

African Languages and Digital Justice

A Landscape on Community-Centered Language Technologies

whose knowledge?

AUTHORS
Claudia Pozo
Maari Zwick-Maitreyi

ADVISORS
Tigist Shewarega Hussen
Anasuya Sengupta
Kathleen Siminyu

TRANSLATORS
Translated to Tigrinya by Berhane Achame
and proofread by Fikadu Gebregziabher

Translated to Kiswahili by Melchior Shukuru
and proofread by Exuper Njau

This research-in-action process was conducted in partnership with the Distributed AI Research Institute (DAIR).

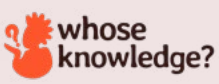
Made possible through the generous support of The Internet Society Foundation (ISOC).

Graphic design and layout by Ellena Ekarahendy, with art from Studio Ping Pong.

Issued under the Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0).

2026







Summary 01

Introduction 05

What is this report about? 06

1. History and context 08

The roots of digital language erasure 09

Digital coloniality: Big Tech's extractive production framework 11

2. Research philosophy and methodology 13

3. Community-centered language technologies 16

What are they? 17

What is community? 19

Technologists as facilitators 22

Motivations for prioritizing language technology 24

Digital invisibility 24

Information gaps 25

To be in community 26

Decolonizing knowledge systems 27

Implementing African relational philosophies in technology design 29

Data collection, representation and fair compensation 31

Online versus on-ground data collection 31

Representation and equity within data collection 33

Gender gap in community-centered language technology 34
Compensation for data contributors 36
Funding challenges 38
Funding sources 39
The complexity of the funding application process 40
Community-centered initiatives and Big Tech funding 42
The influence of funding on practices 42
Funding models that communities aspire to 44
Colonial standardization and dialects 47
Emerging strategies for language diversity 50
The impact of technologist bias 52
The paradox of preservation 52
Open knowledge repositories (OKRs) as digital ecosystems 54
Limitations within open platforms 55
Navigating translation gaps 56
Redefining open knowledge 60
Big Data, extraction, and openwashing 62
Transitioning toward stewardship 63
Carving a path towards digital sovereignty 65
Pragmatism and the African digital context 66
Towards community-in-the-lead 67

4. Conclusion 69
Definitions 72
Gratitude 77
Other Resources 79

5. References 80

Summary



In this report we map and analyze different language technology initiatives across East and West Africa. These community-built technologies and practices demonstrate significant shifts away from Big Tech's extractive models, and toward systems rooted in self-determination and African epistemologies and ontologies, or local ways of knowing and being.

This is a paradigm shift that challenges Big Tech's tendency to equate its own US-centric, English-dominant values with a universal global standard. In contrast, the grassroots initiatives we cover define the "local" as the foundation for epistemic sovereignty, asserting that language technologies must remain anchored to the communities that produce them. This approach does not assume "local" is small or minor by comparison to "global"; rather, it prioritizes situated knowledge in context-specific problem-solving.

We live in a historic moment, marked by compounding crises including climate change. Resisting these global polycrises means we need to interrogate the economic, ecological and social systems underpinning Big Tech dominance even as we actively reimagine the trajectory of technological development.

In particular, we need to re-examine the profoundly contested concept of "Artificial Intelligence" (AI), rooted as it is in Big Tech language technologies and narratives that frequently portray AI as an autonomous, singular, and almost mythological entity.

We prefer to ground our approach in the framing of journalist and author Karen Hao, who defines AI as *"a name for a grab bag of technologies that are supposed to mimic human capabilities"* (Shadel, 2025). Stripping away Big Tech hype, allows us to understand AI as an evolving umbrella of computational methods and emphasize Hao's point that AI is ultimately *"a field shaped by familiar human forces: money, power, and ideologies"*

(Hao, 2025). In fact, centering the human choices, labor, and structural power behind these systems exposes how technological evolution remains deeply intertwined with historical systems of coloniality, capitalism, and other social structures of power.

Even as AI carries significant negative impacts on society and ecology, and holds deep colonial implications, it also holds potential to serve communities positively. This work shows that AI can move beyond LLMs, trendy chatbots and Silicon Valley, in language technologies that provide tangible social and economic benefits for those communities who choose to use it to serve them.

We engage directly with subsets of AI that are language specific technologies such as Natural Language Processing (NLP),[1] Machine Learning (ML), Foundation Models,[2] Automatic Speech Recognition (ASR), and Machine Translation (MT) among others. These are critical technologies that actually allow the preservation, representation and expansion of the languages of the Global Majority,[3] most of which are, at present, existentially threatened and seriously underserved.

The initiatives documented in this study prove that alternative ways of building technology are already being practiced on the ground, offering models of hope, language justice, and epistemic sovereignty.

We use this mapping to define community-centered sociotechnical systems within the contemporary social and historical realities of

[1] Natural Language Processing (NLP) is a subfield of computer science focused on the computational analysis and processing of natural human language.

[2] Foundation Models are a broad category of AI models that can process many types of data (such as text, images, audio, video, etc.) as opposed to LLMs which focus strictly on human language and computer code.

[3] See Definitions section

the regions in Africa, critically analyzing the intricacies of scriptism[4] and the resulting tensions between oral traditions and written-text standardization in digital spaces. We investigate community priorities for ethical data collection, governance, and collective stewardship, all frameworks explicitly designed to safeguard local knowledge from capitalist exploitation. We also explore the deep-seated motivations that sustain grassroots practitioners in the face of structural scarcity, asymmetric funding, and institutional neglect.

This report ultimately charts an actionable path forward, and offers concrete strategic recommendations alongside a "Community-Centered Language Technology: Best Practices Guideline" tailored for funders, policymakers, and technologists.

We at Whose Knowledge?, carried out this work in a partnership with friends and collaborators at the Distributed AI Research Institute (DAIR). Their expertise in language-centered technologies like machine translation and speech recognition, further informed by our shared insights, guides the upcoming stage of our collective work together.

We are deeply inspired by the incredible resilience, creativity, and determination of the African and Global Majority practitioners who contributed to this study. We extend our deepest gratitude to every technologist, scholar, linguist, Wikimedian, and community steward who shared their knowledge with us. Their community practices and hopefulness continuously demonstrate that a fairer, digitally just future is not only possible, it is already being built.

[4] See Definitions section

Introduction



What is this report about?

Language is central to our knowledges, identities, and culture. Different languages can shape and represent different ways of knowing, expressing, and learning about our world. Despite the existence of nearly 7,500 spoken and signed languages globally, almost all online knowledge today is created and accessible only through approximately 12 major languages, predominantly English and other colonial languages.

The continent of Africa, home to more than 2000 languages, representing 30% of global linguistic diversity, is profoundly affected by this imbalance. Our previous research on the State of the Internet's Languages indicates that to engage with today's primary mobile applications and websites, over 90% of people across the continent are forced to use a secondary language. This evidence of linguistic inequity highlights a significant barrier to digital access for African communities (Whose Knowledge? et al., 2022).

This profound level of erasure necessitates a deep re-thinking of the production, governance models, and deployment processes of digital language infrastructures, and an urgent need for more just and community-based language technologies.

To address this, Whose Knowledge?, an organization focused on decolonizing the internet through community-centered participatory methods, and the Distributed AI Research Institute (DAIR), an interdisciplinary organization conducting community-rooted tech research, began collaborating in 2024. Our initiative "Community-

Centered Sovereign Language Technologies for Online Knowledge Repositories" aimed to contribute to creating a safe, trustworthy, and contextually-relevant internet where the multitudes of humans worldwide can have a meaningful online presence. Our central question was how to develop sovereign language technologies[5] for African languages, using people-first design[6] and governance to facilitate content creation for online knowledge repositories like Wikipedia.

The first phase, facilitated by Whose Knowledge?, involved a mapping of existing community-centered language technology efforts, with a specific focus on Africa. The second phase of the work by the Distributed AI Research Institute (DAIR) involves using these insights to build technologies for machine translation and speech recognition in Tigrinya, Kiswahili and their variants.

This report summarizes the findings of the mapping studies, revealing unique features of values, community, design, collaboration, and tech production in African regions that powerfully challenge dominant technological norms and serve as vital ways forward[7] for digital sovereignty and epistemic justice for communities across the Global Majority.

We believe that this work will be useful to communities exploring language technology solutions for their own linguistic contexts, policy makers and state representatives, technologists, Wikimedians, language activists and preservationists, open knowledge advocates, linguists, funders, and anyone invested in language justice for the Global Majority.

[5] See Definitions section
[6] See Definitions section
[7] See our "Community-Led Language Technology: Best Practices Guideline"

1.

History and context



The roots of digital language erasure

The digital landscape is marked by a profound imbalance. Of the nearly 7,500 spoken and signed languages globally, fewer than 15 are well-represented—the majority of which are Eurocolonial.[8] This stems from an internet designed around Western perspectives and epistemological norms that impose a written-word standard on every language. Known as scriptism (Harris, 1980), this bias inherently marginalizes the orally dominant languages[9] spoken by most of the world's population.

For people of the Global Majority, the resulting lack of abundant written and textual corpora restricts the volume of digitizable and therefore machine-readable data. Consequently, a majority of the world's languages have been designated as *low-resource languages*, while a relatively tiny number of largely Eurocolonial languages[10] are called *high-resource languages*. To challenge the non-contextualised use of low-resource languages, we recommend replacing it with **local languages**[11] (Bird, 2022) and **digitally under-resourced languages**.[12] These terms introduce the necessary nuance and capacity to acknowledge and bring forward the specific historical processes that created these enduring disparities.

[8] See Definitions section
[9] See Definitions section
[10] See Definitions section
[11] See Definitions section
[12] See Definitions section

For the Global Majority, this terminology of scarcity requires reconsideration, as this state of under-resourcing is not accidental but a consequence of historical systems of colonial capitalism. Coloniality, in fact, remains a key force in contemporary linguistic dynamics. Across much of Africa, Eurocolonial languages, such as English and French, maintain hegemony over institutional and public spheres, including government, administration, and national education systems. Furthermore, historical and ongoing efforts to "unify" and elevate some local languages over others perpetuate the marginalization and erasure of languages with smaller numbers of speakers. Even where local languages have been officially recognized (such as Kiswahili, Akan, or Amharic), linguistic inequities persist as these dominant local languages eclipse the numerous other local languages spoken within those territories. This unrelenting pressure to conform to dominant languages, whether colonial or local, becomes a driver of declining linguistic diversity.

When local languages are not digitized, the risk is not merely digital decline, but total extinction, a process termed *epistemicide* (Thiong'o, 1986) and likened to *digital language death* (Kornai, 2013). Such loss is not confined to the language itself but extends to the history, collective memory, and vast knowledge systems held by its people. Consider the relative scarcity of online resources, such as educational materials, news, and entertainment media in smaller-scale languages like Dagbani, Nyindu, Kujarge, or Wolane. This severely limits adoption and learning, especially among young people, allowing dominant cultural viewpoints to take over, thus perpetuating hierarchies and resulting in erasures of knowledge, language, and identity.

Digital coloniality: Big Tech's extractive production framework

Today, language technology development is overwhelmingly controlled by a small group of global tech giants, like Google, Microsoft, Meta, Anthropic, and OpenAI, who possess the massive data centers and foundation models necessary for training language AI.

