 2: Academic-Industry Collaboration

Innovation thrives at the intersection of academia and industry. Stakeholders should prioritize collaborations that produce resilient, accessible technologies, embrace open-source principles, and optimize advanced AI techniques.To translate these principles into tangible outcomes, stakeholders should focus on the following strategic priorities:

- Promote Ecosystem Collaboration: Co-develop solutions with key industry players like telecom companies to ensure wide reach, mirroring the success of India's Indus OS keyboard [32, 33].
- Design for Resilience and Accessibility: Create lightweight, offline-first tools suitable for low-bandwidth environments, similar to DeepSeek AI's approach in rural China [34].
- Establish Data Sharing and Open-Source Initiatives: Create shared repositories for African language datasets to accelerate research, following the model of the SEA-LION project in Southeast Asia [35].

- Optimize Multilingual Transfer Learning: Leverage data from linguistically similar, higher-resource languages to enhance model performance, as seen with Vietnam's PhoGPT [36].

Pillar 3: Strategic Funding

A robust funding strategy is critical. By aligning with national priorities, forging tech partnerships, and rigorously measuring impact, the ecosystem can attract the sustained investment needed to scale. To execute this funding strategy effectively, stakeholders should concentrate on the following key actions:

- Align Investment with National Digital Strategies: Secure funding in alignment with national and pan-African digital strategies, as exemplified by India's Digital India Mission [37, 38].
- Leverage Tech-Driven Infrastructure Support: Partner with global tech leaders for access to essential compute resources, which enabled the Masakhane Universal API to power commercial chatbots [39].
- Prioritize Measurable Impact and Holistic Assessment: Track clear success metrics beyond technical performance, including user growth, usability, and social impact indicators like literacy improvements.

Pillar 4: Rapid Prototyping and Domain-Specific Initiatives

Practical impact is achieved by focusing on real-world problems. Stakeholders should use SDG-aligned hackathons to spur innovation, provide support to scale prototypes, and establish standardized benchmarks to drive progress. To translate this focus on real-world problems into tangible results, stakeholders should pursue the following initiatives:

- Focus on Critical Sectors: Prioritize the development of specialized language resources for domains like healthcare and agriculture. AfriLingual's medical translation engine, which reduced patient misdiagnoses in Nigeria, is a prime example [13, 14].
- Launch SDG-Aligned Innovation Challenges: Align hackathons with the UN Sustainable Development Goals (SDGs) to focus creative energy on pressing issues, similar to the Smart India Hackathon [40, 41].

- Scale Prototypes with Post-Event Support: Provide grants and mentorship to help hackathon winners mature their solutions into scalable products, as seen with the AfriLabs Bambara Literacy Tool.
- Establish Task-Specific Benchmarks: Develop standardized, Africa-centric evaluation benchmarks for NLP tasks to systematically measure progress, following the lead of initiatives like IndicNLPSuite and XTREME-R [42, 43].

Pillar 5: Government Advocacy and Infrastructure

Government action is critical for creating a supportive environment. This includes establishing strong data governance, mandating fairness in AI, and leading national initiatives that align technology development with public interest goals. To fulfill this critical role, governments should prioritize the following policy and infrastructure initiatives:

- **Implement Ethical and Open Data Governance:** Governments should lead the implementation of ethical data governance frameworks that prioritize privacy, inclusivity, and open access, drawing inspiration from Rwanda's Digital Umuganda initiative and Singapore's AI Governance Framework [44].
- **Mandate Bias Auditing and Fair AI Representation:** Encourage or mandate regular auditing of AI models for gender, ethnic, and linguistic bias to ensure fairness in deployed systems.
- **Advocate for Digital Language Inclusion Policies:** Develop policy frameworks to formally integrate indigenous languages into official digital services, education, and public communications. These policies should exist within a hierarchy of legal and ethical guidelines, from international principles down to national data laws (see Figure 12).

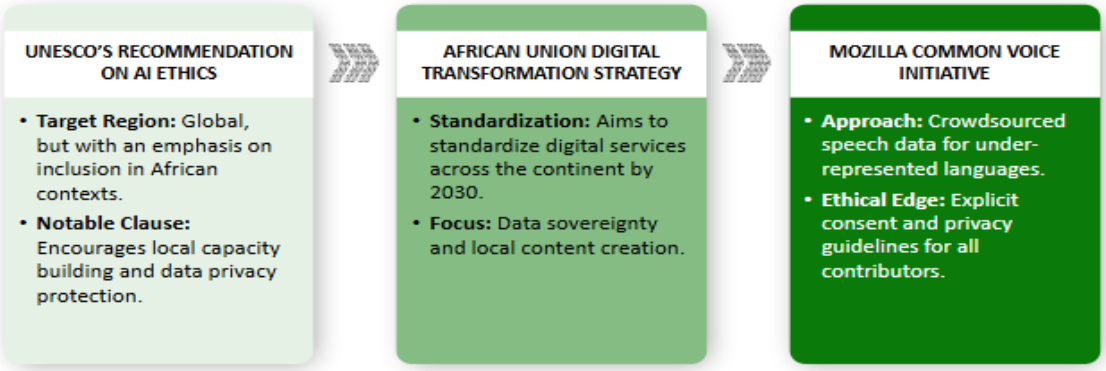

Figure 12: Hierarchy of Ethical and Policy Frameworks for Language Digitization

- Drive Policy-Led Technological Growth: Launch national initiatives to align disparate efforts and drive large-scale progress, modeled on India's successful Bhashini platform [45, 46].

Conclusion: Building the Future of African Languages, Together

This guide started with the assertion that Africa's digital transformation has been successfully charted, away from a future of digital silence and toward one of linguistic inclusion. It has laid out a strategic plan for the journey. This plan details the practical methodology for creating high-quality data and, crucially, illuminates the network of partners—the startups, academics, communities, and governments—who must move in concert.

The core message is clear: no single actor can complete this journey alone. Sustainable progress depends on a holistic approach, a shared mission where all partners advance together, as illustrated in Figure 13.

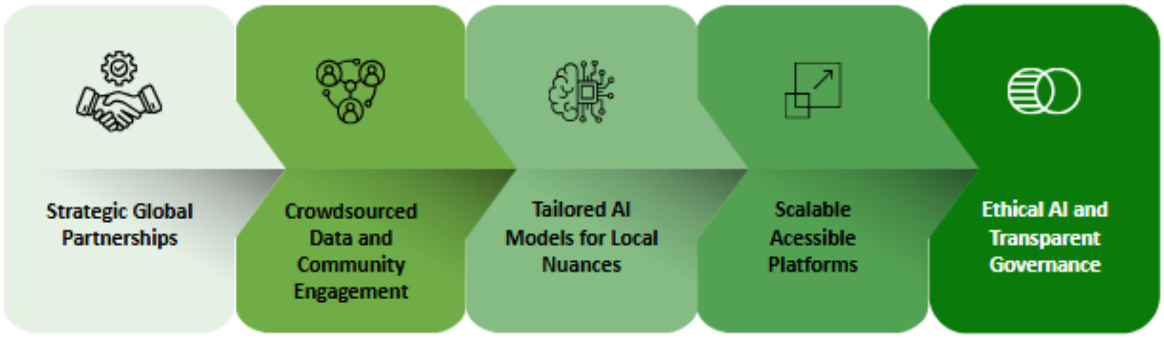

Figure 13: The Strategic Pathway to Sustainable Impact

The methodologies and pillars presented here are more than a technical blueprint; they are a framework for collaboration built on ethical stewardship and deep cultural respect.

Ultimately, this work is a call to action for digital linguistic equity. It is an invitation for all stakeholders to grow this connected ecosystem together, ensuring Africa's rich linguistic heritage is not merely preserved, but secured as a thriving and integral part of our shared digital future.

Chapter 2: Audio Data Collection and Processing

Effective NLP for African languages relies heavily on quality collection and processing of both textual and acoustic data. Both are equally important but most African languages are spoken more than written in practice. Acoustic (audio/oral) data collection will be of great importance in accelerating the democratization and localization of AI for these digital low-resource settings. For communities under this category, spoken language is the main way of sharing and passing down their history, culture, and identity. Since these languages are often missing from modern technology, collecting, cleaning and preserving quality speech data is key. It protects the cultural heritage and ensures that these languages are included in tools like speech recognition, so they can continue to be used and valued by future generations.

Acoustic data can be grouped into read speech (ie a scenario where people read prepared text) and spontaneous speech (ie a scenario where people speak free either on a topic or during an interview). For either of these categories, textual data is very important. These texts serve as scripts for speakers during the recording process for read speech. They also serve as prompts used in spontaneous speech. Text plays a big role in a wide range of NLP applications such as machine translation, sentiment analysis, information retrieval, etc. This data can be gotten from books, reports, blogs, transcripts etc. In low-resource languages, sources of textual data may include folktales, religious text, community newspapers, idioms, etc. Collecting acoustic or textual data for low-resource languages must rely on community-driven methods if we are to accelerate the transition toward the digital inclusion of these languages. The overarching goal must remain the curation of high-quality, ethically sourced data.

This section provides a comprehensive and replicable methodology for audio data collection and processing. This methodology is designed to build upon and expand the current landscape of data assets, which includes foundational corpora, specialized datasets, and the strategic initiatives that support them.

Step 1: Foundational Readiness and Ethical Grounding

The process begins not with technology, but with people. Building trust, defining the scope, and ensuring ethical participation are paramount for the long-term success and adoption of any language technology. Below are the necessary steps to move successfully through Stage 1:

- Community Engagement and Partnerships: Collaborate with local universities, cultural organizations, and community leaders who can act as trusted intermediaries. These partners are invaluable for navigating cultural norms, building rapport, and recruiting participants. For example, the TangaleNLP project in Nigeria partnered directly with the Tangale Community Development Association, ensuring the project was guided by and accountable to the community it served.
- Ethical Framework and Informed Consent: Create a clear, transparent informed consent process. Participants must understand how their voice data will be used, who will have access, and how their privacy will be protected. Provide consent forms in the participant's native language and explain verbally to ensure full comprehension. Collect no personally identifiable information (PII), and ensure all data is securely stored and anonymized.
- Defining Data Collection Types: The dataset's real-world applicability depends on capturing natural speech. A combination of "read" and "spontaneous" audio is ideal.
    - Read Speech: Participants read from a prepared script. This method is useful for collecting a controlled vocabulary and ensuring specific phonetic coverage.
    - Spontaneous Speech: Participants converse freely on a given topic, prompted by text, images, or videos. This approach is crucial for capturing the natural cadence, vocabulary, and code-mixing of everyday language. This methodology prioritizes spontaneous speech to ensure the final model reflects authentic conversational patterns.

Step 2: Ontology Development and Prompt Design

Eliciting diverse and meaningful spontaneous speech requires a well-designed prompting system. These prompts are conversation triggers to encourage natural responses across different contexts and everyday situations. This is achieved by developing a domain-specific ontology, a structured framework that defines concepts and their relationships, to guide the creation of prompts.

- Ontological Framework: The framework organizes conversations from broad themes to specific triggers (see Figure 1).
    1. Dominant Engagement Point: The central axis or theme of the conversation (e.g., Personal Finance).

2. Themes: Foundational categories that establish context (e.g., Money Management).
3. Sub-Themes: Granular focal points within a theme (e.g., Saving and Spending).
4. Visual and Question-Based Triggers: Strategically curated prompts designed to elicit detailed responses. For example, pair an image of a mobile money transaction with the question, "When has a financial decision you made significantly improved your life?"

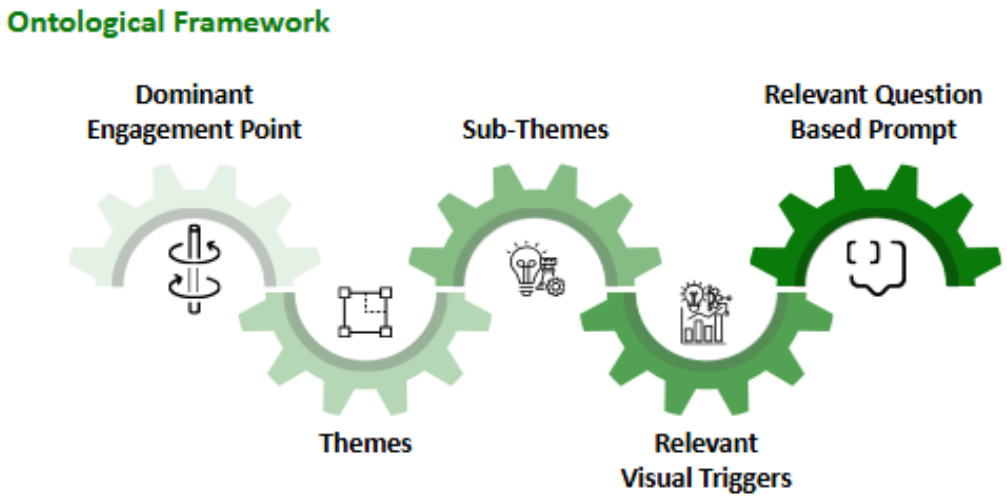

Figure 1: The Domain-Specific Ontological Framework

- Prompting Modalities: To stimulate natural conversation, deliver prompts through various engaging and culturally relevant channels. The key is to present prompts in a common language (like English in the Nigerian context) to avoid pre-biasing the speaker's vocabulary choice in their native language (see Figure 2).
    - Text-Based Ontology: Structured textual prompts designed to stimulate dialogue.
    - Image-Based Ontology: Scene descriptions or contextual images that prompt speakers to articulate their observations.

- Voiceless Video Ontology: Silent videos that require speakers to interpret and narrate the visual story, capturing their descriptive language skills.

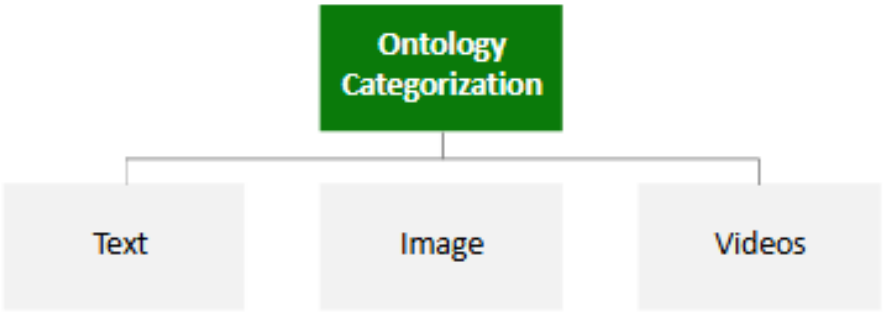

Figure 2: The Domain-Specific Ontological Framework

Step 3: Participant Recruitment and Distribution Logic

The quality of a speech dataset depends on the diversity of its speakers. A model trained on a narrow demographic will fail to perform well in the real world. This requires the following strategic approach to voice sampling:

- Recruit Native Speakers: Prioritize the recruitment of native (L1) speakers, known as Language Consultants (LCs), who possess an authentic command of the target language.
- Ensure Demographic Coverage: Actively recruit a diverse pool of participants, balancing for the following demographic elements:
    - Age Distribution: Sample across a wide range of age groups, with consideration for the population's median age.
    - Gender Distribution: Ensure a gender balance to create an unbiased dataset.
    - Educational and Geographic Diversity: Include participants from various educational backgrounds and geographic locations (urban vs. rural).
- Manage Dialectal Variation: In collaboration with linguistic experts, define a strategy for handling dialects. For initial model training, it is often effective to focus on a primary or

"central" dialect to maintain accent consistency and clarity. You can collect data from other dialects later to enrich the dataset.

Step 4: Data Collection and Technical Quality Assurance

This section outlines the systematic process of gathering both textual and recording resources necessary for building high quality audio datasets. It includes textual data collection where scripts and texts from various sources are prepared for recordings, and acoustic data collection which may require both read and spontaneous speech. In the case of spontaneous speech, speakers get prompts on different sectors of life and can respond freely in their own words without a fixed script. In addition, this section also addresses the technical setup and conditions for effective recording sessions during acoustic data collection.

Textual Data Collection

Textual data is the digital representation of human language. It's the ingredient that powers recent Natural Language Processing (NLP) systems. In any NLP project, the success of downstream tasks like sentiment analysis, translation, or event extraction depends on how well this data reflects real-world language use.

The first step in building a text corpus is strategic data sourcing. This involves drawing information from multiple domains to capture language variety e.g news, social media, education, entertainment, and conversation. For example:

- Yankari Yorùbá Corpus gathered over 51,000 documents from 13 trusted sources including news outlets, blogs, educational websites, Wikipedia etc.
- NaijaSenti gathered tweets across Yorùbá, Hausa, Igbo, and Pidgin.
- MENYO-20k combined news, TED talks, proverbs, and book translations to create an English–Yorùbá parallel dataset.

While collating this data, always aim for transparency in source documentation and permissions. The next phase is data cleaning because this collated text rarely comes clean. A standard text cleaning pipeline includes:

- Text Parsing: Removing unwanted elements such as hashtags, emojis etc.
- Encoding Normalization: Conversion of all text to UTF-8 for consistency.
- Deduplication: Removal of duplicate text.

- Filtering: Excluding very short, non-linguistic, or non-target-language content.
- Normalization: Handling of orthographic inconsistencies and diacritics, especially in tonal languages like Yorùbá.

Recording Equipment and Environments

The choice of equipment used for recording is key to collecting quality data. Noise-reducing devices and high-fidelity microphones (e.g Neumann KM184, AT2020, or equivalent condenser microphone) are essential tools for capturing precise tones and highlighting subtle word-structure details.

After selecting the preferred instrument, the environment where the recordings will be done is also of great importance. Audios need to be recorded in a controlled space that limits echoes and background noise. This controlled environment could be a studio or a specified quiet location.

Recording is done in two phases: first, a small-scale pilot is carried out after the preferred instruments and environment have been selected. Once these have been configured correctly, full-scale recording may proceed.

During the small-scale pilot, test recordings let stakeholders test the setup, make necessary adjustments, and see how well the equipment captures the language's tones and word structures. After this, a complete review of the pilot recording should be carried out to ensure the clarity, consistency, and overall fidelity of the voice samples. Evaluators should identify and correct imperfections, e.g., stray sounds in the background or technical problems. Once the pilot recording and analysis is completed, the team can proceed to large-scale recording.

At the stage of large-scale recording, all selected speakers are recorded following the review from the pilot recording. Recording schedules are set and each participant is briefed on the process to maintain uniform speaking styles, pacing, and microphone distance. During this phase, the team captures the complete list of target materials such as sentences, narratives, or conversational prompts. Technicians monitor audio levels continuously to prevent clipping, distortion, or environmental interference. Once all recordings are done, the dataset is organized and prepared for processing and cleaning.

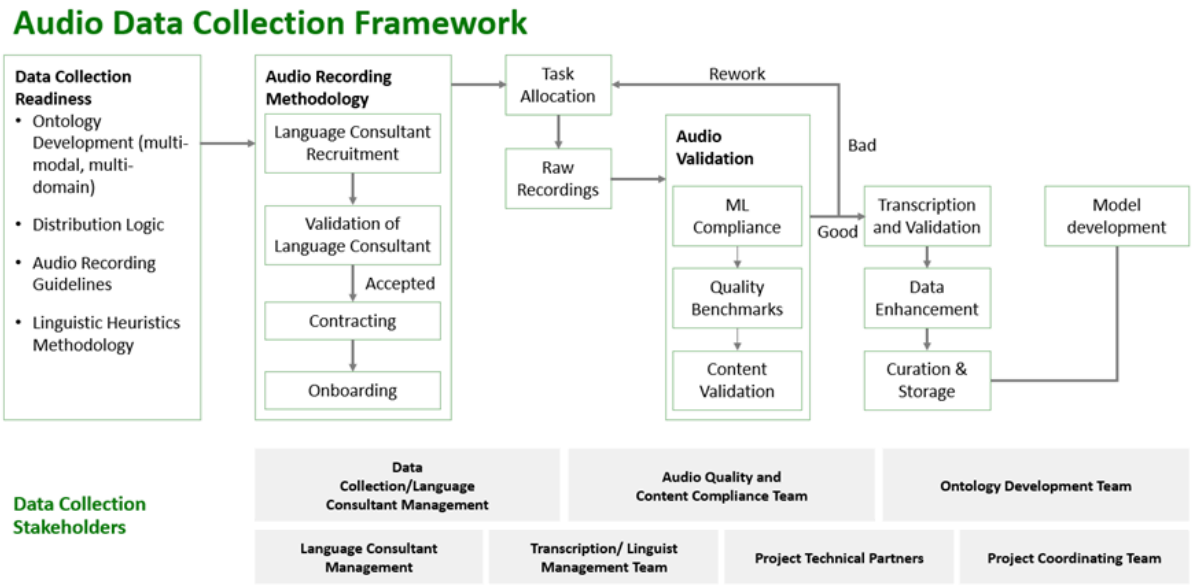

Figure 3: The Audio Data Collection Framework

Step 5: Data Processing and Validation

After collecting the audio data, it must be refined to ensure that it meets high standards of quality and usability. This section covers several key steps: standardization and correction of speech data to fix inconsistencies, file formats and storage to ensure compatibility, annotation and transcription for proper audio-text alignment, noise reduction to improve the clarity of the audio, and filtering low-quality recordings to remove unusable data.

Standardization and Correction of Speech Data

Immediately after the large-scale recordings, standardization should follow. It includes addressing discrepancies in pronunciation and articulation. A major step in this process is the correction of unclear or mispronounced words. Many times, speakers unintentionally use slur words, incorrect phonemes, or inconsistent speech patterns. Such instances could either be corrected by editing the audio or re-recording the entire text or sentence.

Some tonal languages may introduce complexity to the data because the meaning of a word can change based on the pitch, stress, or intonation of the person speaking. It is advisable to ensure that individuals participating in the recording sessions maintain the same intonations throughout. Any inconsistencies in tone or intonation should be carefully reviewed and

corrected. Finally, all files should be converted to a common format (typically 16kHz, 24-bit WAV). This format is widely supported and optimized for any audio task.

File Formats and Storage

While preparing the audio data for use, selecting the correct file format is a crucial step. The type of audio file used impacts the quality of recording. Common formats include WAV, AIFF, FLAC, MP3, and AAC etc. Lossless file types like WAV and FLAC are usually advisable to use because of their ability to maintain the quality of the original recording. They keep all of the subtle details of the sound even during file conversion or storage. MP3 file types are used where storage capacity is more important than overall sound quality. MP3 files are usually smaller than WAV or FLAC (typically 5 to 12 times smaller, depending on the bitrate and number of channels encoded). When file conversion is necessary, an online software called “online-convert (https://audio.online-convert.com/)” can be used. After storing audios in the recommended format, they should be tagged with proper metadata. This metadata includes speaker details, recording conditions, and dialect information.

Annotation and Transcription

After standardizing collected data, annotation is next. Annotation involves adding labels or metadata to each audio file so the AI system can recognize and understand the language correctly. The audio data could be easily annotated with the scripts used for recording. This information serves as the transcription. The transcription process itself is layered to ensure precision (see Figure 4).

- First-Pass Transcription: Trained annotators who are native speakers perform the initial transcription, using tools like ELAN to time-stamp the audio and adhere to established orthographic conventions.
- Tier 1 Validation (Supervisory Review): A supervisory annotator reviews the initial transcription against the audio, correcting errors in spelling, punctuation, or word choice.
- Tier 2 Validation (Linguistic Review): A lead linguist, an expert in the language, performs the final quality check, validating complex linguistic aspects and providing the ultimate sign-off.

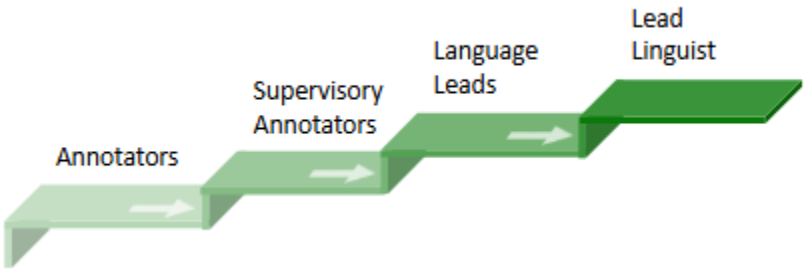

Figure 4: The Multi-Tier Transcription Validation Workflow

For example, during annotation, a Yorùbá sentence such as “Mo máa padà wá lọ́la” is aligned with the exact timestamp where it occurs in the audio (e.g., 0.0–3.8 seconds). Additional labels may be added to capture contextual information, such as dialect (e.g., Ọ̀yọ́ Yorùbá), speaker gender, etc.

Example of metadata include;

- Speaker ID: SPK_YOR_07
- Gender: Female
- Age Range: 25–34
- Dialect: Ọ̀yọ́
- Native Speaker: Yes
- Orthographic transcription: “Mo máa padà wá lọ́la.”
- English translation: “I will come back tomorrow.”
- Speech type: Read speech
- File ID: yor_00123
- Duration: 12.5 seconds
- Audio format: WAV

Noise Reduction and Volume Normalization

To ensure high-quality recordings, audio data should be cleaned and standardized following the steps below:

- Reduce background noise to make recordings clear. Use reliable software tools to help with this e.g Pratt (see Appendix B).
- Filter unwanted sounds by removing very low or very high noises that interfere with clarity. Use tools such as Pratt etc.
- Maintain consistent volume across all recordings so that they are easier to process and use. Use tools such as Pratt, online convert, etc.

Handling Low-Quality and Corrupted Recordings

Even after reducing noise and normalising the volume, some audio can still be compromised. In many cases, these artefacts stem from the use of low-quality recording equipment or unsuitable recording environments, where issues such as overlapping speech, inaudible words, or intrusive background noise (e.g., traffic, multiple speakers, or mechanical hums) are more likely to occur. These elements can distort the intended signal and also confuse AI models that might be used afterwards. Flagging and removing these compromised recordings ensure that the dataset maintains quality.

Synchronization with Transcriptions

After improving the quality of the data, alignment should be done so that every spoken word in the audio is matched with its corresponding transcription. This involves creating a time-aligned dataset. With time-stamped data, the relationship between audio and text becomes clearer and easier to analyze or manipulate.