Historically, Big Tech has neglected most Global Majority languages, deeming them too small or unprofitable. When development does occur, these corporations typically employ a Eurocentric, *one-size-fits-all* engineering model. This approach prioritizes large-scale data scraping and statistical patterns, instead of a genuine understanding of a language's cultural, social, political and linguistic contexts.

Within such a restrictive framework, languages are reduced to mere quantifiable resources to be mined and funnelled into foundation models, which are then exported globally as a universal infrastructure. This view disregards the profound fact that language data is embodied; rooted in humans and human creations. It is generated directly from people and communities (Kovacs et al., 2025).

Instead, using this extractive, top-down approach, Big Tech consolidates their power over the linguistic infrastructures of the Global Majority. This constitutes a form of digital necropolitics, after Achille Mbembe's

work on necropolitics (Mbembe, 2003), decisively determining which languages, and by extension, which communities' knowledges, are allowed to thrive in the digital sphere, and which are effectively condemned to digital death.

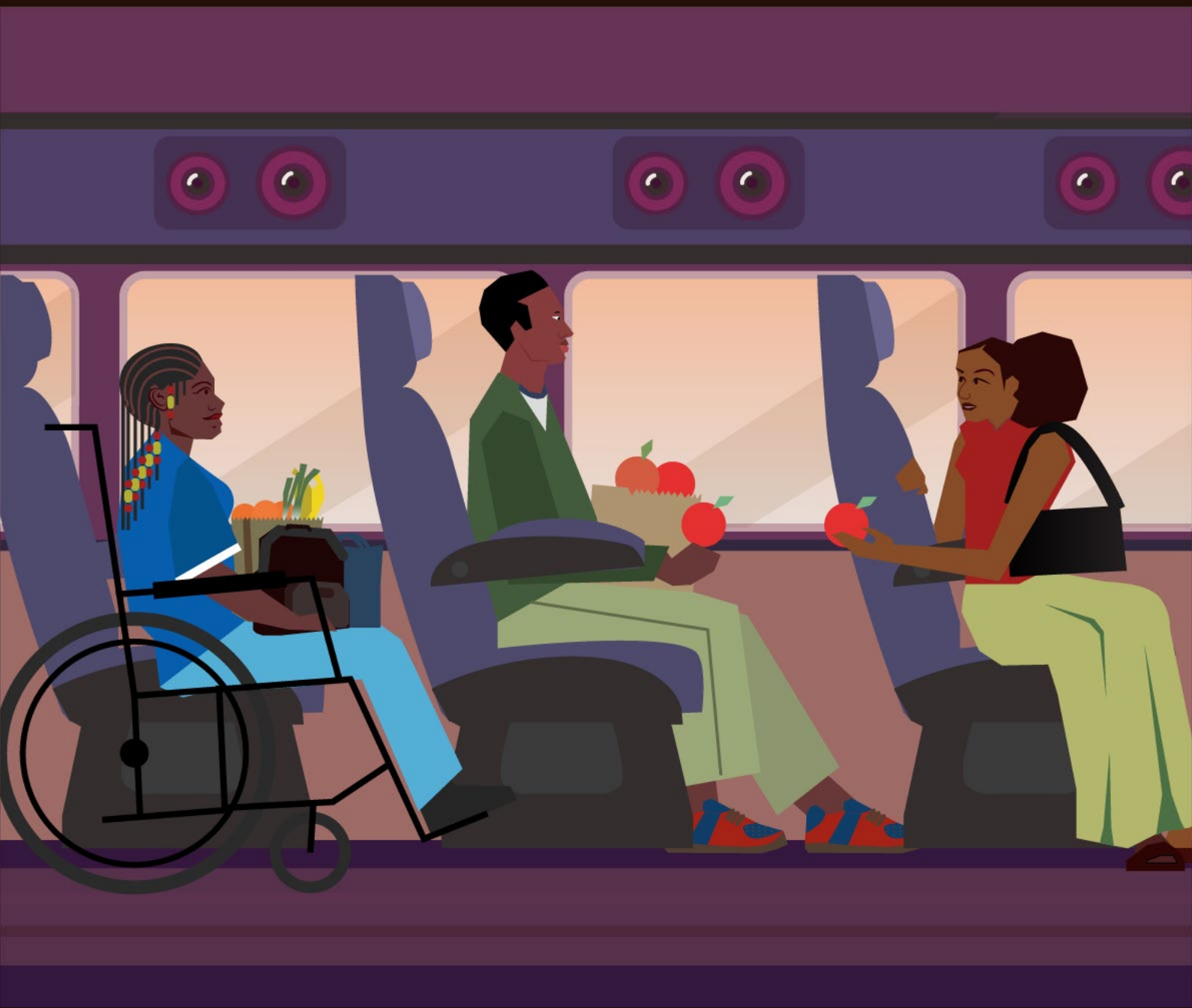
African communities have long countered this oppressive corporate approach with bottom-up responses to technological development. This movement is driven by shared anti-colonial histories, a dedication to social justice, and inspiration from indigenous relational philosophies like Ubuntu, Umuganda, Sankofa, and Teranga among others. At the heart of this framework of community-centered language technologies, is ethos that places people above profit, and process before outputs.

These approaches also ensure that technology is created by and for the speakers themselves, resulting in culturally resonant, and critically—high-quality digital tools.

African communities' experiences, challenges, and solutions offer vital lessons for the rest of the Global Majority, including in the digital lives of our languages, and community-centered technology development more generally.

2.

Research philosophy and methodology



The Language Justice program at Whose Knowledge? is committed to confronting language inequities online by driving structural change through community-centered solutions. Our research process adopted a qualitative approach, guided by feminist, anti-colonial, and disability rights alignments.

Drawing upon decolonizing methodologies studies championed by scholars such as Bagele Chilisa and Linda Tuhiwai Smith, this study was influenced by the belief that research must move beyond the extractive traditions of Western academia. Instead, it should seek to dismantle the unequal power dynamics of the researcher-subject relationship by establishing frameworks that honor and validate people's multiple ways of being, knowing, and doing.

This informed our decision to center the research process on direct, participatory engagement. Our study focused on mapping and documenting existing community-centered initiatives in language technologies by working closely with the practitioners leading them. The goal was to gather and analyze comprehensive insights into the social, historical, personal, and political dimensions of African communities as they develop their own language technologies.

Our strategic advisor and co-researcher, Tigist Shewarega Hussen, was instrumental during this process, helping us to broaden our research community network. Her profound grasp of nuanced local contexts and her feminist theoretical and methodological insights were essential in shaping the depth and direction of this study.

To uphold our core values, particularly our affirmation of multiple forms of experience and expertise, we intentionally engaged a diverse range of participants, ensuring the data was not biased solely toward technical perspectives. All contributors were recognized and compensated for their intellectual labor. Furthermore, our feminist framework for

consent treated it as a dynamic, ongoing dialogue involving continuous information sharing and negotiation.

In the end, it is our firm belief that the narratives surrounding these projects, technologies, and institutions are essentially human stories and that research is not greater than methodology (Chowdhury et al., 2023).

We used a grounded theory approach alongside purposive sampling to emphasize depth over breadth. The core data collection consisted of 22 in-depth semi-structured interviews. We engaged with a diverse group of participants, including technologists, linguists, researchers, and Wikimedians, who are leading community-centered technology creation in languages such as Amharic, Tigrinya, Dagbani, Wolof, Twi, Ewe, Gurune, Kinyarwanda, and Kiswahili. These interviewees represent a broad range of initiatives, including Lesan AI, GalsenAI, Lanfrica Labs, GhanaNLP, Digital Umuganda, EthioNLP, and HausaNLP, among others. They were from countries such as Ethiopia, Senegal, Kenya, Ghana, Rwanda, and Nigeria.

The research prioritized East and West African regions, although it also incorporated perspectives from interviewees located in other geographical areas including Asia and South America.

All interviews were recorded, transcribed, and subjected to thematic analysis. This report presents and explores the main findings drawing from this process.

3.

Community-centered language technologies



What are they?

This subsection defines community-centered language technologies as sociotechnical systems that treat language as collective knowledge rather than an exploitable commercial commodity. It acknowledges the varied stages of digital development for Global Majority languages, before establishing four essential pillars that guide this community-driven approach. These pillars are embodied praxis, epistemic sovereignty, data as collective heritage, and technology for people's benefit. The development of language technologies not merely as a technical effort, but as socially rooted practices designed to empower local communities.

The integration of Global Majority languages into the digital landscape occurs across a spectrum of development. For certain languages, the primary focus is on establishing fundamental textual infrastructure, such as scripts, fonts, keyboards, and Unicode standardization. Others are at the stage of formalizing dictionaries, orthographies, and digital tools for grammar and spell-checking. Simultaneously, numerous communities are advancing toward AI-powered applications such as automatic speech recognition, machine translation, and text-to-speech technologies. Our research largely explores these AI based natural language processing (NLP) technologies by focusing on the communities that develop and sustain them.

We define community-centered language technologies as sociotechnical systems that treat language as collective knowledge and heritage rather than a commodity. We identify key essential pillars that underpin this approach:

Embodied praxis: The technology is built by people who are culturally and linguistically rooted in the community, effectively bridging the gap between developers and users. The user is not a mere subject of study, but is recognized as an active co-creator and human being whose nuanced understanding of social and linguistic contexts shapes the technology's logic.

Epistemic sovereignty: There is a recognition that languages carry specific ways of knowing. Community-centered language technology prioritizes, as much as possible, the preservation of local epistemologies rather than force-fitting them into Western linguistic frameworks.

Data as collective heritage: Language data is not treated as an exploitable commodity, but rather as cultural and historical heritage. This approach is founded on the formation of deep communal trust and continuous consensus building during data gathering.

Technology for people's benefit: These tools are built to solve local problems such as accessing healthcare in one's language of choice, accessing information and infrastructure for agriculture and local industries, documenting endangered oral histories, or enabling local-language education and preservation. Here the social and economic needs of the community take precedence over profit, serving as the primary driver for technological development.

What is *community*?

This subsection explores the multifaceted definitions of *community* among language technology practitioners, highlighting it as a dynamic concept rooted in shared identity, interconnectedness, and collective problem-solving. While the term language "community" should ideally encompass a multi-layered ecosystem of data contributors, elders, linguists, everyday speakers, among others. We note a tendency of technologists to restrictively define it as those within their own immediate professional tech circles, underscoring the need to recenter actual language speakers. Furthermore, we examine the pivotal role of technologists as community facilitators; acknowledging that they frequently operate from positions of relative social and educational privilege. This is a dynamic that inevitably shapes the priorities and decision-making processes in community-centered technologies.

To truly grasp what it means to be community-centered, we must also examine the definition of *community* itself. Among those developing and practicing language technology, the term is far from uniform; instead, it is interpreted through a variety of different perspectives. To clarify what community specifically represents to the people engaged in this field, we prompted participants to share their own definitions.

Amrit Sufi, a scholar, researcher, and language diversity activist spoke of community being linked to identity and collectivity.

> "From my perspective, in a practical sense, community represents practitioners who share a specific identity. Identity is deeply important to me, and as language serves as its carrier, it is within the community that knowledge systems and cultures are safeguarded. When individuals recognize the immense power they hold and unite to harness it, they can revitalize both their language and their culture."
>
> **— Amrit Sufi (Wikimedian)**

Others like Sadik Shahadu, Wikimedian and journalist, describe it as a safe space in which people can come together.