This alignment could be achieved with commonly employed tools such as the Montreal Forced Aligner (https://montreal-forced-aligner.readthedocs.io/en/latest/), ELAN (https://archive.mpi.nl/tla/elan/download), and Praat (www.praat.org). These tools use acoustic models and phonetic dictionaries to automatically map audio segments to written transcripts. But this feature isn't available for many low resource languages like Yorùbá, Hausa, etc. This means that forced alignment has to be done manually for these languages. Trained reviewers are essential for validating and correcting any misalignments or transcription errors that occurred during both manual and automated processing.

Expert Validation and Dataset Review

Expert validation is the final stage in preparing a high-quality acoustic dataset. At this point, the focus shifts from cleaning and alignment to human-led review processes. In this step, all prior processing such as noise reduction, segmentation, transcription alignment, and standardization are verified to ensure the preservation of the original audio content. Listening tests should be performed by native speakers to assess the quality of the pronunciation and intonation. Their evaluations help identify issues that may have been overlooked, such as accent inconsistencies, or mispronounced words. This ensures that the dataset represents the natural language of focus to the highest possible level.

Before the dataset is finalized, released, and used by the public or AI models, it should undergo a comprehensive review to confirm its completeness, consistency, and usability. This review ensures that all the audio files conform to the required format, metadata is accurate, and all corrections and alignments are properly integrated.

Chapter 3: Scale and Ecosystem Sustainability

Introduction and Methodology

The digitisation of African languages is gaining momentum, propelled by advances in Natural Language Processing (NLP) and Artificial Intelligence (AI) (Adebara, 2024). Yet many initiatives remain fragmented and are focusing on single languages or regions, as a result leaving significant gaps in representation and reinforcing the digital divide between African communities and the Global North (Ahmed, 2007). This divide is not only technical but also socio-economic, limiting equitable access to tools, data, and participation. To bridge it, African NLP requires an inclusive and sustainable ecosystem built as digital public infrastructure: one that empowers communities to digitise their languages on their own terms. This chapter in the playbook addresses that task through a combined methodology: a literature review and a continent-wide practitioner survey. The literature review draws on benchmark projects such as MasakhaNER, MasakhaNEWS, AfriSenti, LAFAND-MT, and IrokoBench, which illustrate both progress and persistent gaps in coverage, participation, and licensing. Complementing this, an online survey of 90 practitioners spans; developers, linguists, educators, community advocates, and policymakers across diverse African regions. The online survey was administered through research networks and digital platforms. The survey aimed to map expertise, highlight successes and barriers, and surface perspectives on scaling African NLP. This playbook is guided by a Theory of Change (Weiss, 1995; Gaventa & Barrett, 2010), it argues that scaling digitisation requires more than data growth: it demands participatory, community-centred methodologies, robust taxonomies, and accessible tools that reduce technical barriers. By integrating literature insights with practitioner experiences, the analysis sets out actionable strategies and best practices for building sustainable African language infrastructures.

To contextualise the survey findings, it is important to note the geographical distribution of respondents. As shown in Figure 1 below, participation was strongly concentrated in Southern Africa, with South Africa accounting for nearly half of all responses (45.6%). Nigeria (8.9%), Ethiopia (7.8%), Zimbabwe (5.6%), and Uganda (5.6%) also featured prominently. Smaller but significant contributions came from countries across West, East, and Central Africa, as well as isolated responses from the diaspora (e.g., Switzerland and the United States). This spread highlights both the regional strengths of African NLP activity and the gaps where engagement remains limited, reinforcing the need for intentional inclusion of underrepresented regions.

Detailed Country Breakdown		
Country	**Respondents**	**Percentage**
South Africa	41	45.6%
Nigeria	8	8.9%
Ethiopia	7	7.8%
Zimbabwe	5	5.6%
Uganda	5	5.6%
Kenya	4	4.4%
Ghana	4	4.4%
Eswatini	3	3.3%
Tanzania	2	2.2%
Cameroon	2	2.2%
Namibia	1	1.1%
Mauritania	1	1.1%
Lesotho	1	1.1%
Comoros	1	1.1%
Madagascar	1	1.1%
Benin	1	1.1%
Gambia	1	1.1%
Switzerland	1	1.1%
United States	1	1.1%

Table 1: Detailed Country Breakdown of Survey Respondents. The percentage distribution of participants across countries, highlighting strong representation from South Africa, Nigeria, and Ethiopia, as well as smaller contributions from other regions and the diaspora.

To structure this analysis, the chapter proceeds as follows. It begins by examining the human infrastructure of African NLP, mapping the diverse roles and expertise that sustain the ecosystem. It then turns to language coverage and distribution, highlighting both representation and persistent gaps across regions. The next sections focus on capacity building and success, unpacking how roles, resources, and access shape participation and outcomes. This is followed by an exploration of language resources and tools, patterns of sharing, and the ethical challenges surrounding accessibility, archiving, and licensing. The chapter also considers barriers to development and the varied methods of community engagement, reflecting on their ethical implications. A literature scan consolidates lessons from benchmark projects to ground the discussion in evidence. Finally, the chapter concludes with reflections and best practices for building inclusive and sustainable African language infrastructures.

Human Infrastructure and Role Mapping in African NLP

Infrastructure in African NLP is not only technical but also human, built on the skills and commitments of researchers, developers, linguists, educators, advocates, policymakers, translators, librarians, and speech-language therapists. These actors collectively sustain the ecosystem by creating resources, raising awareness, and connecting communities with technology. Figure 2, based on survey data from 90 practitioners, shows how these roles are distributed, highlighting both strengths and gaps in expertise.

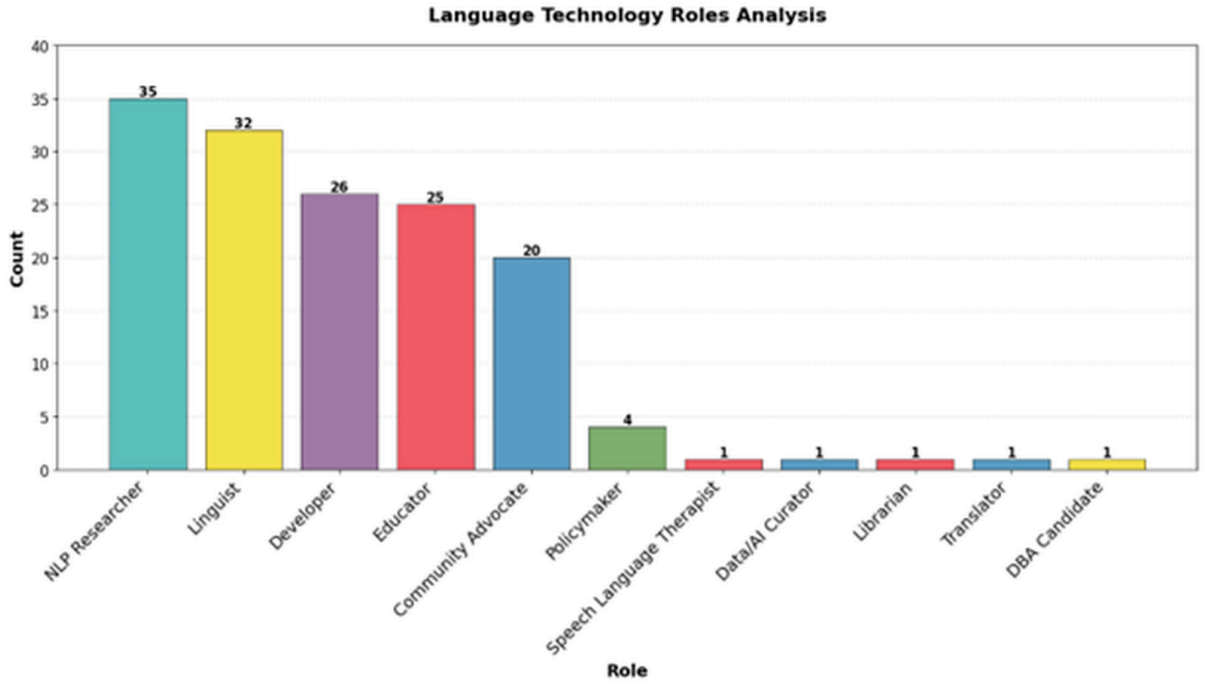

Figure 2: Role Distribution Among African NLP Practitioners. Diversity of roles reported by survey respondents, including NLP researchers, linguists, developers, educators, community advocates, policymakers, and graduate students. It highlights the broadness of expertise within the ecosystem while also pointing to groups that may be less visible or not fully captured by the survey's reach, such as indigenous community members.

Survey results reveal that the majority of respondents identified as NLP researchers (23.8%) and linguists (21.8%), with smaller proportions of developers, educators, and community advocates. Community advocates (13.6%) and policymakers were notably underrepresented in the responses, yet this does not necessarily reflect their absence in the wider ecosystem. The survey's reach shapes who participated, and numerical counts of roles cannot fully capture the balance of expertise in practice, as developers and researchers are often more visible simply due to institutional affiliation and access. What the findings do demonstrate is the breadth of actors involved, from technical experts to educators, translators, cultural custodians, and activists. Educators in particular play a distinctive role: they serve not only as instructors of computational or linguistic methods but also as mediators who contextualise language technologies for students and communities, bridging academic curricula with lived social practice. Likewise, DBA (Doctor of Business Administration) candidates who participated bring insights from organisational management and policy innovation, underscoring the value of cross-disciplinary expertise. At the same time, we must ask who is missing from this picture,

especially indigenous community members and cultural custodians whose lived expertise is central to language sustainability but may not have been reached through this survey. Supporting all these actors to leverage their knowledge and contribute meaningfully in participatory ways is essential for building a truly inclusive African NLP ecosystem.

Language Coverage and Distribution in African NLP

Language coverage, the extent to which individual languages appear in research, is a key indicator of representation and equity in African NLP. It reflects policies, funding, colonial histories, and institutional priorities. Survey responses from 90 practitioners reveal both diversity and imbalance in African NLP. While the findings point to wide ambitions including cross-border and multilingual work, they also expose persistent gaps, especially for minority, oral, and underrepresented languages in Central and North Africa. Mapping this coverage is therefore strategic: it shows not only what exists but also who is excluded, emphasising the need for intentional investment to ensure no language or community is left behind. Figure 3 presents an overall word cloud of all responses, Figure 4 isolates languages reported by non-technical practitioners (linguists, educators, librarians, cultural custodians), and Figure 5 highlights those targeted by technical practitioners (developers, technologists, NLP researchers). Together, they reveal overlaps and divergences in focus, showing how practitioner roles shape which languages gain visibility in African NLP.

Figure 3: African languages word cloud by both technical and non-technical practitioners. Word cloud visualising the range of languages practitioners are working on both technical and non-technical practitioners showing dominant languages such as isiZulu, Swahili, Yoruba, and Hausa, alongside less-documented ones, highlighting diversity as well as underrepresentation.

Figure 4: African languages word cloud by non-technical practitioners. Word cloud illustrating the languages that non-technical contributors such as language practitioners, librarians, and cultural workers engage with in African NLP, highlighting both dominant and less-documented languages.

Figure 5: African languages wordcloud by technical practitioners. Word cloud representing the languages most frequently worked on by technical contributors such as developers and researchers showing concentration around isiZulu, Kiswahili, isiXhosa, and isiNdebele, alongside a wide spread of less-represented languages.

Recommendations for Building Equitable Capacity

To ensure that African NLP grows in ways that are just and inclusive, we recommend the following:

- Democratise technical training: Create accessible, multilingual, role-sensitive training materials for linguists, educators, advocates, and developers.
- Bridge disciplines: Support interdisciplinary teams where developers learn from linguists, and community members co-design alongside technologists.
- Mentorship and peer networks: Establish cross-regional mentorship programmes, especially for underrepresented roles and geographies.
- Tool accessibility: Develop NLP tools and platforms that are intuitive, open-source, and designed with non-specialist users in mind.

- Measure inclusion: Progress should not only be judged by the number of datasets or languages covered. What matters just as much is how the data was created and who was involved. Suggested questions to measure inclusion are:
    - What proportion of participants co-authored the work?
    - Were under-documented languages represented and how many?
    - Was the process fair, transparent, and respectful of local knowledge?

These recommendations directly respond to the inequities outlined above. They move beyond recognising disparities to proposing actionable steps for redistribution of skills, resources, and opportunities. By embedding inclusivity into training, design, and evaluation, the ecosystem can ensure that no one is excluded from the digital future of their own language.

Language Access Shapes Success and Ethical Participation in African NLP

In any effort to digitise African languages, we must look at which languages are being worked and who is involved, who is succeeding, and who is excluded from success due to structural or linguistic barriers. This section highlights how roles, resources, and language access shape people's ability to participate fully and meaningfully in African NLP projects. The goal is to ensure that the digital public infrastructure we are building is not just technically sound, but socially and ethically inclusive, offering pathways for different kinds of contributors to succeed on their own terms. Survey responses to the question, "What have been your successes?", revealed that success is not evenly distributed across the African NLP ecosystem. Respondents who reported measurable progress or outcomes tended to be those with access to technical skills, institutional resources, or computational tools often identifying as developers, researchers, or members of university-affiliated teams.

Their successes included:

- Creating working translation or classification models,
- Improving alignment between machine outputs and linguistic nuance,
- Building tools that engage speakers in learning or vocabulary exploration.