> "So community to me is an avenue where people can feel safe and comfortable in sharing their ideas and collaborating with others to achieve a common goal."
>
> **— Sadik Shahadu (Dagbani Wikimedians User Group)**

This idea of coming together with purpose is also reflected in the actual naming of companies like Digital Umuganda, which derives its identity from the relational Rwandan philosophy of Umuganda. Samuel Rutunda, Digital Umuganda's Chief Technology Officer, elaborates:

> "Every last Saturday of the month as Rwandans, we come together to do some community work. We fix some damaged roads, we build houses of those who are vulnerable within the community. We clean some dirty communal places. So we do communal activities. And then after that we are together, and we discuss issues within the community. This practice of solving problems through community is called Umuganda."
>
> **— Samuel Rutunda (Digital Umuganda)**

Community is also a necessarily dynamic idea in this work. E.M. Lewis-Jong, Founder and CEO of the Mozilla Data Collective, talks about it:

> "We must be able to translate/cascade from community as an ethos, to community as a function. In response to the question 'who is the community?', we work through concentric circles of impact and closeness to the stakeholder. For larger decisions, such as determining if a specific dataset should be open for contribution on Common Voice, we adopt a broad perspective, consulting community members beyond our immediate user base. A case in point is our handling of the Māori language: because we recognize that Māori sovereignty structures do not agree on the use of open datasets, we choose not to localize the language, even upon request from individual Māori speakers. Conversely, for more targeted decisions, like naming specific language variants, we use a narrower consultation model where the Common Voice community manager typically collaborates with top contributors. These ways of navigating prevents community from just being a buzzword or hitting a consensus paralysis."
>
> **— E.M. Lewis-Jong (Mozilla Data Collective)**

Like Lewis-Jong, others also conceptualized community as a multi-layered entity that expands and contracts across various groups, including direct contributors, volunteers, NLP technologists, core organizing members, elders providing profound linguistic guidance, lay language speakers, academic linguists, among others.

Despite these aspects, during our discussions, it became evident that when prompted to define *community*, many technologists adopted a

restrictive perspective. Their definitions were frequently confined to their closest professional circles: other technologists, data contributors, and volunteers with whom they collaborate regularly, exchange technical skills, and develop technology. This narrow focus frequently results in a tech-centric definition of community. There is a clear necessity to broaden this scope to center and empower the actual speakers who serve as the primary stewards and ensure the survival of these languages.

Though various interpretations exist, the predominant understanding of community within this report is anchored in the essential concepts of identity, interconnectedness, and collective action toward a common goal.

Technologists as facilitators

Technologists frequently serve as essential community facilitators within the development of community-centered language technologies. As people who have significant expertise, they often occupy a position of relative privilege within the tech development lifecycle. Their role involves convening community members, surfacing values and needs, leading the technical processes, and frequently managing access to resources required.

Here technologists are themselves members of the local language communities they serve. Their own lived experiences and connections offer them a natural insight into community needs, existing gaps, and required solutions. Although there are power dynamics inherent in these roles, most of the technologists we spoke to expressed a strong commitment to broad community engagement at every stage. They advocate against isolated, individualistic methods, preferring instead to integrate collective perspectives.

It is important to note that a majority of the technologists come from socially dominant groups within their own local contexts. This reflects the broader trend that those with the necessary education, technical skills, and with access to financial resources for this work are often from relatively privileged backgrounds. Although these disparities in positionality are not always drastic and do not necessarily prevent the creation of inclusive local technology, it is important to recognize that they exist, as they shape the priorities and decision-making processes involved in technology building.

Motivations for prioritizing language technology

This subsection examines deeply personal motivations for building technology, showing how people are compelled to combat digital invisibility and critical informational gaps that cause tangible real-world harm to people. Beyond addressing technicalities, there is an underlying desire for communal solidarity and resilience, proving that grassroots commitment can successfully navigate systemic resource constraints. Ultimately, in this subsection, we reframe technology making as a profound act of decolonization, demonstrating that the work is not merely about building tools, but about reclaiming cultural identity, preserving collective knowledge, and actively dismantling the profit-driven, ongoing coloniality of Big Tech.

We asked what motivates people to work on their languages, particularly through technology. We found that motivations are shaped by a diverse array of elements, frequently anchored in people's life stories.

Digital invisibility

There is often a sense, as members of the Global Majority, of not being able *to see ourselves* on the internet or find the information that is needed and relevant to our contexts, cultures and in our languages. As Sadik Shahadu reflects, this digital invisibilisation of Dagbani compelled him to intervene:

> "So for me, I was kind of moved by the fact that my language, Dagbani, is not visible on the internet. I saw that very little information in Dagbani could be found on the internet at large. And so often, I see our people asking why they cannot find resources in Dagbani, why they don't have Google Translate or search support in Dagbani. Just very basic internet needs."
>
> **— Sadik Shahadu (Dagbani Wikimedians User Group)**

Information gaps

For various other technologists, the motivation for their work begins with a similar foundational awareness of the informational gaps in their own lives. This perspective is further strengthened by their recognition of the internet's power to drive transformative change. As mathematician and computer scientist Chris Emezue says:

> "If you are not an American or not a white person, you come from a different community. Let me put it like that. It's very clear to see when the technology is not built for you, like, when it's not recognizing you. When I first saw that, I was shocked. Simple things like NLTK,[13] the most popular engine for many things like semantics, does not support your native language. You're just like, wow!"
>
> **— Chris Emezue (Lanfrica Labs)**

[13] NLTK is a platform for building Python programs to work with human language data.

This non-recognition is not just a cultural grievance, it creates concrete obstacles to accessing essential information in fields like health, education, and agriculture. A clear example occurred during the peak of the COVID-19 pandemic, when people were unable to immediately access vital, life-saving information in their local languages, or languages of choice. This lack of information caused, and continues to cause, serious personal, health, social, and economic-related damages for people.

To be in community

For some participants, systemic adversities served as a primary motivator for their work. Having navigated the intense challenges of war or other social oppressions, these individuals often credit their own communities as vital support systems that made their survival possible. Consequently, their dedication to building technologies for their languages is framed as an act of resilience and a profound commitment to give back to their communities, rather than simply being a professional interest or a personal passion.

Interestingly, at times, the initial motivation for getting involved with language and technology is not actually the vision of the final technological output. Instead, it is rooted in an underlying aspiration to foster and participate in a collective human community. We feel this is especially true of interviewees like Samuel Rutunda who emphasized his motivation to work alongside like-minded individuals, driven by the conviction that complex problems demand collective solutions and that every member contributes unique value. For him, the act of tech making is continuous, natural and flows from social and political everyday relationships.

The drive to continue these initiatives is strengthened by witnessing and experiencing the tangible positive impact they have. The work of grassroots organizations like Masakhane, a community-led collective dedicated to advancing NLP research for African languages, has demonstrated that a strong, community-focused effort can even mitigate many of the challenges posed by limited financial resources, as academic, researcher and co-founder of HausaNLP Shamsuddeen Hassan said:

> "We started in Masakhane, coming together, working. It's not something being paid. It is more like—do you care about your own [people]? Yes? Come let's work together, create the datasets, and now create a solution."
>
> **— Shamsuddeen Hassan (HausaNLP)**

Decolonizing knowledge systems

In the context of Africa, we highlight that post-colonial nations have continued to grapple with the deep entanglement between their pursuit of liberation and the enduring dominance of colonial languages, a legacy that creates what scholar Tigist Shewarega calls a distinct *colonial rupture of identity*.

Languages being at the very heart of indigenous and local identities means that for many language technology communities, the work transcends mere technical development. These initiatives also represent attempts to mend this colonial rupture and counter historical wounds by validating local knowledge systems and populating digital spaces with currently under-represented languages.

These efforts have to be considered in the context of works like "Decolonising the Mind: The Politics of Language in African Literature" by Ngũgĩ wa Thiong'o, where he describes the connection between language, identity and colonialism (Thiong'o, 1986) in these terms:

> "The biggest weapon wielded and daily unleashed by imperialism against that collective defiance is the cultural bomb. The effect of a cultural bomb is to annihilate a people's belief in their names, in their languages, in their environment, in their heritage of struggle, in their unity, in their capacities and ultimately in themselves."
>
> **— Ngũgĩ wa Thiong'o, Decolonising the Mind: The Politics of Language in African Literature.**

By asserting the right to digitally encode their own knowledge systems and ensuring they flourish in a tech-driven future, these language communities are actively shifting the politics of knowledge. They serve to dismantle the contemporary cultural bomb and constitute a direct challenge to colonial structures and therefore, as acts of decolonization.

In contrast to this, the evidence gathered in this study provides an underlying critique of the incentives driving Big Tech. The findings suggest that these corporations are motivated primarily by profit, the consolidation of power through the production of information, and a pervasive disregard for the community needs—all of which perpetuate a state of ongoing coloniality (Quijano, 2024).

Implementing African relational philosophies in technology design

This subsection highlights the emerging integration of African relational philosophies into technology design, demonstrating how language tech communities are rooting their workflows and organizational structures in indigenous interconnectedness and collective ownership. Using Digital Umuganda's communal problem-solving model as an example, we illustrate how traditional practices can directly reshape technical development to ensure local relevance and communal well-being.

At the core of many African relational philosophies is the understanding that human identity, knowledge, and well-being are fundamentally relational and intertwined. Here the individual is not a solitary entity but a node within an intricate web of interdependencies that connects people to one another, to their ancestors, to other species, and to the environment. Some examples of these philosophies include Ubuntu, Umuganda, Sankofa, and Teranga, among numerous others.

In alignment, some language technology communities are attempting to develop models, methods, and build organizations rooted in local social histories, indigenous relationality, and kinship.

An example is Digital Umuganda, which bases its community-centered approach on the Rwandan philosophy of Umuganda, discussed earlier by Samuel Rutunda and translated as *coming together in common purpose*. This traditional practice of regular communal gathering for collective problem-solving and labor is directly integrated into their technical workflow. This methodology ensures that the resulting language tools are not only technically proficient but are also locally relevant, held in collective ownership and inextricably linked to communal well-being. The effectiveness of this community-driven approach can be evidenced by projects like the Mbaza chatbot, which delivered essential COVID-19 information in local languages.

Despite this example, connecting African philosophies with language technologies is still a nascent endeavour. Communities can explore further and design specific strategies that move these relational philosophies from ideals into practical methodologies where the logic of the technology is aligned with the philosophies of the communities they serve.

Data collection, representation and fair compensation

This subsection bridges community ideals with the operational realities of building equitable datasets, proving that data ethics directly dictate technology quality. We highlight the strategic tension between resource-intensive on-ground data collection and cheaper, less nuanced online methods. Furthermore, we confront the steep financial and societal barriers to true diversity, particularly unpacking the pervasive gender gap where women, the primary traditional custodians of language, are frequently excluded from the male-dominated technical pipeline. We also show that fair compensation and representation are not optional technical features, but are the cornerstone of truly representative and fair language technologies.

Online versus on-ground data collection

On a functional level, technologies like spell-checking, speech-to-text, and machine translation operate by identifying patterns within extensive training datasets. Consequently, the acquisition of data remains a critical pillar for technological projects aimed at supporting languages that are digitally under-resourced.

However data is not just about quantity; it is also about quality through the ethics of collection which includes attempts to bring further equity, represent embodied experience, and the ethics of collection, or how the data is collected.

Our interviews reveal that technologists employ distinct strategies for online and on-ground data collection as they work to develop new language solutions.

Direct and intensive engagement between technologists and communities, frequently enabled by local relationships, distinguishes on-ground methodologies. Researchers conduct physical field visits to spend time with community members, gaining a deep understanding of their perspectives, localities and power dynamics. This approach is essential for establishing trust and collecting nuanced, more representative data. This combination of positionality, motivation, and methodology of on-ground data collection offers a powerful means to address and mitigate the systemic racial, gender, and other biases that are often hard-coded into language technology through data.

Asmelash Teka Hadgu, Co-Founder and Chief Technology Officer of Lesan AI, explains:

> "Many people may not know how to use the internet but we don't want to limit our data contributors to only people who have access to it. So we take on these tasks ourselves, putting people on the ground with audio recording devices to reach out to people without internet access especially in remote areas."
>
> **— Asmelash Teka Hadgu (Lesan AI)**

Online data collection, by comparison, emerged largely as a necessity. Although it is currently regarded as a convenient methodology that

provides the required large datasets quickly and serves as a budget-friendly substitute for field-based initiatives which require substantial funding for equipment and travel, it has drawbacks. Notably, online collection encounters obstacles in maintaining the agency of contributors without the support of face-to-face engagement.

Because of this, nearly all technologists expressed a preference for physical, on-site practices. However, the capacity to implement such field-based methods is almost universally hampered by a lack of financial resources.

When African language solutions are informed by a profound, lived understanding of local settings, they yield superior tools that often outperform the generic products released by major Big Tech corporations for these languages. The effectiveness of these community-driven efforts is illustrated by how small organizations like Lesan.ai, an Ethiopian startup, are able to machine translate Amharic and Tigrinya at a level superior to both Google Translate and Microsoft Translator (Deck, 2023). In other NLP spaces, smaller models like LLaMA 3 or Mistral 7B built on localized, community-specific datasets have demonstrated the potential to significantly outperform Big Tech LLMs, outperforming them at a tiny fraction of the operational cost (Alhanai et al., 2024).

Representation and equity within data collection

Data collection efforts prioritize representation through active community involvement even while acknowledging the inherent difficulty in capturing the full breadth of any community. Many technologists identify several critical dimensions for representation, including geographical location, ethnicity, age, gender, caste, tribal

affiliation, and accessibility, alongside linguistic variations like accents and dialects.[14]

The most significant barrier to attaining a deeper level of representation is a pervasive shortage of resources required to identify, involve, and provide incentives for a diverse range of contributors. Ensuring diversity costs resources. For example, in regions with limited connectivity, funding is essential simply to facilitate basic participation by providing internet access, data packages, and hardware. Furthermore, dedicated resources are necessary to design and implement specialised programs that encourage involvement from under-represented groups.

Gender gap in community-centered language technology

The language technology industry, especially at the level of linguists, content developers, technologists, and contributors, continues to be male-dominated. Our research suggests that many community-oriented technologists, even when facing financial limitations, are prioritizing using resources to improve gender balance. These proactive measures include creating specific *women's wings* and establishing award schemes to celebrate *outstanding women contributors*, all aimed at fostering greater female participation in the field.

> "So I think we look at it based on reality. So most of the datasets we create are limited within a finite, let's

[14] See Definitions section

> say timeframe or a budget. So we have to set some priorities. One thing that we really do and we work on is gender representation. We set a hard boundary here. If representation of women is lacking, those are also the technical boundaries we shouldn't go beyond."
>
> **— Samuel Rutunda (Digital Umuganda)**

Samuel Rutunda's perspective highlights a deep-seated commitment to achieving balanced gender representation up to the technical layer. Nevertheless, the level of participation among women continues to be shaped by broader social dynamics. They frequently encounter systemic patriarchal barriers such as limited access to education in technical fields, and the disproportionate burden of domestic and caregiving duties, which results in lower engagement of women compared to men.

When the most visible voices in a community space are men, the resulting technological "standard" can marginalize the lived experiences and linguistic variations used by women, queer individuals, and other oppressed groups.

A prevalent imbalance in gender representation creates a poignant contradiction, particularly in many African contexts where women are recognized as the primary educators, preservers, and disseminators of culture and language for new generations. It is precisely because they are the conduits of linguistic and cultural transmission that some traditional family structures still prioritize marrying within the tribe to ensure children learn their people's heritage. Yet, despite their central role in preserving the language and culture, women are frequently absent from high-level discussions regarding that language's future. This raises a pressing question: how can a language's future be secured while systematically excluding the very people responsible for its survival?

It is essential to recognize that gender is only one facet of this broader issue: numerous other dimensions of representation and access remain largely unaddressed by the language technology efforts highlighted here. These challenges of representation and equity are central to ongoing debates in this field and require further investigation.

Compensation for data contributors

For Chris Emezue, recognizing the embodied, human nature of data is fundamental to creating a language technology development cycle that is both fair and meaningful:

> "I always wanted to try and build something that does something differently by factoring the data providers into the pipeline. And that creates a different conversation and ecosystem. Because if you just say I want to just build and code and then deploy and I'm done. If you're not thinking about 'Okay, this data that I'm using, where did it come from? Who wrote the article that got script by common crawl that got put into C4, that I'm now downloading?', when you think about that, this data becomes more than just an item to throw into an AI, it becomes a story of different people and their connections. I'm focusing on and very passionate about bringing in the data providers as people into the pipeline. And we call that 'data farming'."
>
> **— Chris Emezue (Lanfrica Labs)**

Chris Emezue posits the concept of *data farming* as a counter to *data mining*. He characterises *data mining* as an exploitative, coercive, and

profit-centric method of gathering data for language technology, while *data farming* is envisioned as a process deeply rooted in language communities that centers on ethical accountability and fair compensation, ensuring that the benefits, including career paths, resources, and overall well-being, are channeled directly to the people involved.

Similarly, other language technology initiatives we engaged with expressed a profound awareness of the necessity for equitable community compensation. This includes ensuring equitable pay for data sharing, intellectual labor, and time spent, while also guaranteeing significant participation across every phase of the technical lifecycle. Despite the widespread recognition of these needs, nearly all these efforts face a critical limitation in funding which remains a primary obstacle for fair compensation. This constraint is explored further in the following section.

To fill the resource gaps, we found that people try to compensate in social ways that extend beyond financial payments. For example, for participants who are students from lower socioeconomic backgrounds, there exists an implicit understanding that their voluntary efforts may eventually lead to educational or professional networks or connections, even if they aren't paid monetarily. The issuance of participation certificates is one practical example of this.

While such measures are sometimes necessary in resource-constrained environments, our interviewees almost universally expressed the belief that direct monetary payment remains the most ethical and ideal standard for honoring this labor.

Funding challenges

This subsection exposes the fundamental economic bottlenecks and power imbalances that prevent community-centered language technologies from fully realizing their ethical practices. Local organizations remain deeply dependent on Global North capital, creating a profound funding imbalance that forces grassroots initiatives to compromise on fair contributor compensation, community engagement, and resource-intensive on-site field visits. This scarcity distorts project priorities toward fast, marketable outcomes and cheap digital data collection, which inadvertently encodes biases and reinforces the very digital invisibility technologists aim to dismantle. Ultimately, in this subsection we bridge the gap between ethics and harsh material realities, highlighting a shift toward sovereign funding models, including hybrid commercial licensing and collaborative Pan-African alliances like the Huniki Federation, designed to secure true local autonomy and long-term resource sustainability.

Local technical organizations often serve as the primary facilitators for many language technology communities. Operating under various structures, from non-profit to for-profit, these entities are generally dependent on external financial support largely provided by Global North sources, such as private capital from donors and Big Tech, non-profit organizations, and state-sponsored development agencies.

A profound funding imbalance exists, with African projects receiving significantly fewer resources than equivalent American or European endeavors. This disparity and consequent scarcity of resources often compels African organizations to compromise on community-centeredness like compensations or on-ground work to meet the

specific budgets, goals, schedules, and other requirements dictated by Global North funders.

Funding sources

Funders within the language technology funding ecosystem operate differently and fulfill specific roles:

Developmental and philanthropic partners: Institutions such as the Lacuna Fund[15] and the German Institution, Deutsche Gesellschaft für Internationale Zusammenarbeit[16] (GIZ) have been financing foundational language efforts in Africa, supporting critical work, like the building of datasets, which may lack immediate commercial viability but are essential for the field's long-term growth.

Big Tech: While these corporations provide substantial funding, there are serious concerns regarding the inequitable distribution of these resources, the nature of their partnerships with local technologists, and the potential exploitation of community-generated open-source data for corporate gain.

Grassroots models: Approaches like those practiced by Mozilla, for instance through the Mozilla Common Voice project, use small contributions from many organizations and individuals. This model aims to decentralize power, ensuring that no single funder exerts

[15] Before transferring its leadership and management to partner institutions in Africa and Latin America, the Lacuna Fund had become a significant source of capital, allocating millions toward the creation of African datasets. However, the fund was also noted for an application process that was frequently described as overly complex.

[16] German Society for International Cooperation.

undue influence and that the communities involved maintain a sense of ownership and agency over the resulting technology.

National governmental support: While there is a trend toward coordination among national entities in the African AI space, government funding often has to prioritize voters' immediate needs like schools and hospitals over technology. Consequently, many local technologists remain dependent on external donors for consistent support.

Ultimately, the current funding landscape for this work is characterized by insufficiency in available funding. This shortage of resources has reached such critical levels that some technologists have felt compelled to use their own personal economic resources to ensure the continuation of their work.

The complexity of the funding application process

Our data reveals that the tedious nature of funding applications poses significant obstacles, particularly for smaller organizations and initiatives focused on marginalized languages.[17]

Often people find themselves in situations that reflect a lack of consistency and care from funders. Paul Azunre, founder and Director of Research at GhanaNLP, highlights this burden:

[17] See Definitions section

> "Their process can be really tough. We apply for funding through a very complex process that requires us to submit a lot of documentation, but then they ghost you, don't even reply, then program managers change and we have to start all over again until the next cycle."
>
> **— Paul Azunre (GhanaNLP)**

There is also a preference among funders for supporting established, large-scale organizations. Paul Azunre observed that capital tends to flow toward *"very large brands like Masakhane or Ashesi [University]"*, which can inadvertently centralize language technology efforts and establish hierarchies in the tech ecosystem rather than support a diverse range of local initiatives. To mitigate these imbalances, large entities like Masakhane do use resource redistribution strategies, such as providing subgrants to researchers and smaller organizations that cannot manage the administrative burdens of traditional funding. Nevertheless, the concentration of resources in a few well-resourced structures can continue to limit opportunities for emerging groups.

The challenge of securing funder interest often relies on existing connections and relationships, a dynamic that disproportionally benefits well-known organizations. Chris Emezue highlighted this difficulty by describing the significant time spent *"trying to figure out how to get into the network and play the game"* in order to understand funders expectations. This sense of isolation for smaller groups is symptomatic of a broader funding model where access to resources is gated by human networking opportunities, reducing access to funding to a privileged few.

Community-centered initiatives and Big Tech funding

The relationship between Big Tech and local technologists is characterized by a complex blend of vital support and significant obstacles. On the positive side, corporations like Nvidia and Google have issued grants to foster open-source initiatives and expand African datasets. Although these may serve as profit-oriented ventures for these corporations, such funds are often indispensable in an already resource constricted environment.