Most African NLP practitioners are open to sharing resources, showing strong collaboration despite limited tools. Where hesitation exists, it reflects caution such as incomplete work or licensing concerns rather than refusal. Sustaining this culture requires mentorship, ethical guidance, and structures that value in-progress contributions. Another pattern emerged in responses: several participants left questions blank or struggled to explain their successes and challenges clearly. While this silence may have multiple causes, one probable factor is language. The survey was conducted in English, and the questions included terms like machine translation, annotation, and small language models phrases that are not always intuitive, especially outside formal technical or academic environments. This made it harder for participants, particularly those from community or educator roles, to express themselves with confidence or clarity. This raises an ethical concern: if people are excluded from full participation because of linguistic barriers, then the process itself is not truly inclusive. African NLP cannot be a space where only the English-proficient or tech-savvy are able to contribute meaningfully. If you want participation from diverse contributors, you must make language itself accessible. This includes how you design forms, explain workflows, and document tools. Clarity is not just a communication tool; it's a condition for inclusion.

Best Practices:

- Write in plain language with examples that make sense to local users.
- Translate key materials into widely spoken African languages.
- Allow responses in multiple languages or formats (e.g., voice notes, local scripts).
- Value oral storytelling and contextual explanation as valid forms of communication

Designing for Role Diversity and Capacity Building

The varied definitions and expressions of success across roles point to a deeper issue: African NLP needs to create multiple entry points for people to join, contribute, and grow. If only developers are able to build, test, and measure progress, then the ecosystem will remain limited. To scale sustainably, we must equip linguists, educators, and community actors with tools and support that match their context and strengths.

Best Practices:

- Build interdisciplinary teams where linguists and communities take a central role, shaping workflows and approaches in ways that reflect African contexts.

- Create modular training resources: short tutorials, mobile-accessible guides, or offline toolkits.
- Offer workshops or mentoring tailored to non-developers.
- Document processes in ways that others can replicate without needing advanced coding skills.

In Practice: Moving Toward Inclusive Success

African NLP will only succeed if it reflects the full complexity and capacity of its contributors. Success must be understood as both technical achievement and social impact. Projects should be evaluated not just by what models they build, but by how they build them, and who gets to participate in that process. A playbook for digitising indigenous languages must offer concrete, practical strategies for inclusion starting with language itself. Accessibility, clarity, and cultural responsiveness are not luxuries. They are foundational design choices that shape whether a language project is just or extractive, empowering or exclusionary.

As you build, train, and deploy, ask:

- Who can participate here?
- Can they describe what they're doing?
- And can they succeed on their own terms?

If the answer is yes, you're helping build a truly inclusive digital infrastructure for African languages.

Understanding Expression in a Multilingual Context

As we deepen our analysis of the African NLP ecosystem, it becomes increasingly clear that language itself, including how people express themselves and which language they use, plays a central role in shaping participation. Many contributors to this ecosystem are not first-language English speakers, yet they are often asked to describe complex work, challenges, and aspirations in English. This raises important questions about how multilingual expression shapes participation and whose voices are more easily heard within the ecosystem.

Expression and Complexity: A Cry to Be Heard

Some survey responses were long, indirect, or emotionally layered. They may, at first glance, appear difficult to analyse. But this "messiness" is not a flaw. It is a reflection of people trying to make meaning through a language that does not fully hold their thoughts.

"It's not because they want to sound more knowledgeable… it's merely a cry to be heard."

This insight, shared during project discussions, captures the emotional and linguistic labour required to participate in an ecosystem that does not operate in your home language. The responses show not confusion but effort in a willingness to engage, to be seen, and to contribute despite the limitations of dominant language norms. As a result, we must resist the instinct to reduce, correct, or overinterpret what people say. Instead, we must build NLP systems and research environments that make room for non-standard expression, ambiguity, and cultural variation.

Visualising the Challenge Landscape

To further understand how language intersects with difficulty, the following bar chart was generated from open-ended responses to the question, “What have been your biggest challenges in working with African languages and NLP?”

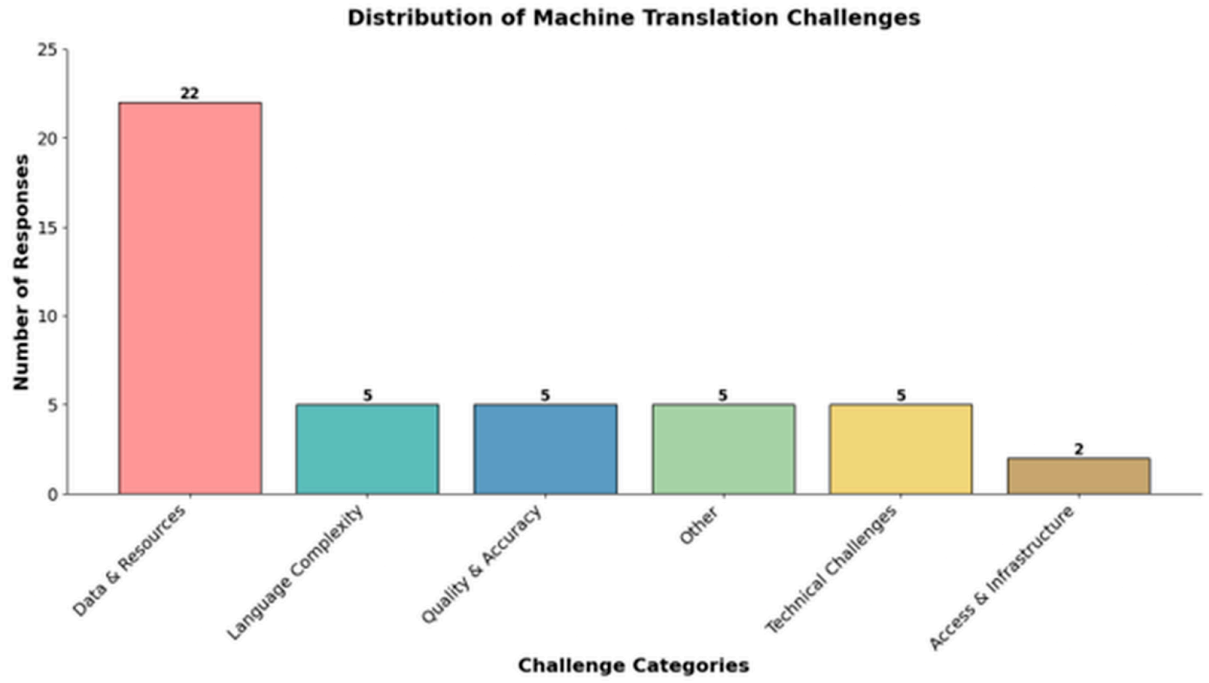

Figure 6: Common Challenges Reported in African NLP Work. Bar graph illustrating the most frequently mentioned issues. Words like data, language, lack, translation, and challenge dominate. These are not just technical terms, they reflect structural inequalities in access, training, and representation.

Success in African NLP: Roles, Outcomes, and What It Takes

While much of this chapter has focused on challenges and structural barriers, it is equally important to reflect on what success looks like across the African NLP landscape and who is achieving it. Success is not uniform. It takes many shapes depending on role, access, and position within the ecosystem. This section unpacks survey responses to the question, “What have been your biggest successes?”, and explores how different actors define and experience impact.

Patterns of Success by Category

Responses were grouped into seven high-level categories:

- Model Development (27.8%)
- Data & Resources (22.2%)
- Other (22.2%)
- Performance Improvement (11.1%)
- Platform & Applications (5.6%)
- Community & Collaboration (5.6%)
- Research & Academic (5.6%)

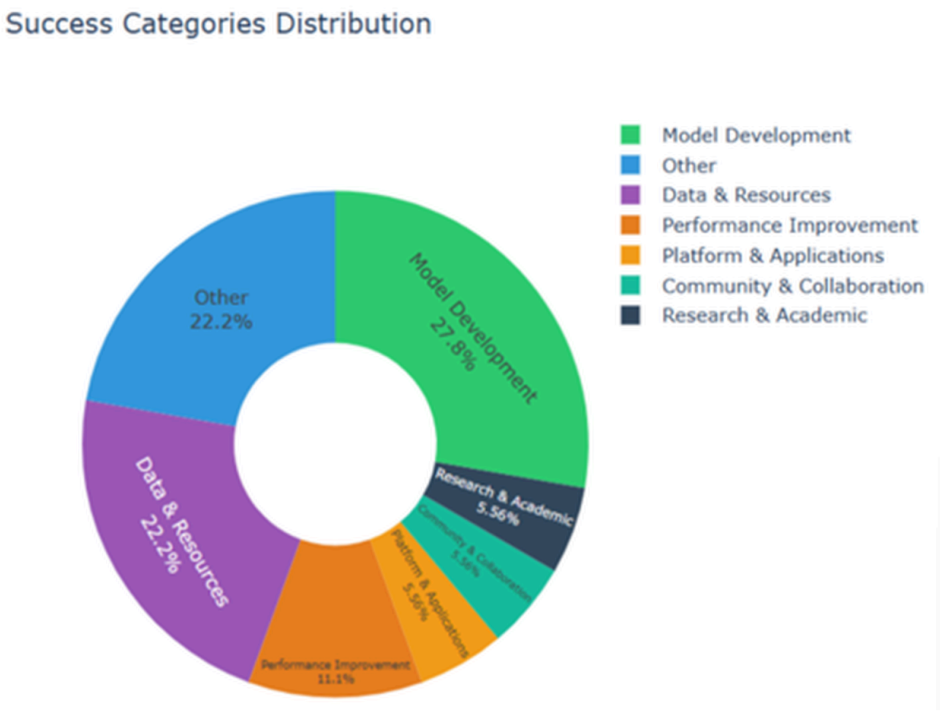

Figure 7: Distribution of Reported Successes in the African NLP Ecosystem
The visualisation above shows a strong concentration of success in technical areas: nearly 50% of achievements fall

under model development or resource creation. These reflect successes typically reported by developers, data scientists, or NLP researchers who have access to computing power, institutional support, and collaborative tools.

These successes include:

- Building or fine-tuning machine translation models,
- Publishing speech-to-text systems or small language models (SLMs),
- Contributing to datasets through platforms like Common Voice or Masakhane.

These are high-visibility outcomes that signal ecosystem maturity—but they are only one part of the story.

Whose Success Counts? The Role of Access and Visibility

Survey responses from linguists, educators, and community advocates told a different story.

Their definitions of success were more relational or structural. Examples included:

- Gaining institutional recognition for local languages,
- Introducing linguistic nuance into broader datasets,
- Advocating for the inclusion of tonal variation or dialects,
- Organizing awareness campaigns or training workshops.

These achievements are harder to quantify but no less essential. In fact, they represent what might be called the soft infrastructure of NLP: the work of translation, mediation, cultural framing, and pedagogy that underpins the technical layer. Yet these forms of success are rarely rewarded in traditional funding or evaluation models.

Role-Specific Success: Why It Matters for Scaling

If we fail to differentiate how success manifests across roles, we risk two outcomes:

- Invisible labour: The contributions of non-technical actors go unrecognized or underfunded.
- Stalled scale: Efforts to grow the ecosystem focus narrowly on coding or model building, leaving gaps in awareness, uptake, and community engagement.

Recognizing this, the playbook recommends supporting role-specific markers of success, such as:

- For developers: model performance, deployment reach, code reuse.
- For linguists: dialect coverage, descriptive insights, validated resources.
- For educators: curriculum design, student engagement, public communication.
- For community advocates: network-building, local ownership, multilingual outreach.

Success cannot be reduced to benchmarks, alone it must reflect the ecosystem ecology: the interplay of knowledge, trust, tools, and community.

Playbook Guidance: Redefining Success for Inclusive Growth

To build a scalable and sustainable ecosystem, stakeholders should:

- Develop layered success frameworks that accommodate both technical and cultural wins;
- Design mentorship programs that match people across roles not across skill levels;
- Include indicators of impact, not just output, in project evaluations;
- Make room for invisible labour ,the work of listening, translating, caring, and connecting as part of what sustains the ecosystem.

Language Resources and Tools: Patterns of Use, Access, and Visibility

The tools and datasets that African NLP practitioners reach for are more than just instruments—they are reflections of what is accessible, visible, and supported within the ecosystem. In response to the question, “What language resources or tools have you used or created?”, participants listed a variety of assets, ranging from globally recognised platforms to custom-built or unpublished local datasets. These responses were cleaned and consolidated to reduce naming overlaps and visualized to highlight usage frequency.