At the same time, there is profound concern over the appropriation of community-developed open-source data by major technology corporations. Meta, specifically, has faced criticism for its role as an extractive funder, using this data to create proprietary products while failing to provide communities with appropriate recognition or financial compensation. Such practices raise serious ethical concerns about the exploitation of local knowledge and labor. These tensions are further discussed in the open knowledge section of this report.

The influence of funding on practices

Financial limitations significantly impact the practical execution of language technology initiatives, frequently necessitating a departure from ideal community-driven approaches. A prominent example of this pressure is the difficulty in offering fair remuneration to those providing data. Many initiatives are restricted by their budgets to providing only nominal payments.

> "Regarding NaijaVoices, we were aware from the start that our resources were insufficient to compensate individuals at professional rates for their contributions. We chose to be transparent about this rather than pretending the pay was adequate. We recognized that while we provided some funds, it did not reflect the professional value of the work; one cannot expect professional services for such minimal compensation."
>
> **— Chris Emezue (Lanfrica Labs)**

Moreover, these economic hurdles limit the capacity for essential in-person collaboration and shared learning. For example, while practitioners emphasize the importance of grassroots, local gatherings, existing funding architectures tend to value more and prioritize large-scale international events that many communities may find difficult to attend.

The choice of projects is further distorted by limited resources, which leads to a prioritization of fast road initiatives that produce quick, marketable outcomes. This shift comes at the expense of slow road foundational efforts, which frequently struggle to secure sufficient support despite being essential for long-term sustainability.

Additionally, financial constraints frequently force a departure from fairer and more representative design principles. Integrating diverse perspectives like those of women, rural-origin speakers, and minoritized groups, requires time-intensive, labor-intensive, on-ground engagement that is fundamentally incompatible with the restrictive lean budgets imposed by many funders.

When projects are forced to bypass these community-centered methodologies and directions, this compromises the final quality of the data, producing models that are exclusionary, and reinforcing different forms of digital invisibility that communities seek to dismantle.

Funding models that communities aspire to

The drive for financial sustainability fundamentally shapes how data licensing strategies are chosen. Because conventional funding structures frequently lead to temporary, one-off data collection projects, NaijaVoices adopted a unique hybrid approach utilizing both commercial and non-commercial frameworks, demonstrating the creative ways teams must use to adapt and overcome systemic financial barriers. Chris Emezue explains this model in the following way:

> "We realized that a lot of data efforts are short-lived once you get funding, you create the data, that's the end. But we wanted to create a recurring wheel—you get funding, you create data, you get returns, you create data next year, and so on. So we spent a lot of time thinking about it, and this made us change our licensing scheme. We had to also go back and forth with our funders and ended up with a model where we used both non-commercial open-source licenses and also commercial licenses. For some, you can use it, download it and it's open access. For commercial interests, the commercial organization has to donate to the community. The donation is used to provide jobs for the community to create more data. Everyone benefits. Organizations get their data. More data and jobs are also created. Local people benefit."
>
> **— Chris Emezue (Lanfrica Labs)**

The desire for local autonomy within language technology communities is fueled by the conviction that Big Tech's broad, generalized models often miss critical local contexts. Consequently, these communities are working to transition away from the constraints of top-down funding

originating in the Global North. Technologists, such as Samuel Rutunda, maintain that true innovation stems from those with a profound grasp of these localized, ground-level realities:

> "So the Big Tech can only do things at scale so they will come and they will [try to] do a solution for everyone that's general. But someone, a developer who is studying his graduate studies in Rwanda, they understand the local needs and the realities on the ground. So with that information given that they will have the tool easily, they can create a custom-made solution based on reality."
>
> **— Samuel Rutunda (Digital Umuganda)**

Our data analysis reveals a powerful push and a significant demand from those in the language technology sector for funding frameworks that prioritize local contexts, sustainability, and the reinvestment of resources to foster local economic well-being and social agency. Consequently, numerous community technologists are exploring private sector paradigms to build a resilient industry for this work, seeking to move away from a dependence on grants for non-profits and financial support from Big Tech. While the exact transition path remains under exploration, several pioneering experiments are currently underway.

It also remains to be seen how private entities and funders uphold their accountability to local communities and redistribute resources fairly.

Language technology communities throughout Africa are also exploring the creation of Pan-Africanist federations to facilitate the exchange of financial resources and the development of collaborative strategies for obtaining new resources. Initiatives such as the Huniki Federation aim to harmonize the mechanisms of philanthropic exchange while facilitating the international distribution of resources across different

countries and initiatives. These efforts are in line with the values of The Africa Declaration on Artificial Intelligence, adopted by 52 African nations, driven by Pan-Africanist ideals, ensuring the continent shapes its own digital destiny together independently of the Global North.

Furthermore, there are also urgent appeals from language technology communities for increased local government involvement. State funding has the potential to harmonize local social objectives with language technology work on a larger scale. Many African governments are taking these requests seriously and there is expectation that this funding stream may grow in the years ahead.

Colonial standardization and dialects

This subsection exposes the critical friction between machine learning's technical demand for standardized data and the historical legacy of colonial standardization, which forces local dialects into rigid Western linguistic frameworks. We address why most modern language technologies bypass dialectal diversity in favor of dominant macro-languages, revealing that choosing which linguistic variant to digitize is a deeply political act that can inadvertently exclude non-dominant speakers, a power currently wielded unchecked by Big Tech corporations.

We also unpack a paradox of preservation: the very methodologies used to digitize and protect endangered languages often force developers to standardize data, risking the erasure of the multilingualism and oral traditions they seek to save. Ultimately, we emphasize that technologists must critically examine their own internal biases and adopt community-led, collaborative strategies to ensure language technologies serve as an enabler of linguistic diversity rather than an agent of modern assimilation.

The historical impulse to select and standardize a single language for conformed use has been driven by colonial administrations. European missionaries deliberately selected specific regional dialects to codify into standard systems, prioritizing the pragmatic need to mass-distribute translated Bibles and administer colonial education over the natural, fluid variations spoken by local communities (Makoni & Pennycook, 2007). Alexander Andrason explains:

> "Colonizers wanted to standardize and select one variety because they sought either to spread the Bible or to disseminate their own knowledge. Standardization—whether of Shona, isiZulu, or any other language—largely disregarded internal linguistic diversity"
>
> **— Alexander Andrason (University of Cape Town)**

Andrason also highlighted the fact that linguistic variation represents the norm across African languages.

> "For many African speakers, variation is not normally a problem. The fact that your accent is different, or that certain conjugations or inflections vary, is not perceived as an error or as something wrong. Instead, it reflects that you come from a particular village or region of the country. This prescriptive approach to grammar is largely a product of European and Western philosophical traditions, which tend to privilege uniformity over diversity. As a result, there is often a perceived need to select a single dialect in order to protect the language."
>
> **— Alexander Andrason (University of Cape Town)**

Although linguistic variation is a standard feature of African languages, in our interviews, we have not encountered many language technologists working on dialects; most of them choosing to focus on large, dominant languages instead. This persistent gap is driven by several underlying factors.

The definition of what constitutes a language versus a dialect is seldom defined by linguistics alone; rather, it is a profoundly sociopolitical construct. Popularized by the maxim that *a language is just a dialect with*

an army,[18] these distinctions typically revolve around issues of power, geography, and political recognition. This reality creates complex challenges for technology developers, as determining which dialects to prioritize often carries political weight. Since communities are not homogeneous, disagreements tend to arise over how to structure macro-languages like Kiswahili which encompasses multiple dialects across various regions and countries.

It remains crucial to recognize that communities have different preferences for how their dialect should be supported. The Manding language in West Africa, encompassing varieties like Bambara, Maninka, and Dyula, provides a compelling example of a community choosing cross-border standardization over dialect-specific language technology. Rather than building separate, fragmented NLP processing tools for each regional dialect, developers and digital activists have chosen to use N'Ko, a unifying script invented in 1949 specifically to bridge these dialects (Doumbouya et al., 2023). This strategic preference prevents the fragmentation of limited tech resources, fosters a shared digital sovereignty across multiple West African nations, and helps the Manding-speaking community resist digital marginalization by dominant colonial languages like French.

The consequences of failing to support dialects undeniably carries profound real-world implications, often leading to the socio-economic exclusion of speakers of non-dominant varieties and jeopardizing their general welfare. Regrettably, as Rising Voices Director Eddie Avila points out, the authority to make these critical linguistic choices frequently rests with major technology firms rather than the communities themselves:

[18] Popularized by Max Weinrich, 1946, author unknown

> "Quechua is found in major AI language platforms, but which variant is being used? Those are political decisions, and it seems like those decisions are just made for you. What does that mean for communities that speak a certain variant? Who decides, and how is that communicated? These are important conversations to have, and protocols are needed for how those decisions are made. This isn't unique to Quechua. Many Indigenous language communities are being left out of these decisions."
>
> **— Eddie Avila (Rising Voices)**

Emerging strategies for language diversity

In addressing the intricacies of linguistic variation, some interviewees have developed collaborative frameworks that provide a space for discussion among community members, together with linguists and social scientists. These discussions are often led by strong facilitators who help navigate the complexities of language structuring. As E.M. Lewis-Jong noted:

> "What's hard about that approach is, communities, as I say, are not homogeneous, and they do not always agree on how to structure a language. And so when there is disagreement, we try and hold space for the discussion"
>
> **— E.M. Lewis-Jong (Mozilla Data Collective)**

E.M. Lewis-Jong also mentioned that in certain cases, community consensus on a language-related decision can be prioritized even over

academic linguistic advice. For example, even when a linguist classifies a language as a dialect based on academic considerations, the community's decision to maintain it as a language is respected.

Furthermore, some technologists view linguistic variation not as an obstacle, but as a technological opportunity. Asmelash Teka Hadgu, suggests that for languages like Tigrinya, spoken across two countries and various regions, technology can serve as an enabler to amplify everyone's voice and bridge differences without erasing internal variations, rather than a problem to be solved.

It can also be beneficial to evaluate the specific technological requirements needed for supporting dialects. Dialects with only speech variation, and without variations in script or orthography, can be more easily supported through the development of specialized automatic speech recognition tools. Paul Azunre describes this approach as finding a *"middle ground"* within the intricate realities of dialectal diversity, especially when navigating constraints in funding and available resources.

In addition to the methodological variations and complexities previously noted, Kathleen Siminyu, a researcher at the Distributed AI Research Institute (DAIR), specialised in NLP, highlights a significant challenge in language technology: the dominant methodologies for building speech technology are typically designed for languages that are already established in written form. This creates a significant barrier for many Global Majority languages which remain primarily oral. Developing speech technology for these languages requires a substantial amount of foundational work, such as the creation of formal orthographies, before any speech technology can be built.

The impact of technologist bias

As much as there is a commitment to work toward language justice and representation, we have to come to terms with the fact that technologists themselves may have a personal bias toward their own language variants. These personal preferences can unintentionally lead to developers making unilateral decisions about which language versions are integrated into technology, potentially coding power imbalances that neglect broader community needs. In this case, the developer is an insider to the community and also an insider to the language technology development process, and these intricate dynamics must be acknowledged. Consequently, it is important for technologists to critically examine their own role and positionality throughout the technology-creation process.

The paradox of preservation

Often language digitization and AI development require the standardization of a language. When a language is standardized for machine learning purposes, a substantial amount of linguistic and cultural nuance, particularly dialectal diversity, risks being discarded. To make diverse datasets usable, developers and researchers often normalize the data, specifically through Western documentation methods which require creating dictionaries and standard grammar rules, which can result in an artificial freeze of the language, killing the fluid, natural variation, orality and multilingualism that defined it in the first place.