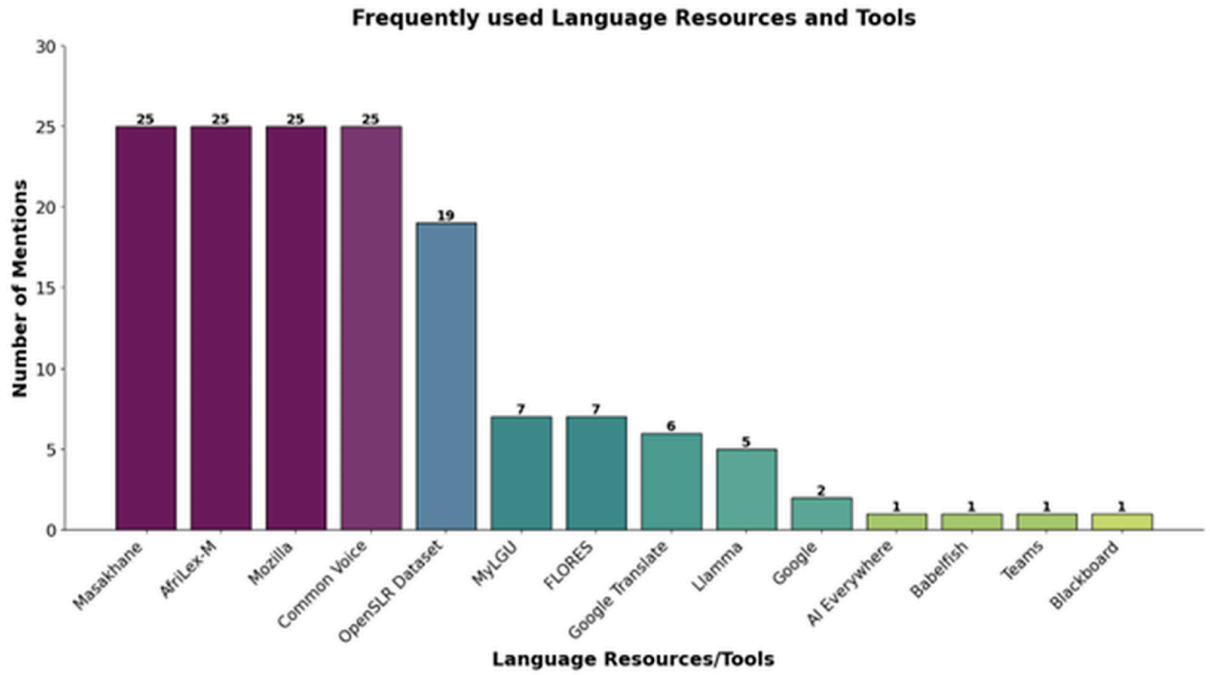

Figure 8: Frequency of Mentioned Language Resources and Tools

A few tools dominated responses, with Masakhane, AfriLex-M, Mozilla, and Common Voice cited most frequently, reflecting their accessibility, documentation, and community support. Beyond these, the landscape fragmented, with many resources mentioned only once often due to limited visibility, documentation gaps, or projects still in development. Importantly, “not yet” sharing reflected care and readiness rather than unwillingness. The prominence of widely used tools highlights that community engagement and ease of access are as crucial as technical sophistication. At the same time, the reliance on tools like Google Translate underscores ongoing issues of accessibility, connectivity, and familiarity, pointing to the need for open, locally rooted, and usable resources.

Recommendations for Sustainable Resource Ecosystems

To support ethical, inclusive, and sustainable use of language tools, we propose the following best practices:

- Keep documentation simple, clear, and visible.

- Use consistent names for tools and resources.
- Translate key materials into African languages and plain English.
- Value and share work-in-progress contributions.
- Ensure ethical, respectful processes when publishing community data.

Together, these practices emphasise not just creating more resources, but ensuring they are accessible, equitable, and trustworthy within the African NLP ecosystem.

Willingness to Share: Trust, Readiness, and Ecosystem Ethics

The survey revealed that most African NLP practitioners are willing to share the resources they have created, reflecting a strong culture of collaboration and solidarity in a context where tools and datasets remain scarce. This openness demonstrates not only generosity but also a shared responsibility to strengthen African languages in digital spaces. However, a smaller group expressed hesitation, often due to valid concerns such as incomplete or developing work, team-based ownership, licensing restrictions, or a lack of confidence. These responses suggest that a “not yet” should be read as caution and care rather than refusal, pointing to the need for supportive structures, mentorship, and ethical guidance. Building a sustainable sharing culture therefore requires encouraging contributions when practitioners feel ready, valuing partial or in-progress work, and creating safe, respectful processes for licensing and hosting. What matters most is not only openness but also the conditions that make sharing inclusive, ethical, and empowering for all contributors.

Barriers to Development: Building Capacity Across the Ecosystem

One of the most consistent themes in the survey was the presence of barriers that slow or complicate efforts to develop African language datasets and tools. Respondents were asked to describe what challenges they face in accessing or creating language resources. Their answers revealed more than technical difficulties—they pointed to deeper issues of access, role identity, and uneven support structures across the ecosystem.

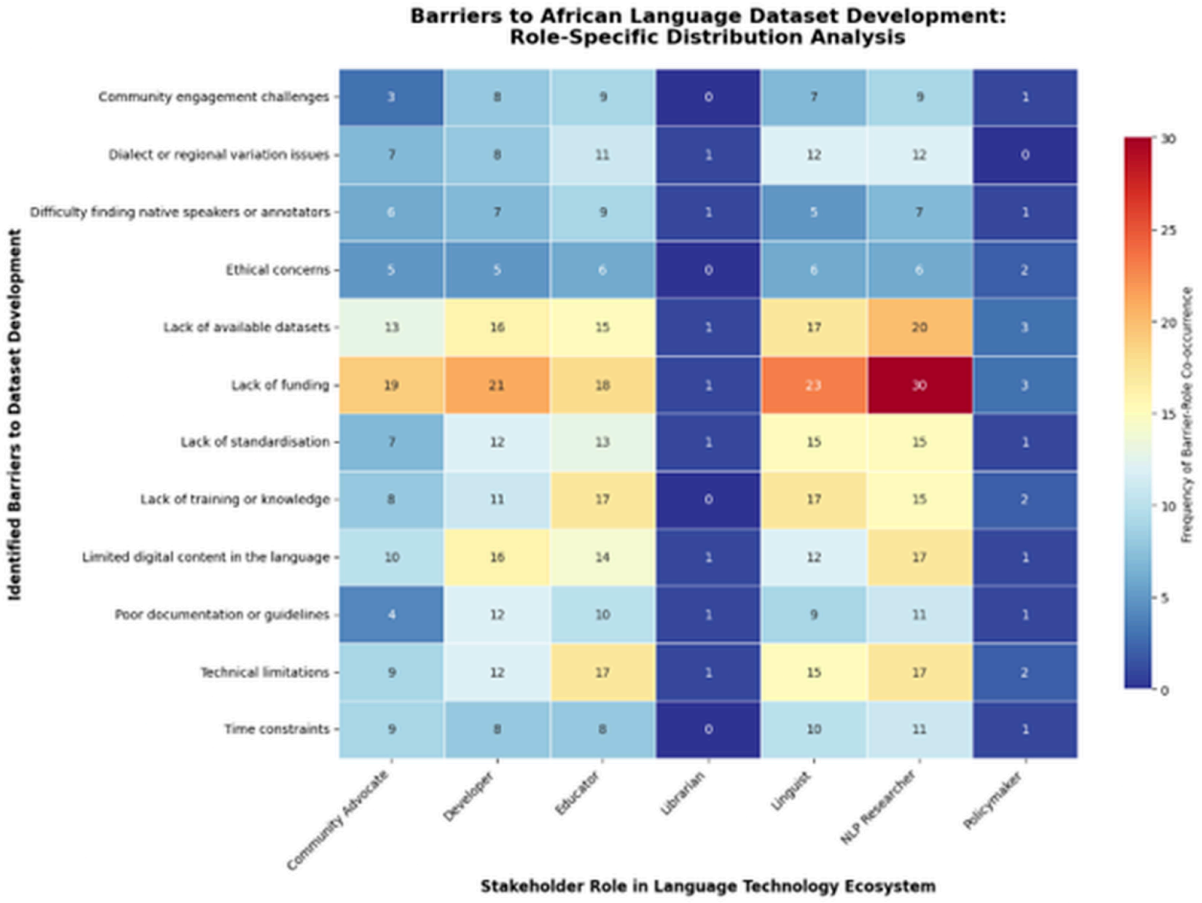

Figure 6: Illustration of the distribution of barriers across different roles in the ecosystem

The visualization above shows how these barriers are distributed across different roles in the ecosystem. It highlights a recurring pattern: linguists, educators, and community advocates report the most persistent challenges, particularly in accessing infrastructure, community support, or training. While developers and NLP researchers certainly face difficulties, their access to tools and networks often gives them a slight advantage. This suggests that capacity in African NLP is not just about what tools are available but also about who is equipped to use them.

Types of Barriers

From the qualitative responses, three main categories of barriers emerged:

- Technical barriers were among the most frequently reported. These include the absence of computing resources (such as GPUs), limited internet access, poor documentation, or

the complexity of using certain NLP frameworks. For many, even the basic act of annotating or training a model becomes a logistical challenge.

- Linguistic barriers relate to the nature of the languages themselves. Respondents working with tonal languages, those without standard orthographies, or multiple dialects often struggle to establish baseline data that meets the needs of NLP models. This is especially difficult for under-documented languages that exist mainly in oral form.
- Community barriers involve issues of trust, ethical concerns, or lack of access to native speakers. Some respondents described hesitation in collecting data without formal protocols, or fatigue within communities that have been asked to contribute with little feedback or recognition.

Role-Based Capacity and Imbalance

The impact of these barriers is not uniform across roles. Survey responses suggest that linguists and community actors face a greater share of challenges not because they lack expertise, but because they are often cut off from institutional support or digital infrastructure. By contrast, developers and NLP researchers, especially those working in academic or research institutions, tend to have greater access to the tools, mentorship, and datasets necessary to work through technical challenges. Their successes, while earned, are also the result of being more connected to global NLP ecosystems. This reveals a broader structural issue: the tools, resources, and knowledge required to digitise African languages are not equally distributed. Without intentional support, some communities and contributors may remain stuck behind barriers that others barely notice.

Best Practices for Strengthening the Ecosystem

- Collaborate, not isolate: Pair linguists and developers to foster shared growth.
- Train across disciplines: Use co-teaching and peer-led tutorials to bridge skills.
- Lower barriers: Provide open-source, no-code tools with clear documentation.
- Value in-progress work: Encourage sharing even before projects are complete.
- Use data for support: Map barriers to guide targeted funding and mentorship.

A resilient African NLP ecosystem must be structured not just around technology, but around care. Success should not depend on who has the most access, but on who is supported to grow.

Licensing: Protecting, Not Controlling

Licensing determines how a dataset can be used, adapted, or cited. It communicates the creator's intent and protects the rights of both contributors and communities. Yet in our survey, licensing practices were inconsistent. Some resources had no license at all; others were shared informally. This ambiguity undermines trust and discourages ethical reuse. Commonly used licenses like Creative Commons Attribution (CC-BY) and CC-BY-NC (Non-Commercial) offer important protections but may not fully account for the power asymmetries and communal rights present in African NLP. As highlighted in the Masakhane and LAFAND-MT studies, standard open licenses risk enabling external actors (e.g., well-resourced tech companies) to profit disproportionately from African datasets often without benefit flowing back to the source communities. To address this, new frameworks like the Nwulite Obodo Open Data License (NOODL) have emerged. NOODL supports equitable data sharing, allowing dataset creators to indicate desired benefits in-kind (e.g., attribution, community support, or downstream access). It provides a nuanced structure that balances openness with community-defined safeguards because a dataset is not just a file. It's a record of language, culture, and community. Ethical archiving and licensing ensure that this record is treated with the respect it deserves. In addition to NOODL, emerging licenses like the Esethu Framework promote sustainable dataset governance and equitable benefit-sharing. The Esethu Framework focuses on community-led curation and licensing tailored for low-resource African languages, safeguarding local creators' interests while addressing gaps in language technology.In addition to NOODL, emerging licenses like the Esethu Framework promote sustainable dataset governance and equitable benefit-sharing. The Esethu Framework focuses on community-led curation and licensing tailored for low-resource African languages, safeguarding local creators' interests while addressing gaps in language technology.

Visual Aid: Comparing Licensing Approaches

To support dataset creators in making informed choices, we offer the following comparison:

License Type	Strengths	Limitations	Best Used When…
CC-BY	Broad reusability with attribution	Risk of commercial exploitation	Wide adoption and citation are priorities
CC-BY-NC	Prevents commercial misuse	Still allows institutional overreach	Sharing with academic and community users
NOODL	Context-aware, supports benefit-in-kind	Requires community guidance and effort	Emphasis on equity and local empowerment

Ethical Archiving and Licensing Checklist

- Use stable, public repositories - e.g., Zenodo, Masakhane GitHub, or university archives
- Include full metadata and contextual documentation - contributors, purpose, source agreements, annotation method
- Select licenses based on community values - e.g., CC-BY, NOODL; clarify what kind of use is permitted
- Establish consent and feedback mechanisms - contributors should understand how their data will be used
- Avoid extractive practices - do not publish without meaningful community engagement

Reflections from Literature and Practice

Findings from the literature reinforce the importance of ethical archiving and licensing:

- MasakhaNER and MasakhaNEWS emphasize participatory approaches where contributors are credited and engaged throughout the process.

- LAFAND-MT documented not only dataset creation but licensing protocols, ensuring contributors were named as co-researchers.
- AfriSenti and IrokoBench extended licensing conversations to multilingual benchmarks and sentiment datasets, addressing the complexity of representing culture and expression.