Resource and other constraints cause people to be forced to select and elevate a single regional variety such as specific dialects within a language continuum. Consequently, the very act of standardizing and protecting the language inadvertently erases its natural internal variation and quality. This is something that communities have to contend with deeply.

Alexander Andrason identifies this as challenge as inherent to language preservation:

> "This is a paradox that is difficult to avoid—and a very sad one. When we seek to preserve, maintain, and promote the use of a language such as Dogon, practical considerations often lead to one or two varieties being selected and given prominence. In protecting the language, we may inadvertently undermine its internal variation. This is extremely difficult to counteract, particularly in Africa, where resources are limited and efforts must often focus on preserving at least some part of this rich linguistic diversity."
>
> **— Alexander Andrason (University of Cape Town)**

Open knowledge repositories (OKRs) as digital ecosystems

This subsection evaluates OKRs, such as Wikimedia projects, as the primary digital infrastructure currently driving language preservation, demonstrating how these open platforms act as both catalysts and areas of dispute for language justice. We highlight how grassroots collectives successfully use OKRs as pipelines to document knowledge, archive media, and train localized AI tools. However, we expose deep systemic limitations, including epistemic friction caused by Western-centric, and text-heavy citation standards that inherently marginalize oral traditions. Ultimately, we underscore that the democratic promise of open-knowledge ecosystems cannot be fully realized until platforms dismantle structural barriers, redirect funding from Big Tech to local creators, and ensure that language communities are able to customize these tools to reflect their own unique epistemologies.

OKRs, such as the Wikimedia projects, have become essential ecosystems for the development of language technologies. For many African communities, these platforms transcend their role as simple digital archives; they provide holistic environments dedicated to the active documentation and revitalization of their linguistic heritage.

The primary value of OKRs stems from their ability to seamlessly integrate content creation with data curation. A notable illustration of this is the Dagbani Wikimedians User Group, which utilizes the

Wikimedia ecosystem as a comprehensive pipeline: they document knowledge through Wikipedia articles, build lexicographical data (Masssly & Abdulai, 2026) on Wikidata, and archive multimedia assets on Wikimedia Commons. By bridging the gap between open-knowledge ecosystems and technical development, this integrated approach facilitates partnerships with local creators, exemplified by the use of GhanaNLP's Khaya app (named after Khaya or the African Mahogany tree) for Automatic Speech Recognition (ASR), thereby turning these repositories into indispensable tools for community use.

Limitations within open platforms

Despite their potential, these ecosystems frequently generate epistemic friction due to their Western-centric and text-based frameworks. Oral language speakers face significant hurdles, as the demand to write in a standardized academic style demands specific resources and expertise. Sadik Shahadu notes:

> "We all speak the language because it's an oral language, but when it comes to writing it, you need another set of skills and energy. So we rely heavily on academic institutions and students who are already learning the language."
>
> **— Sadik Shahadu (Dagbani Wikimedians User Group)**

This tension extends to verifiability. Pau Giner, Principal User Experience Designer at the Wikimedia Foundation, highlights the difficulty of applying traditional citation standards to oral traditions, questioning who defines a reliable source in contexts where written records are scarce.

Importantly, gaps in digital infrastructure limit universal access to OKRs. These technical obstacles—ranging from high data costs to security protocols like IP blocks that, while intended to curb vandalism, inadvertently exclude legitimate contributors in regions with poor connectivity or sensitive political climates. These gaps threaten the fundamental "openness" of these platforms.

Navigating translation gaps

The built-in translation tools within the Wikimedia ecosystem present similar limitations. Pau Giner of the Wikimedia Foundation describes available translation tools as a quick start meant to facilitate initial content creation rather than a one-to-one manual translation. Content created through these tools[19] still must be reviewed by humans and must have citations from reliable sources. Some Wikimedians have already benefited from this, noting the high efficiency of Wikipedia's content translation tools in streamlining the creation of new articles. Mariana Fossatti, a Wikimedian who utilizes these tools for translations between Spanish and English, describes the process:

> "A great tool for translators because you have two columns, one of them with the original article and the other in the targeted language[...] automatically translated with a translation automated tool. And then you can edit the translation easily and publish an article easily. It is a simplified process with all the things you

[19] We note that the choice of source language is also an epistemic justice issue, as it embeds specific biases and perspectives into the content. For example, translating an article in English into Kiswahili preserves an western-centric framing in contrast to one that is translated from Kiswahili to English. Thus, the direction of translation fundamentally alters how a subject is understood.

> need at hand, the original text, the translation, and the interface to edit the translation and to publish the translation."
>
> **— Mariana Fossatti (Whose Knowledge?)**

However, it is notable that many other languages do not share this seamless digital experience. For these language groups, digital inclusion is still advancing through different phases, including font and script development, Unicode recognition, and the creation of custom keyboards along with other essential infrastructure.

As mentioned earlier, bringing languages into digital environments often requires a formal standardization process, particularly for oral languages that must develop technical vocabularies to support their own linguistic frameworks on digital ecosystems like OKRs. To find culturally resonant concepts, some interviewees recommend drawing from the daily lives of the community. For example, Sadik Shahadu notes that in Dagbani, the absence of a direct term for file, as in a computer file, led contributors to examine ancestral methods for protecting valuables. They explored terms such as Kpanjɔɣu or Kɔligu, a type of leather pouch, and eventually repurposed Kɔligu to represent a digital file. Developing sustainable infrastructure requires that technologists focus on localizing terminology within specific cultural settings, avoiding a reliance on overused transliterations.

> "When we started doing the translation on Translate Wiki, we encountered challenges where the committee had to argue and agree on a particular meaning of a word. You know, it's just because the language is an oral language. We don't have these technical terms translated anywhere. So as a committee, we've been able to, I would say for the first time, create many

> translations or definitions for most-used Wikimedia terms and messages on MediaWiki."
>
> **— Sadik Shahadu (Dagbani Wikimedians User Group)**

The unavailability of translation tools for most of the Global Majority languages remains a primary obstacle for content creation. Amrit Sufi, for example, highlights that machine translation for Angika is unavailable both in external services and within the Wikimedia ecosystem. Moreover, African technologists report that even when such tools exist, they can be misleading; despite user-friendly interfaces, the linguistic quality is often poor and flawed, failing to capture essential cultural nuances or accurately convey intended meanings. Consequently, some argue that leaving an article untranslated is preferable to publishing unedited machine translations. This suggests that for many languages, the translation efficiency remains low.

Pau Giner noted that the Wikimedia Foundation doesn't develop its own machine translation models due to the intensive computing requirements. Instead, it offers a central content translation service that brings together existing machine translation tools from both community and corporate organizations.

However, the Wikimedia ecosystem is currently facing a call for transformation, with growing advocacy for the adoption of more community-developed tools and the implementation of equitable compensation models. Paul Azunre criticizes the current practice where some funds are allocated to Big Tech (like Google Translate and Facebook models) for what he describes as *inferior solutions*. He contends that, whenever possible, financial resources should be directed toward local communities as payment for their expertise and data, instead of offering *handouts*.

The Wikimedia Foundation is currently collaborating with various communities to tackle these challenges. Recognizing the limitations of external tools, organizations are increasingly engaging in active feedback loops with platforms like Wikipedia to adjust thresholds, provide human-validated corrections, and ensure that communities retain the right to disable any tools that fail to meet their specific linguistic needs.

In the end, Wikimedia projects and similar OKRs represent powerful, if imperfect, sites for contestation. Their success as tools for language justice depends not just on the volume of content, but on the ability of local communities to customize these tools to reflect their own epistemologies, maintain agency over their data, and ensure that the "open knowledge" they contribute supports local well-being rather than merely supplying data for international AI development.

Redefining open knowledge

In this subsection we deconstruct the assumption that open data is inherently equitable, exposing how Western-centric frameworks of openness frequently act as a gateway for Big Tech extraction and corporate openwashing. While the open-source movement successfully democratized digital access, it failed to redistribute structural power, often relying on the privileged labor of uncompensated volunteers while allowing multi-billion-dollar corporations to harvest communal data for proprietary profit. We outline a paradigm shift away from extraction and toward community-centered stewardship and data sovereignty, championed through innovative, indigenous licensing frameworks like Kaitiakitanga, NOODL, and Esethu. Ultimately, we reframe language data not as a free-for-all digital commodity to be mined, but as an embodied, collective knowledge system that requires protective legal boundaries to ensure technology serves community well-being rather than Global North corporate monopolies.

Financial obstacles are frequently intensified by Global North funders who impose inflexible requirements for open knowledge results. We note that some practitioners feel that the standard definitions of *openness* pushed by funders can inadvertently restrict a community's need to develop sustainable and equitable licensing frameworks. In this regard, E.M. Lewis-Jong argues:

> "I think that some funders still do insist on a particular version of open. I worry that it's coercive without

> meaning to be. If you want to create African language datasets, you have to open source them, [that] is roughly the line. I do think that's problematic."
>
> **— E.M. Lewis-Jong (Mozilla Data Collective)**

It is also a reality that community-based language technology development relies heavily on open-source ecosystems, particularly for under-resourced languages of the Global Majority who are entrenched in efforts towards digitized data. However, while the open knowledge movement successfully established the sharing of culture and information as a universal right, it has now reached a critical juncture needing deeper deliberation as broader access to open content alone has failed to redistribute power. The underlying challenge is that dominant entities in the Global North continue to maintain control over digital knowledge, even within systems labelled as "open". As Mariana Fossatti articulates:

> "I think that the open knowledge, open culture, activism and advocacy in certain ways reached its goal, because access is broader now, but at the same time, have we really won the most important part of the battle? Who is the producer? Who is the holder? Who has control over our culture? That didn't change easily with the open knowledge movement"
>
> **— Mariana Fossatti (Whose Knowledge?)**

Another critical point raised by community technologists is that many contributions under open-source licenses, especially within OKRs ecosystems, are a *product of privilege* rather than purely noble intentions. Contributors with stable employment have the luxury to volunteer time, while those with fewer resources are often expected to contribute labor without compensation. As Subhashish Panigrahi, an open knowledge language contributor and founder of OpenSpeaks, observes:

> "And if somebody thanks me for the open work I have put in, I'd be very happy for that. But at the same time, one also should not take away the fact that I have certain privileges. And if I didn't have those privileges, I wouldn't have contributed. And that's the same case for many people that contribute to open projects as volunteers"
>
> **— Subhashish Panigrahi (OpenSpeaks)**

Big Data, extraction, and openwashing

The persistence of power imbalances in language technology necessitates a shift in focus from open data rhetoric to the underlying legal structures of AI. Governance in this field is defined by a multi-layered licensing framework: it begins with the software code (typically under MIT or Apache licenses), moving to training data (frequently embroiled in fair use or copyright conflicts), it then incorporates specialized AI model licenses such as OpenRAIL, and concludes with provider Terms of Service for the resulting outputs. This structural complexity across code, data, models, and outputs result in a fragmented legal patchwork. Consequently, identifying the true owner of a digital product is often obscured, which in turn presents significant hurdles for the advancement of equitable technology.