These initiatives demonstrate that ethical licensing is not about limiting access, it's about defining access with care. It's about ensuring that the people who create and power African language technologies are protected, credited, and empowered.

Literature Scan: Participatory Modeling and Resource Equity in African NLP

This section draws on five foundational studies; LAFAND-MT, MasakhaNER, MasakhaNEWS, AfriSenti, and IrokoBench to establish an evidence base for best practices in African language digitisation. These studies inform many of the recommendations, frameworks, and ethical commitments found throughout this playbook. The reviewed work demonstrates that building infrastructure for African NLP is not simply a technical project; it is a socio-technical, participatory effort that hinges on equitable access, ethical data stewardship, and community-centered design. These studies reflect both the urgency of digital inclusion and the opportunity to preserve linguistic diversity through open, collaborative innovation.

Coverage and Language Diversity

Drawing on five benchmark efforts—LAFAND-MT (MT parallel data), MasakhaNER (NER), MasakhaNEWS (news topic classification), AfriSenti (sentiment), and IrokoBench (LLM evaluation)—over 30 African languages are represented, spanning the Niger-Congo, Afro-Asiatic, Nilo-Saharan, and Creole families. These include widely spoken languages like Swahili, Yoruba, and isiZulu, alongside lesser-studied languages such as Ghomálá', Fon, and Naija (Nigerian Pidgin). They also use diverse scripts—Latin, Arabic, and Ge'ez—and typologies, including tonal, agglutinative, and isolating structures. While these efforts mark real progress, smaller and endangered languages remain notably underrepresented. This signals an ongoing need for targeted support, especially in Central, North, and island nations where digital language activity is limited.

Participatory Methods and Ethical Engagement

The Masakhane community's work stands out for its co-creation ethos. In contrast to extractive approaches, these projects have:

- Paid annotators,
- Recognized contributors as co-authors,
- Used collaborative workshops to define priorities.

This participatory model challenges colonial data dynamics and provides a compelling ethical precedent. Building with communities, not for them but emerges as a critical lesson, echoed throughout this playbook's sections on licensing, consent, and benefit-sharing.

Modeling Approaches and Theory of Change

The reviewed studies demonstrate that even small datasets when ethically collected and community-led can yield strong results. Strong results here refer not only to competitive model performance, but also to the broader value these datasets bring in terms of usability, inclusivity, and cultural relevance. For example:

- LAFAND-MT fine-tuned M2M-100 with just a few thousand examples,
- MasakhaNEWS used few-shot learning to enable fast deployment,
- IrokoBench introduced benchmarks for typologically diverse languages across multiple tasks.

These methods align with a Theory of Change focused on creating digitally inclusive ecosystems where African languages serve learning, commerce, expression, and governance and as cultural and cognitive tools.By demonstrating that even small but community-curated datasets can drive usable translation systems (LAFAND-MT), or that few-shot learning can accelerate deployment across multiple languages (MasakhaNEWS), these studies show that scale is not only about quantity but about inclusivity of process. Similarly, IrokoBench's emphasis on typological diversity illustrates how evaluation frameworks can expand whose languages are recognised in digital spaces. Together, these approaches embody a Theory of Change that sees African NLP not simply as a technical exercise, but as a pathway to cultural preservation, economic participation, and equitable knowledge production.

Lessons for Language Selection and Ecosystem Design

From these studies, key lessons emerge for this playbook:

- Choose languages not only based on data availability, but community readiness and willingness to co-create;
- Foster shared ownership of data through local annotation teams and collaborative licensing;
- Recognize that community-led contributions often yield richer, culturally valid outputs.

These principles underpin many of this playbook's practical guidelines on archiving, resource development, and taxonomy design. This literature scan reinforces the need for community insight, ethical alignment, and equitable participation in African NLP. It offers practical evidence that ethical, inclusive, and low-resource approaches can scale effectively when rooted in community collaboration. As such, it forms a foundational backdrop to the broader recommendations of this playbook.

Concluding Reflections: Building with, Not for

This chapter has explored what it means to build African language technologies at scale, drawing from data, tools, and models as well as people, relationships, and lived realities. This chapter has explored what it means to build African language technologies at scale, drawing on survey findings, practitioner insights, and lived experiences. We mapped uneven access to resources, differences in role-based capacity, and the power of communities to shape meaningful linguistic futures. These reflections return us to the Theory of Change that underpins this playbook: Scale is not simply about producing more datasets or tools, but about cultivating inclusive, sustainable ecosystems where care and participation are central. The best practices outlined throughout the chapter from interdisciplinary teams to ethical archiving, and illustrating a relational view of care, where success is measured not only by technical outputs but by how communities are supported and respected in the process. This resonates strongly with the African philosophy of Ubuntu, which recognises that "a person is a person through other people," reminding us that language technologies must be built with, not for, those whose voices they seek to preserve.

Practical precedents across regions. Proven implementations show how these practices work in the real world: MasakhaNER/2.0 (pan-African) demonstrates distributed, community-led annotation at scale; Mozilla Common Voice (East Africa) shows how mobilisation and validation

raise both quantity and quality of speech data; Te Hiku Media (Aotearoa, Māori) illustrates community-governed licensing that protects cultural value while enabling ASR; and AI4Bharat (India) shows how open pipelines accelerate multilingual MT at national scale. Together, these cases offer transferable patterns, community governance, and open, well-documented pipelines that African projects can adapt. First, infrastructure is not neutral. The languages that are most visible in datasets are not necessarily those most in need; they are often those with easier access to tools, funding, or technical collaborators. Closing these gaps means expanding our taxonomies, redirecting our attention to underserved regions, and centering co-creation. Second, roles matter. Developers may report more measurable technical outputs, but linguists, educators, and community leaders contribute to essential relational and cultural values. A sustainable NLP ecosystem will require these roles to work in partnership, not in silos.

Community Engagement: Evidence-Based Practical Guidelines for Digitising Indigenous Languages

A focused desktop review scan of recent African language technology research including key works by Siminyu et al. (2023), Ajuzieogu (2023), Obasa (2024), Dev et al. (2024), and Smart et al. (2025), reveals critical insights into the impact of community engagement on technical outcomes. Across these studies, a consistent pattern emerges: while the underlying technical architectures of machine translation and natural language processing models may remain relatively stable, the degree and quality of community involvement significantly shape the data, evaluation, and ethical frameworks that underpin these technologies. Siminyu et al. (2023) emphasize that many challenges in African NLP arise from a lack of foundational language tools such as digital dictionaries, keyboards with diacritics, and culturally-aware spellcheckers that are essential for producing accurate datasets. Without these, datasets are prone to errors that degrade model performance. Ajuzieogu (2023) further demonstrates that early and ongoing consultation with language communities, as shown in their Amharic project, leads to more trusted, contextually relevant data and smoother technical workflows.

Obasa (2024) raises important ethical considerations, arguing that many global AI ethics guidelines are grounded in Western values and often fail to address the complex cultural and historical realities of post-colonial African societies. Without community input, models risk reinforcing biases and misrepresentations that can cause measurable harm. Dev et al. (2024) add that local social nuances such as stereotypes and identity categories are often invisible to large language models and require direct community participation to be effectively identified and addressed. Finally, Smart et al. (2025) highlight that the definition of “good” translation is socially constructed and varies between communities. Their work advocates for ethnographic

methods, local partnerships, and co-design processes to ensure evaluation metrics capture what truly matters to language users, leading to greater trust and adoption. Together, these studies emphasise that community engagement is not a peripheral activity but a core driver of data quality, bias mitigation, cultural sensitivity, and ultimately, the technical success of language technologies. This desktop review informs the following practical guidelines designed to embed community engagement deeply into the process of digitising and developing indigenous African languages.

List of best Practices for Community-Engaged Digitisation of Indigenous Languages

- Define the Language Context Clearly: Identify language varieties, dialects, and specific community needs before technical work begins.
- Engage Communities as Co-Creators: Involve speakers, educators, and cultural experts early to co-design objectives and data decisions.
- Audit Existing Resources: Review and assess existing datasets for quality, coverage, and gaps to avoid duplication.
- Collect Data Ethically and Authentically: Prioritise natural, consent-based language use with rich contextual metadata.
- Co-Develop Annotation Guidelines: Collaborate with experts to ensure annotations respect linguistic and cultural norms.
- Use Evaluation Methods That Reflect Real Use-Cases: Combine automated metrics with human evaluation led by community members.
- Compare Modelling Approaches Transparently: Benchmark models with and without community-involved datasets to show impact.
- Detect and Mitigate Bias: Identify harmful patterns and validate mitigation with community feedback.
- Document Decisions Clearly: Maintain accessible records of data sources, annotation, and evaluation processes.

- Plan for Sustainability: Develop strategies for ongoing maintenance, community ownership, and capacity building.

Summary

The objective of this chapter is to explore how scale and ecosystem sustainability can be built for African NLP in ways that are inclusive, ethical, and community-centered. Key themes include representation and coverage of languages, the role of human infrastructure, accessibility and multilingual expression, and the ethics of participation, archiving, and licensing. Activities combined a literature review of benchmark projects (MasakhaNER, MasakhaNEWS, AfriSenti, LAFAND-MT, and IrokoBench) with a practitioner survey of 90 respondents across diverse roles and regions. Outcomes include the identification of role-based imbalances, the reframing of success to include social and cultural impact alongside technical outputs, and the synthesis of best practices that emphasize interdisciplinary teams, accessible tools, and ethical sharing. The chapter concludes by advancing a vision of care-centered digital public infrastructure, rooted in Ubuntu, where technologies are built with, rather than for, African language communities.

Appendices

Appendix A: Comprehensive Glossary

Accent Standardization	The process of ensuring linguistic coherence by prioritizing a specific, widely understood regional accent during initial data collection to enhance model performance.
Adaptive Acoustic Models	Machine learning models that dynamically adjust to variations in speech patterns, accents, and background noise to improve speech recognition accuracy.
African NLP	The study and development of natural language processing tools and models for African languages, addressing low-resource challenges.
AI (Artificial Intelligence)	The simulation of human intelligence in machines that are programmed to think, learn, and mimic human actions.
AI Translation	The automated conversion of text or speech from one language to another using AI algorithms.
AI-driven Pedagogy	The use of artificial intelligence to enhance and personalize teaching and learning methods.
Annotation	The process of labeling data (such as text, images, or speech) with meaningful tags to train and evaluate AI models.
Annotator	An individual, typically a language expert, responsible for reviewing, labelling, or transcribing data for accuracy.
API (Application Programming Interface)	A set of rules and tools that allows different software systems to communicate and exchange data with each other.

ASR (Automatic Speech Recognition)	Technology that converts spoken language into written text.
AUDA-NEPAD	African Union Development Agency – New Partnership for Africa's Development.
B2B (Business-to-Business)	Refers to transactions or business conducted between two companies, rather than between a company and an individual consumer.
Bias Auditing	The process of systematically evaluating AI systems for unfair, prejudiced, or inequitable outcomes, especially regarding gender, ethnicity, or dialect.
Bit Rate	The number of bits processed per second in an audio file, affecting its quality and size.
BLEU (Bilingual Evaluation Understudy)	A metric for evaluating the quality of machine-translated text.
BLEU Score	A score from 0 to 1 that measures how closely a machine-translated text matches a set of high-quality human translations.
BMGF	Bill and Melinda Gates Foundation.
Code-Mixing	The practice of alternating between two or more languages or dialects within a single conversation or sentence, common in multilingual communities.
Context-aware AI	AI systems designed to take surrounding context (such as user history or location) into account to improve accuracy and relevance.
Contextual Language Modelling	Machine learning techniques used to understand and predict words based on the surrounding text or speech context.

Corpus (Plural: Corpora)	A large and structured collection of texts or audio recordings used for linguistic analysis and training language models.
Crowdsourcing	The practice of obtaining data, services, or ideas by enlisting contributions from a large group of people, typically via an online platform.
Cultural Embedding	The process of integrating cultural elements, norms, and nuances into digital products to ensure local relevance and user acceptance.
Cultural Sensitivity	The practice of respecting cultural norms and values by avoiding potentially offensive or inappropriate content during data collection and technology design.
Data Enrichment	The process of enhancing a dataset by adding more diverse and representative data to improve an AI model's performance and reduce bias.
Data Governance	The overall management of the availability, usability, integrity, and security of data within an organization or project.
Data Privacy	The protection of personal information, ensuring that no sensitive details (e.g., names, phone numbers) are collected or exposed.
Demographic Coverage	The practice of ensuring diversity in a dataset by including participants from various ages, genders, education levels, and geographic backgrounds.
Dialectal Variations	Differences in pronunciation, vocabulary, and grammar that exist within a single language, often tied to region or social group.
Digital Equity	The goal of ensuring that all individuals and communities have fair access to and use of digital tools, resources, and technologies.
Digitization	The process of converting information from a physical or analog format into a digital one.

Domain-specific Corpora	Datasets of text or speech focused on a particular subject area, such as healthcare, finance, or agriculture.
DSI	Department of Science and Innovation (South Africa).
ELAN (EUDICO Linguistic Annotator)	A professional software tool used for creating complex, time-aligned annotations on video and audio resources.
Ethical AI	The practice of developing and deploying AI in a way that adheres to ethical principles, including fairness, transparency, accountability, and respect for human rights.
Ethical Data Collection	The process of gathering data responsibly by obtaining informed consent, protecting privacy, and ensuring fairness to participants.
Evaluation Benchmarks	Standardized datasets and metrics used to consistently assess and compare the performance of different AI models or systems.
Fine-tuning	The process of taking a pre-trained AI model and further training it on a smaller, specialized dataset to adapt it for a specific task or domain.
Gamified Learning Platforms	Educational tools that incorporate game-like elements (e.g., points, badges, leaderboards) to enhance user engagement and motivation.
Heuristic Evaluation	A qualitative assessment method where experts review a system or dataset against a set of established principles (heuristics) to identify potential issues, such as in linguistic coherence or logical flow.
Human Post-editing	The process where a human translator reviews and corrects text generated by a machine translation system to improve its quality and accuracy.
Human-annotated Dataset	A dataset where labels, tags, or transcriptions have been added or verified by people, ensuring a high level of accuracy for training supervised models.

Hybrid MT	A machine translation system that combines two or more different MT approaches (e.g., rule-based, statistical, and neural) to leverage their respective strengths.
ICT (Information and Communications Technology)	An umbrella term for technologies and resources used to handle telecommunications, broadcast media, information management systems, and network-based control.
Impact Assessment	The process of measuring the real-world effects of a project, beyond technical metrics, to include social, economic, and user-centric outcomes.
Informed Consent	A formal process where participants are given full information about a project—including risks, benefits, and how their data will be used—before they agree to participate.
Language Consultant (LC)	An individual, typically a native speaker, who is recruited to provide voice recordings or other linguistic data for a collection project.
Language Lead	An expert linguist who oversees the transcription and validation process for a specific language, sets orthographic guidelines, and performs the final quality assurance checks.
Linguistic Bias	Systematic errors or skewed outcomes in AI models that result from unrepresentative or imbalanced training data, potentially disadvantaging certain linguistic groups.
Linguistic Inclusivity	The principle of ensuring that language technologies are accessible and functional for speakers of all languages, especially those from marginalized or low-resource communities.
LLM (Large Language Model)	A type of deep learning AI model trained on vast amounts of text data to understand, generate, and process human language at a sophisticated level.

Low-resource Language	A language for which there are few publicly available digital resources, such as datasets, computational tools, or research.
LSP (Language Service Provider)	A company that offers professional language services, including translation, interpretation, and localization.
ML (Machine Learning)	A subfield of AI that focuses on building systems that can learn from and identify patterns in data to make predictions or decisions with minimal human intervention.
Monolingual Text	A body of text written in a single language, often used to train language models to understand grammar and context.
MT (Machine Translation)	The use of software to automatically translate text or speech from one language to another.
Multi-Tier Data Validation	A structured quality control process involving multiple layers of automated and human review to ensure data accuracy and reliability.
Multilingual Parallel Corpus	A dataset containing the same text aligned sentence-by-sentence in multiple languages, essential for training translation models.
Multilingual Transfer Learning	A machine learning technique where knowledge gained from training on a high-resource language is applied to improve performance on a related low-resource language.
Multilingual User Interface	A digital interface that can be displayed in multiple languages, allowing users to select their preferred language.
Native Speaker (L1 Speaker)	An individual who has acquired a language as their first language from birth.
NCAIR	National Centre for Artificial Intelligence and Robotics (Nigeria).
NER (Named Entity Recognition)	An NLP task that involves identifying and classifying named entities (e.g., persons, organizations, locations) in text.

NLP (Natural Language Processing)	A field of AI focused on enabling computers to understand, interpret, process, and generate human language.
NMT (Neural Machine Translation)	The current standard approach to machine translation, which uses deep neural networks to translate entire sentences.
Offline-first	An approach to software design where an application is built to function effectively without a constant internet connection, crucial for low-bandwidth environments.
Open-source	Software or data that is made freely available for anyone to use, modify, and distribute, fostering collaboration and transparency.
Orthographic Conventions	The standardized rules for writing a language, including spelling, punctuation, capitalization, and hyphenation.
Parallel Text/Corpus	A collection of texts, each presented in two or more languages, where the texts are translations of each other.
POS (Part-of-Speech) Tagging	The process of marking up a word in a text as corresponding to a particular part of speech (e.g., noun, verb, adjective).
Post-hackathon Phase	The period following a hackathon event, dedicated to providing mentorship, funding, and support to help teams develop their prototypes into viable products.
Resource-efficient Architecture	AI models designed to require less computational power and memory, making them suitable for deployment on mobile devices or in low-resource settings.
SADiLaR	South African Centre for Digital Language Resources
SEA-LION	Southeast Asian Languages in One Network.

Sentiment Analysis	The computational task of identifying, extracting, and categorizing opinions and emotions expressed in a piece of text.
Signal-to-Noise Ratio (SNR)	A measurement that compares the level of a desired signal (speech) to the level of background noise. A higher SNR indicates a cleaner recording, with a benchmark of 40dB or higher often used for quality data.
Speech Dataset	A structured collection of recorded spoken language, often paired with transcriptions, used to train and evaluate speech technologies.
Speech Recognition	See ASR (Automatic Speech Recognition).
Speech-to-Text Conversion	The process of converting spoken language into written text using automated algorithms.
Supervisory Annotator	An experienced reviewer who oversees the work of initial annotators, ensuring consistency and accuracy in language data.
Task-specific Benchmarking	The practice of creating and using standardized evaluation metrics and datasets tailored to a specific NLP task (e.g., translation, sentiment analysis).
Technical Criteria	The specific parameters—such as bit rate, sampling rate, and SNR—used to objectively assess the quality of an audio recording.
Theory of Change	A framework that maps out how specific activities or interventions are expected to lead to desired long-term outcomes.
TTS (Text-to-Speech)	Technology that converts written text into synthesized spoken voice output.
Transcription Validation	The process of verifying the accuracy of a transcribed text against its original audio recording.

Transfer Learning	A machine learning method where a model developed for one task is reused as the starting point for a model on a second, related task.
Translation Memory	A database that stores previously translated segments (sentences, paragraphs) to aid human translators and ensure consistency.
Validation and Transcription Teams	The groups responsible for checking audio quality and ensuring that transcriptions accurately reflect the recorded speech.

Appendix B: Ecosystem Actor Directory

This directory provides a non-exhaustive list of key actors within the African language technology ecosystem, categorized by their primary role. The descriptions are based on their activities as cited in this report.

Startups and Companies	• Lelapa AI (South Africa): An AI research and product lab focusing on building enterprise-grade speech and text processing tools for African languages like isiZulu and Afrikaans. • EqualyzAI (Nigeria): A startup building hyperlocal, culturally grounded Small Language Models (SLMs) and hybrid translation APIs for Nigerian languages such as Yoruba and Hausa. • African Translation: A specialized Language Service Provider (LSP) offering certified translation and interpretation services for a wide range of African languages. • AfriLingual: A company that developed a medical translation engine for Nigerian languages, including Kanuri and Tiv, to reduce patient misdiagnoses.
Academic Institutions and Research Labs	• NITHub (University of Lagos, Nigeria): A university-based technology hub that serves as a strategic partner connecting academic research with scalable technological deployment. • Africa Centre of Excellence on Technology Enhanced Learning (ACETEL): A research center that partnered with Data Science Nigeria on the TangaleNLP project to create a machine translation system for an endangered language. • Note: This list is representative. Numerous universities and research centers across the continent are engaged in NLP research, language documentation, and digital linguistics.
Community Movements and	• Masakhane: A pan-African, grassroots NLP community that creates open-source datasets (e.g., MasakhaNER) and

Open-Source Initiatives	models for African languages through the collaboration of thousands of volunteers. • African Languages Lab (All Lab): An initiative that aggregates voice data for dozens of African languages through its crowdsourcing application, the All-Voices App. • GhanaNLP: A community-driven initiative in Ghana focused on developing NLP resources, including speech datasets for languages like Akan and Ga, to enable inclusive technology solutions. • TangaleNLP Project (Nigeria): A community-engaged project that developed a machine translation system for the endangered Tangale language by directly involving native speakers in the data creation process. • UlizaLlama (Kenya): A community co-creation project that developed culturally sensitive Swahili medical chatbots to deliver relevant healthcare advice. • Digital Umuganda (Rwanda): A government-supported initiative that built an open-source Kinyarwanda corpus through community data collection, promoting transparency and accessibility.

Funders, Enablers, and Global Collaborators	This category includes a diverse group of organizations that provide financial, technical, and strategic support to the ecosystem.
	• Philanthropic and Multilateral Funders
	  ○ Lacuna Fund: A major philanthropic funder that sponsors the creation of open machine learning datasets for low-resource languages, particularly in critical domains like agriculture and health.
	  ○ Bill and Melinda Gates Foundation (BMGF): A global foundation that provides funding for targeted impact projects, such as hackathons designed to address gender bias in AI for African languages.

	○ UNESCO: An intergovernmental organization that advocates for linguistic diversity online and supports the development of policies for digital language inclusion. ● Governmental Bodies and Initiatives ○ National Centre for Artificial Intelligence and Robotics (NCAIR) (Nigeria): A Nigerian government agency that provides grants, cloud credits, and strategic support to local AI startups and researchers. ● Ecosystem Builders and Networks ○ Data Science Nigeria (DSN): A leading AI network in Nigeria that acts as a key ecosystem builder by connecting research, policy, education, and industry through projects and advocacy. ○ AfriLabs: A pan-African network of technology and innovation hubs that supports the ecosystem by organizing and hosting events like impact-focused hackathons. ● Corporate and Global Tech Collaborators ○ Google: A global technology company that provides infrastructure support (e.g., cloud computing), research collaboration, and integration of African languages into its global platforms. ○ Meta (formerly Facebook): A global technology company involved in collaborative research, hackathon sponsorship (e.g., Llama 3.1 Impact Hackathon), and fine-tuning LLMs for African languages. ○ Microsoft: A global technology company that provides infrastructure, research grants, and support for language digitization efforts and AI development on the continent. ○ OpenAI: An AI research and deployment company collaborating on projects with local partners to fine-tune

	large language models for specific African languages like Wolof. ○ Orange: A major telecommunications operator involved in public-private partnerships to develop and deploy language models for its customer base in various African markets

Appendix C: Budget, Timelines, and Resource Management Toolkit

With a clear goal and a defined methodology, the next step is to build a practical blueprint for execution. This is where your vision meets the real-world constraints of time, money, and people. A well-considered plan not only guides your project but also becomes your most powerful tool for attracting funders, partners, and community contributors. This section provides a framework for creating a realistic budget, setting achievable timelines, and strategically managing your resources.

Budgeting Beyond the Bottom Line

In the African language technology ecosystem, a budget is more than a spreadsheet of expenses; it's a holistic account of all the resources that fuel your project. Successful projects learn to value and leverage three distinct types of capital.

- **Financial Capital:** This is the direct funding you receive from grants or investment. When applying for grants from organizations like the **Lacuna Fund** or national bodies like Nigeria's **NCAIR**, your financial request should be detailed and justified. Break down costs clearly: Are you paying for data annotators, cloud computing services, or stipends for community researchers? A transparent budget demonstrates foresight and builds trust with funders.

- **In-Kind Capital:** These are essential non-monetary contributions. A prime example is the provision of cloud computing credits, a resource offered by **NCAIR** that can significantly reduce a project's operational costs. Other examples include access to university facilities, software licenses donated by a partner, or legal advice offered pro bono. Always account for the dollar value of these contributions in your project plan to

reflect the true scale of your resources.

- **Community Capital:** This is the immense value generated by volunteers and grassroots contributors. The work of **Masakhane** is a powerful testament to community capital, where thousands of hours of skilled labor are voluntarily contributed to build open-source datasets and models. While you don't assign a salary to volunteers, you must budget for their support: plan for the costs of community management, training workshops, and the infrastructure needed to coordinate their efforts.

Setting Realistic Timelines

Timelines in this ecosystem are not one-size-fits-all. The scope of your project will determine its cadence. Align your expectations with one of these three common project archetypes:

- **Sprint Projects (Weeks to Months):** These are short, high-intensity efforts focused on a single, well-defined goal, such as a data collection drive, a translation hackathon, or adapting an existing model for a new dialect. They are ideal for community-driven initiatives and for producing quick, demonstrable results.

- **Marathon Projects (6-18 Months):** This timeline is typical for formal research and development projects, often aligned with academic semesters or grant cycles. Building a new foundational model, creating a comprehensive dataset like the one for the **TangaleNLP** project, or developing a version 1.0 product fits within this timeframe. It requires structured project management with clear milestones.

- **Endurance Projects (Ongoing):** These are long-term initiatives focused on building and sustaining infrastructure. The work of startups like **Lelapa AI** or national programs like South Africa's **SADiLaR** falls into this category. The goal is not a single deliverable but continuous improvement, platform maintenance, and ecosystem growth.

Pro-Tip: Document Everything from Day One
Your project's documentation is one of its most valuable assets. A well-documented process makes it easier to onboard new team members, justify your budget to funders, and allows others to build on your work. Use free tools like GitHub for code and version control, and maintain a clear, accessible record of your data collection and cleaning

decisions. This practice saves time, prevents knowledge loss, and is a hallmark of a mature project.

Managing Your Core Resources

Finally, your success depends on how well you manage the three most critical resources at your disposal: your people, your data, and your tools.