Big Tech corporations have skillfully exploited these intricacies of licensing to build multi-billion dollar industries. In their early stages, many companies working with language technology utilized permissive open-source licenses to develop their foundational infrastructure, effectively capitalizing on the unpaid contributions of countless independent developers. At present, these same corporations leverage

openly licensed data from platforms like Wikipedia and Reddit for training their proprietary models, yet they often decline to release the subsequent technological advancements under similar open-access conditions.

Similarly, corporations such as Meta and Mistral employ strategic openwashing as a method to market their offerings as collaborative or transparent, despite keeping them under strict proprietary restrictions. Often utilized as a tactical move against competitors like OpenAI, this approach involves the use of selective open licensing to build vast developer networks around corporate models, effectively repurposing community-centered efforts to fuel the growth and expansion of Big Tech. This strategy has led to significant frustration among community technology groups, who observe Big Tech firms misappropriating their localized, high-quality open models to mirror their effectiveness in closed, commercial products, sometimes within a matter of days.

Mariana Fossatti describes this environment as a "wild competition in artificial intelligence", where an urgent need for training data drives corporations to harvest content by any available means. This often involves the opaque collection of protected data and the use of coercive Terms of Service to capture user communications, a process that disregards creator's rights and compromises individual privacy.

Transitioning toward stewardship

Community members increasingly emphasize that openness requires a strategic, nuanced, and cautious approach. This perspective is driving a significant transition toward community-centered stewardship, evidenced by the emergence of specialized licenses, such as Kaitiakitanga

(Te Hiku Media, Aotearoa), the Nwulite Obodo Open Data License—NOODL (Masakhane & African NLP Researchers), and Esethu (Lelapa AI, South Africa).

These indigenous frameworks prioritize community guardianship and data sovereignty and are purposely constructed to stop the exploitative harvesting of information from marginalized or native groups. Unlike Western legal systems that view data as a personal, transferable commodity, these systems define data as a shared, collective resource fundamentally tied to a people's heritage, geography, and selfhood.

While these protections are established at the point of data collection, they are legally designed to maintain oversight as that information evolves into language models and subsequent outputs. This structure ensures that the ethical standards of the community are embedded into the technical infrastructure itself.

Communities must also grapple more explicitly with the complex consequences of bringing their languages and knowledge online, especially in the current environment of surveillance and extraction, as the online presence of a language allows new social and ethical challenges to emerge. Communities must reflect if they are prepared to face these challenges or if they want to exercise their right to Active Refusal.[20]

Ultimately, the success of the open movement must no longer be measured merely by the volume of knowledge released. As Mariana Fossatti argues, we must recognize that knowledges are not just digital information; they are people,they are embodied, and they are a reflection of power.

[20] See Definitions section

Carving a path towards digital sovereignty

This final subsection outlines a practical path toward digital self-determination, shifting the paradigm from state-level territorial control to a radical framework of community-in-the-lead sovereignty. It addresses the infrastructural deficits that force Global Majority technologists to adopt open data ecosystems out of pragmatic necessity, while bypassing Western requirements through decentralized, collaborative Pan-African alliances like the Huniki Federation. Ultimately, we synthesize the technical, political, and ethical threads of the previous pages into an actionable roadmap; demonstrating that true language and digital sovereignty is a fundamental act of decolonization and epistemic freedom.

Digital sovereignty remains a complex term with many definitions. In this context, and within post-colonial frameworks, it is frequently defined as a state's right to govern its own data and digital futures.

For the Global Majority, a demand for digital sovereignty often develops as a reaction to the Global North treating us as a *data mine*. Within this framework, digital sovereignty is often equated with territorial control, emphasizing the need for domestic control over physical infrastructures. This includes servers and hardware alongside robust legal structures to ensure data governance remains within national boundaries.

However, the pursuit of digital sovereignty is a global phenomenon, not limited to Africa or the Global Majority. This is clearly demonstrated in the

European Union, where approximately 70% of the cloud infrastructure market is dominated by three major providers: AWS, Microsoft Azure, and Google Cloud (Gineikytė-Kanclerė et al., 2025). Despite these common challenges, the struggle for sovereign infrastructure remains a critical priority for many nations seeking to protect their economic security and maintain political independence.

From our results, however, it is obvious that sovereignty cannot be reduced to a matter of infrastructure and geographic borders alone. In fact, the Global Majority is actively re-imagining digital sovereignty through the lens of indigenous and relational philosophies, which characterize it as the fundamental right of communities to own and oversee the use of data related to their people, territories, and resources.

Pragmatism and the African digital context

For many African technologists, the concept of digital sovereignty is processed with a degree of cynicism and suspicion. This skepticism is rooted in the practical realities of a severe lack of foundational data and essential large-scale infrastructure combined with insufficient enforceable legal mechanisms. As a result, most African technologists view maintaining open data practices not as an ideal, but as a pragmatic necessity for the advancement of their work.

However, they are not waiting for a single, perfect model of sovereignty to emerge. Instead, they are exploring new forms of self-determination, as we mentioned earlier, leaning towards a Pan-Africanist approach of sharing strategies, ideas, and resources. Asmelash Teka Hadgu expands on this:

> "And you know, we are coming together [...] towards that effort. For example, through this Federation called the Huniki Federation, it is a small exercise. What we're doing is like startups. Let's say, Lesan, around East Africa and GhanaNLP in West Africa and Lelapa in South Africa are kind of joining forces to say, we're all independent, but let's also have APIs that interact together. So if somebody wants to serve the African market, they could hit that one single API. You can have dozens of organizations sharing API in this way so that if you wanted to have a Wikipedia for each of these African languages, there would be just one AI African technology doing it. But each community is doing it in its own way, in whatever way works for them. That vision is compelling to me. And yeah, this language sovereignty [idea] kind of ties to that, because we could do this as a whole African community."
>
> **— Asmelash Teka Hadgu (Lesan AI)**

Towards community-in-the-lead

We find that to move towards digital sovereignty, we must shift the focus from the concept of a *human-in-the-loop* to what Natasha Malik has called *community-in-the-lead* (Malik & Malik, 2025). Ultimately, sovereignty within the African context is linked to the assertion of decolonization and the exercise of epistemic freedom and therefore the right to build, theorise, and develop technology unencumbered by Eurocentrism (Ndlovu-Gatsheni, 2018).

We believe true sovereignty in the field of language technology will not be found in imitating the North, but in the radical act of communities being empowered to imagine and build technology in their own ways.

4.

Conclusion



Throughout this study, we have moved beyond the critique of extractive commercial frameworks to document the vibrant, on-the-ground reality of community-centered technological development. We have seen that language technologies are never neutral; they are embodiments of the power, ideologies, and values of those who build them. By prioritizing the relational, the local, and the sovereign over the extractive and the standardized, African practitioners are proving that alternative technological futures are an ongoing reality.

We approach AI as a set of computational methods shaped by human money, power, and ideologies. We center the human choices and labor behind these systems, and focus on specific language-centered technologies like NLP, which offer critical ways for preserving and expanding the languages of the Global Majority.

We have explored the systemic frictions inherent in this work—from the funding bottlenecks and the paradox of preservation to the epistemic friction of text-heavy digital ecosystems. Yet, we have also seen that these challenges are being met with community innovation and collaborative, Pan-African alliances. Community-centered language technology initiatives demonstrate that digital sovereignty is about the right to build, theorize, and interpret the world from one's own ontological foundations.

The findings presented here make clear that true knowledge and tech justice requires more than the inclusion of Global Majority languages in existing, Western-centric tools and platforms. It demands a fundamental restructuring of power. It calls for funders to transition from extractive grant-making toward long-term, equitable investment; for policymakers to recognize digital language infrastructures as a public good, especially from and for a "public" comprising the minoritized majority of the

world;[21] and for technologists to embrace the profound responsibility of stewardship over collective cultural heritage and knowledge systems. In our "Community-Centered Language Technology: Best Practices Guideline", we incorporated a core set of principles designed to guide the development of more effective and equitable language technologies. Crucially, these insights will also support our partners at the Distributed AI Research Institute (DAIR) in advancing their machine learning initiatives. As we enter the next phase of our joint work, our focus will center on building language tools tailored for dialects and regional variants, while exploring their technical execution, usability, and epistemological integration into platforms like Wikipedia. We believe both the practical tools and the underlying philosophical explorations are essential for supporting Global Majority communities in creating content rooted in their own knowledges, cultures, and histories.

Ultimately, this report serves as evidence to the resilience and creativity of the Global Majority. While the digital landscape has long been shaped by histories of coloniality, erasure, and standardization, the futures of the internet are not yet fixed. Through the leadership of the communities documented here, we see the road map for just futures—one where technology is no longer a tool of assimilation or extraction, but an instrument of epistemic justice and freedom, supporting communities to imagine, build, and govern their own digital lives.

[21] See Definitions section

Definitions

Throughout our work we used certain terms in specific ways. We began with the development of a Language Justice set of definitions for our foundational report, the State of the Internet's Languages. Building upon that initial vocabulary, we now expand this lexicon to define the following key terms featured in this report:

Active Refusal

Active refusal is a deliberate political stance taken by marginalized or Indigenous communities to withhold their language data from technological collection, serving as a counter-strategy against the assumption that AI development is inherently beneficial to all. These communities assert their data sovereignty, choosing instead to protect sacred or traditional linguistic practices from exploitation, scraping, digitization and standardization into Western datasets. This stance reclaims power by challenging techno-solutionism and insisting that not all knowledge should be open-access or digitized, prioritizing community autonomy over the flattening of language into commercial products.

Digitally under-resourced languages

Digitally under-resourced languages, commonly of the Global Majority, are those that have been marginalized or excluded from mainstream technological development and digital presence. This framing shifts the focus away from "low-resource languages" and therefore any perceived deficit of the language itself, and instead, acknowledges the histories of colonialism and other structural inequities that now result in the systemic exclusion of these languages from the digital landscape.

Dominant languages

Languages that are either the languages spoken by the majority of the population in a certain area, or that dominate through specific forms of historical power and validation, through legal, political, or cultural forces. For example, Hindi is a dominant language in South Asia, in comparison with many other languages, especially considering that Hindi itself is a family of languages or what some call “dialects”. Similarly, Mandarin Chinese is a dominant language in China, through government policy, in comparison with other forms of Chinese as well as other Indigenous languages in the region. Some dominant languages are also “official” or “national” languages in a region or country.

European colonial languages / Eurocolonial Languages

Languages from Western Europe that spread across Africa, Asia, the Americas, the Caribbean and the Pacific Islands through the processes of colonization by Western European companies and governments, from the 16th century onwards. These include English, Spanish, French, Portuguese, Dutch, and German. It’s important to note that these languages were also “colonizer” languages for the Indigenous peoples of North America, not only Latin America (Central and South America).

Global Majority

In this report, we use largely the term Global Majority, originated by Afro-Caribbean scholar Rosemary Campbell-Stephens, to denote the world's minoritised majority. This designation includes women, Black, Indigenous and other people of color, the caste-oppressed, LGBTQIA, and the disabled, often concentrated in, but not restricted to, formerly colonised regions. We find this term to be more specific and inclusive than alternatives such as BAME (Black, Asian and minority ethnic), POC (people of color), or Global South.

Languages and dialects

We consider any structured system of expression between humans, whether by voice, sound, sign, gesture, or writing, to be language. Some linguists define a "dialect" to describe what sound like different varieties of the same language, but that may be "mutually intelligible"—understood by all speakers of these different varieties who can then talk to each other. However, because the difference between how "language" and "dialect" is defined most commonly is a political (rather than linguistic) choice—based on historical processes of power and privilege—we have rarely used the term dialect in this report. We prefer to use the term "language family", showing that there are many languages that may have similar histories but that also have different characteristics, like the family of languages that is Arabic, Chinese, or Hindi.