- **Your People:** A project is only as strong as its team. Identify the key roles you need to fill: Are you looking for linguists, software engineers, community managers, or a combination? The **TangaleNLP** project, for example, succeeded by integrating academic experts with native speakers from the community. Define these roles and responsibilities early on.
- **Your Data:** Before starting a new data collection effort, conduct a thorough search of existing resources. Platforms like the **African Languages Lab (All Lab)** and **Hugging Face** may already host datasets you can use or adapt, saving you months of work. If you must create a new dataset, plan for the entire lifecycle: collection, cleaning, annotation, documentation, and—critically—sharing it back with the community.
- **Your Tools:** Leverage the open-source nature of the ecosystem. You don't need to build everything from scratch. Use pre-trained models from the community as a starting point and fine-tune them for your specific language. Take advantage of crowdsourcing platforms like **Mozilla's Common Voice** for speech data collection. This approach accelerates your timeline and allows you to focus your limited resources on what makes your project unique.

Appendix D: Existing data and their sources

Name	Type	Language	Task
YANKARI	text	Yoruba	Machine Translation (MT), Part-of-Speech (POS) tagging,Named Entity Recognition (NER), Language Modeling, Text Classification
MENYO-20k	text	Yoruba	Machine Translation
Open source yoruba	speech	Yoruba	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)
Hausa Corpus	text	Hausa	Language Modeling,POS Tagging,NER, Text Classification
IroyinSpeech	speech	Yoruba	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)
Hausa Speech Dataset	speech	Hausa	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)
NaijaRC	text	Hausa, Igbo, and Yoruba	Question Answering, Information Retrieval
MasakhaNews	text	Hausa, Igbo, Yoruba, Pidgin and other African languages	Text Classification (News Topic Classification)
BibleTTS	speech	Yoruba, Hausa and others	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)

Yorùbá Flickr Audio Caption Corpus	speech	Yoruba	Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Speech-based Image Captioning
Kencorpus	text and speech	Kiswahili, Dholuo, Luhya-Lubukusu, Luhya-Logooli, Luhya-Lumarachi	Automatic Speech Recognition (ASR), Text-to-Speech (TTS),Language Modeling, POS Tagging,NER,Text Classification
KenSpeech	speech	Swahili	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)
MAFAND-MT	text	Bambara, Ghomala, Ewe, Fon, Hausa, Kinyarwanda, Luganda, Dholuo, Mossi, Chichewa, Nigerian-Pidgin, chiShona, Setswana, Twi, Wolof, and isiXhosa	Machine Translation
Financial Inclusion Speech Dataset for some Ghanaian Languages	speech	Akan (Akuapem Twi, Asante Twi, Fante) and Ga	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)
IgboSynCorp	speech	Igbo	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)
Bayelemabaga Aligned Bambara-French Corpus for	text	Bambara	Machine Translation

Machine Translation			
KALLAAMA	speech	Wolof, Pulaar, and Serer	Automatic Speech Recognition (ASR), Text-to-Speech (TTS)
AFRIDOC-MT	text	Amharic, Hausa, Swahili, Yorùbá, and Zulu	Machine Translation
Masakhane-NLU	text	Amharic, Ewe, Hausa, Igbo, Lingala, Luganda, Oromo, Kinyarwanda, Shona, Sesotho, Swahili, Twi, Wolof, Xhosa, Yoruba and Zulu	NER,Sentiment Analysis,Question Answering,Text Classification
Expanding a parallel corpus of Portuguese and the Bantu language Emakhuwa	text	Emakhuwa	Machine Translation
Lacuna PII Multilingual Dataset	text	Luganda, Lumasaba, Hausa, and Kanuri	Named Entity Recognition (NER), Language Modeling, Text Classification

Appendix E: Tools

Tool	Description
Pratt	Praat is a free, cross-platform software used for speech analysis and annotation. It allows researchers to annotate, segment, and label audio files using TextGrid annotation tiers. You can also record sounds for annotating and editing of noise, speech manipulation etc. Download Praat via this link

Appendix F: Chapter 3 Survey Questions

1. What is your primary role in language technology or African languages?
2. Which African languages do you work with or have an interest in?
   (You can list up to 10 of your most frequent.)
3. Have you worked with machine translation or small language models for African languages?
4. What language resources or datasets have you used in your work?
5. Would you be willing to share these resources with the broader community?
6. What barriers have you faced when accessing or developing African language datasets?
7. How do you engage with local NLP communities or research networks when developing language technologies?
8. How do you engage with communities who speak the local language when developing language technologies? Why or why not? Elaborate on what worked or didn't work.
9. What are the top three improvements you would like to see in African language technology?
10. What ethical considerations should be prioritised when collecting or using African language data?
11. Are there specific policy gaps that hinder the development and adoption of African language technology?
12. How can machine translation or NLP tools better serve African communities?
13. What are the biggest risks or challenges of developing NLP for African languages?
14. Would you be interested in collaborating with others working in African NLP? If yes, what kind of support or partnerships would be most beneficial?
15. Any additional thoughts on how to advance African language technology?

16. Would you like to be contacted for further discussion?

References

[1] "Lelapa AI and Vodacom: A Case Study," Lelapa AI Blog, 2024.

[2] "Vambo AI: Multilingual Customer Service Solutions," vambo.ai, 2024.

[3] "EqualyzAI: Hybrid Translation for Nigerian Languages," EqualyzAI Whitepaper, 2024.

[4] "Masakhane: A Decade of Progress in African NLP," Journal of African Language Technologies, 2024.

[5] "All-Voices App: Crowdsourcing African Language Data," African Languages Lab, 2023.

[6] "NITHub and Scalable Deployment," University of Lagos, 2023.

[7] "Meta and Orange Collaborate on African Language Models," Meta AI Blog, 2023.

[8] "TangaleNLP: Community-Driven Language Preservation," Data Science Nigeria, 2023.

[9] "Masakhane Community Survey 2024."

[10] "GhanaNLP: Building Speech Corpora for Akan and Ga," Proceedings of LREC 2024.

[11] "ACETEL and Data Science Nigeria Partnership," ACETEL News, 2023.

[12] "African Translation Services," africantranslation.com.

[13] "AfriLingual Medical Translation Engine: A Pilot Study in Northern Nigeria," Journal of Health Informatics in Africa, 2023.

[14] "Reducing Misdiagnosis Rates with Local Language Health Information," World Health Organization Bulletin, 2024.

[15] "Data Science Nigeria Annual Report 2023."

[16] "NCAIR Grant Program," NCAIR Nigeria, 2023.

[17] "SADiLaR Annual Report on Digitisation," South African Department of Science and Innovation, 2023.

[18] "OpenAI and Wolof Language Model," OpenAI Blog, 2023.

[19] "Lacuna Fund Impact Report 2023."

[20] "OpenAI, Orange, and the Wolof Language," Orange Press Release, 2023.

[21] EqualyzAI Website, https://equalyz.ai/

[22] "Masakhane Hackathons," https://www.masakhane.io/hackathons

[23] "Africa Innovation Week," UNESCO, https://en.unesco.org/events/africa-innovation-week

[24] "AfriLabs Llama 3.1 Impact African Hackathon," AfriLabs, https://www.afrilabs.com/afrilabs-is-thrilled-to-announce-the-llama-3-1-impact-african-hackathon-for-ai-innovators-a-pan-african-project-funded-by-meta-and-bmgf/

[25] "AI for Good Impact Africa Summit," ITU, https://aiforgood.itu.int/event/impact-africa/

[26] "DSN AI Bootcamp & Hackathons," Data Science Nigeria, https://datasciencenigeria.org/2025_aibootcamp/

[27] "AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages," Lanfrica, https://lanfrica.com/record/afrisenti-a-twitter-sentiment-analysis-benchmark-for-african-languages

[28] "IrokoBench: A Human-Translated Benchmark for 16 African Languages," arXiv, https://arxiv.org/abs/2406.03368

[29] "UlizaLlama: Co-creating Culturally Sensitive Health Chatbots," Proceedings of CHI 2024.

[30] "Vodacom Rural Connectivity Report 2023."

[31] "MaLLaM: A Blockchain-based Framework for Fair Compensation in Language Data Crowdsourcing," IEEE Transactions on Computational Social Systems, 2022.

[32] "Indus OS: Reaching the Next Billion Users," TechCrunch, 2019.

[33] "The Impact of Local Language Interfaces on Digital Adoption in India," Digital India Foundation, 2021.

[34] "DeepSeek AI's Lightweight Models for Rural China," DeepSeek AI Technical Report, 2023.

[35] "SEA-LION: A Collaborative Approach to Southeast Asian NLP," Proceedings of ACL 2023.

[36] "PhoGPT: A Pre-trained Language Model for Vietnamese," VinAI Research, 2023.

[37] "Digital India Mission: Language Technology Grants," Ministry of Electronics and Information Technology, Government of India, 2022.

[38] "The Economic Impact of India's Digital Language Ecosystem," NASSCOM Report, 2023.

[39] "Masakhane Universal API: Powering Multilingual Chatbots in Africa," Masakhane, 2024.

[40] "Smart India Hackathon 2022: Problem Statements and Outcomes," Government of India, 2022.

[41] "From Hackathon to Governance: Integrating AI in Public Services," e-Gov India Magazine, 2023.

[42] "IndicNLPSuite: A Suite of Resources for Indian Languages," AI4Bharat, 2022.

[43] "XTREME-R: A Massively Multilingual Benchmark for Cross-lingual Transfer," Google AI, 2021.

[44] "Singapore's AI Governance Framework," Infocomm Media Development Authority, 2020.

[45] "Bhashini: India's National Language Translation Mission," bhashini.gov.in.

[46] "Crowdsourcing Language Data at Scale: The Bhashini Model," Proceedings of the Digital Government Society, 2023.

References from Chapter 3

Adebara, I. (2024). Towards Afrocentric Natural Language Processing.

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., et al. (2021). MasakhaNER: Named Entity Recognition for African Languages. Transactions of the Association for Computational Linguistics, 9, 1116–1131. https://doi.org/10.1162/tacl_a_00416

Ajuzieogu, U. C. (2023). Ethical data augmentation techniques for low-resource language AI: A framework for African languages. Research Report. July 2023.

Ahmed, A. (2007). Open access towards bridging the digital divide–policies and strategies for developing countries. Information Technology for Development, 13(4), 337–361. https://doi.org/10.1002/itdj.20067

Alabi, J., Adelani, D. I., Ruiter, D., & España-Bonet, C. (2022). A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. arXiv preprint arXiv:2205.02022. https://arxiv.org/abs/2205.02022

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

https://doi.org/10.1145/3442188.3445922

Dev, S., Jha, A., Goyal, J., Tewari, D., Dave, S., & Prabhakaran, V. (2023). Building NLP evaluation resources with LLMs and community engagement for scale and depth. Paper presented at the 5th Deep Learning Indaba Conference (DLI 2023).

Gaventa, J., & Barrett, G. (2010). So what difference does it make? Mapping the outcomes of citizen engagement. IDS Working Papers, 2010(347), 01–72. https://doi.org/10.1111/j.2040-0209.2010.00347_2.x

Muhammad, S. H., Ruiter, D., Orife, I., Adelani, D. I., Emezue, C. C., et al. (2023). AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. EMNLP 2023. https://aclanthology.org/2023.emnlp-main.862/

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., et al. (2020). Participatory research for low-resourced machine translation: A case study in African languages. arXiv preprint arXiv:2010.02353. https://arxiv.org/abs/2010.02353

Ruiter, D., Emezue, C. C., Orife, I., Ogayo, P., Adelani, D. I., et al. (2023). MasakhaNEWS: News Topic Classification for African Languages. arXiv preprint arXiv:2304.09972. https://arxiv.org/abs/2304.09972

Siminyu, K. (2022). Licensing African Datasets: A Participatory Framework for Data Ownership and Use. Carnegie Endowment for International Peace (CEIP) Working Paper. work-pub-88417

Siminyu, K, Abbott, J, Túbọ̀sún, K, Mthembu, A. T, Ramkilowan, A, Oladimeji, B. (2023). Consultative engagement of stakeholders toward a roadmap for African language technologies. Patterns, 4, 100820. https://doi.org/10.1016/j.patter.2023.100820

Smart, A. (2024). Socially responsible data for large multilingual language models. In ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2024). arXiv:2409.05247 [cs.CL]. https://doi.org/10.48550/arXiv.2409.05247

Soria, C., Pretorius, L., De Schryver, G.-M., & Monachini, M. (2012). PanLex and the lexical infrastructure for the world's languages. Language Resources and Evaluation, 46(2), 191–209. https://doi.org/10.1007/s10579-012-9173-0

Tapo, K., Adelani, D. I., Kreutzer, J., Emezue, C. C., Orife, I., & Ruiter, D., et al. (2023). IrokoBench: Benchmarking African Languages on Multilingual Language Models. arXiv preprint arXiv:2406.03368. https://arxiv.org/abs/2406.03368

Obasa, A. E. (2024). Large language models through the lens of ubuntu for health research in sub-Saharan Africa. South African Journal of Science, 120(5/6), Article #16814. https://doi.org/10.17159/sajs.2024/16814

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell et al. (Eds.), New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts (pp. 65–92). Aspen Institute.