Local languages

In this report, we adopt a definition of local languages that moves beyond the number of speakers. Our framing is informed by the work of scholar Geneva Smitherman, who views these languages not as regional curiosities but as historical products of resistance and survival (Smitherman, 1986), as well as Sinfree Makoni's conceptualization of "disinventing" the concept of language itself to truly understand local linguistic realities, without relying on the same colonial classification systems that created the term "local" in the first place (Makoni & Pennycook, 2007). Within this framework, we use the term "local languages" loosely to describe predominantly oral languages that are deeply woven into the social practices and cultural fabrics of specific geographies. Most importantly, we use the term to challenge the non-contextualized use of "low-resource languages", and to introduce necessary nuance, allowing us to acknowledge and address the specific historical processes, such as colonial capitalism, that have created linguistic disparities and the state of under-resourcing of languages throughout the Global Majority.

Marginalized languages

In this report, marginalized languages are those languages that are not prominent on the internet in terms of language support or content, i.e. information and knowledge in that language. These languages are marginalized by historical and ongoing structures and processes of power and privilege, including colonization and capitalism, rather than by the population or the number of speakers. Some marginalized languages are already endangered in the world (like many Indigenous languages). But some marginalized languages are spoken by a significant majority of peoples in their region or the world, and yet under-represented online (like Punjabi and Tamil, or Hausa and Zulu, as just a few examples from many dominant languages in Asia and Africa).

Minoritized majority of the world

Historical and ongoing structures of power and privilege result in the discrimination and oppression of many different communities and peoples, across the world. These forms of power and privilege are often interlocking and intersecting, so some communities are disadvantaged or oppressed in multiple ways: for instance, by gender, race, sexuality, class, caste, religion, region, ability, and of course, by language. Whether online or in the physical world, these communities make up the majority of the world by population or number, but they are often not in positions of power, and therefore they are treated like a minority. In other words, they are the “minoritized majority” of the world.

Minority and majority languages

Minority languages are those spoken by a minority (in terms of numbers) of the population, in any described territory or region, while a majority language is spoken by the majority (in numbers) of that population.

People-first design

A people-first design methodology prioritizes human agency and community needs machines. It ensures that technical systems meet community needs, emerge from local lived realities and knowledges, and maintain accountability to community.

Scriptism

Scriptism is an epistemological bias rooted in Western perspectives that elevates writing as the "ideal" form of linguistic representation, despite the reality that most of the world's languages are orally dominant. Scriptism positions oral cultures as deficient, driving an impulse to reduce languages to standardized written forms.

Sovereign language technologies

Sovereign language technologies are sociotechnical systems that treat language as collective cultural heritage rather than an exploitable market commodity. They use a “people-first” design which integrates local social histories, politics and epistemologies directly into the design and technology. Sovereign language technologies are working towards or ensuring that language communities ultimately hold absolute authority over how their digital assets are stored, governed and used.

Gratitude

We are deeply grateful for our partnership with the Distributed AI Research Institute (DAIR), especially Kathleen Siminyu and Hellina Hailu Nigatu. Collaborating with them and the DAIR team has been essential in grounding our research in community-rooted, interdisciplinary AI practices.

We extend our special thanks to Tigist Shewarega Hussen, our Strategic Advisor and co-researcher. Her crucial participation and insights during outreach, and data collection, were instrumental in shaping the direction and depth of this work.

We extend our deepest gratitude to the Whose Knowledge? team, especially our Director and Co-founder Anasuya Sengupta, and to all the communities, friends, and colleagues who have supported this work through encouragement, feedback, and shared labor. Thank you for your continued solidarity and partnership in the journey toward language justice.

We also want to thank the translators and proofreaders whose dedication ensured that our research respects the linguistic diversity we advocate for, helping us reach wider communities.

Finally, this report is profoundly indebted to the many interviewees who generously shared their time, expertise, and lived experiences. Their insights are the heart of these findings. We feature some of them below:



Alexander Andrason
University of Cape Town



Amrit Sufi
Wikimedian



Anushah Hossain
Script Encoding Initiative (SEI)



Asmelash Teka Hadgu
Lesan AI



Awa Ly
GalsenAI

Digital Umuganda



E.M. Lewis-Jong
Mozilla Data Collective



Eddie Avila
Rising Voices

EthioNLP



Kalika Bali
Microsoft Research India



Kathleen Siminyu
Distributed AI Research Institute (DAIR)

Lanfrica Labs



Mariana Fossatti
Whose Knowledge?



Meseret Alemayehu
RDLS Ethiopia



Pau Giner
Wikimedia Foundation



Paul Azunre
GhanaNLP



Sadik Shahadu
Dagbani Wikimedians User Group



Seid Muhie Yimam
EthioNLP

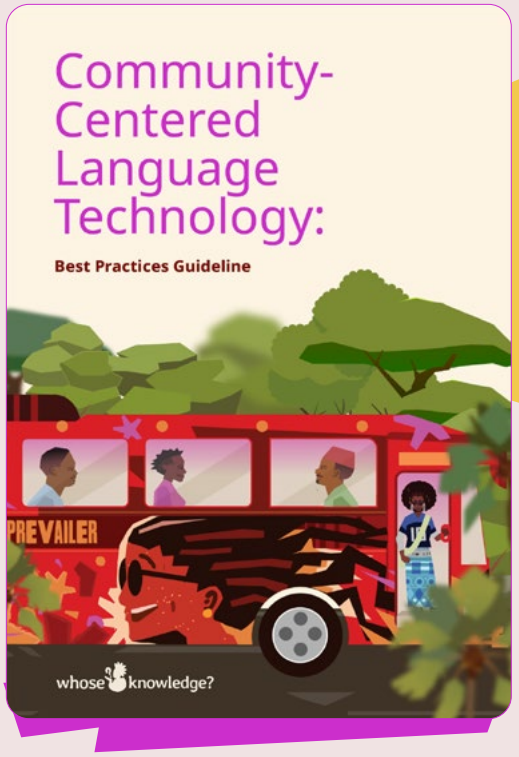


Shamsuddeen Hassan Muhammad
HausaNLP



Subhashish Panigrahi
Openspeaks

Other Resources



[Community-Centred Language Technologies: Best Practices Guideline]

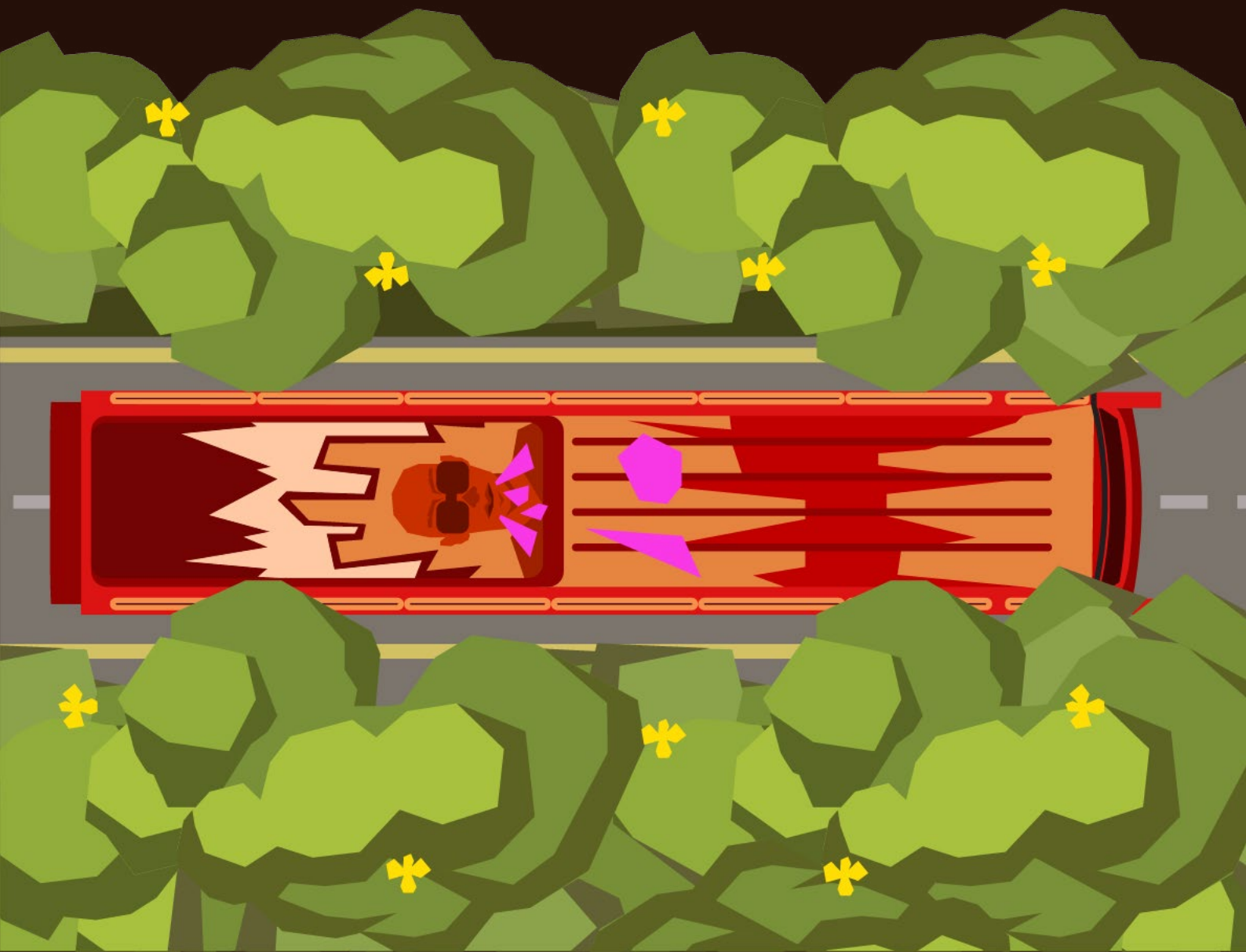
[Community-Centred Language Technologies: A storytelling experience]



[langtech.whoseknowledge.org]

5.

References



Alhanai, T., Kasumovic, A., Ghassemi, M., Zitzelberger, A., Lundin, J., & Chabot-Couture, G. (2024). Bridging the gap: Enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27), 27802–27812.

Bird, S. (2022). Local Languages, Third Spaces, and other High-Resource Scenarios. *Proceedings of the 60Th Annual Meeting of the Association for Computational Linguistics*, 1, 7817–7829.

Chowdhury, J. S., Wahab, H. A., Saad, M. R. M., Mathbor, G. M., & Hamidi, M. (2023). *Research Is Not Greater Than the Methodology: Ways of Seeing and Analyzing Ubuntu*. 31–71.

Deck, A. (2023). The AI startup outperforming Google Translate in Ethiopian languages. In Rest of World. https://restofworld.org/2023/3-minutes-with-asmelash-teka-hadgu/

Doumbouya, M. K. B., Diané, B. M., Cissé, S. F., Diané, D., Sow, A., Doumbouya, S. M., Bangoura, D., Bayo, F. M., Condé, I. S., Diané, K. M., Piech, C., & Manning, C. (2023). *Machine translation for NKO: Tools, corpora and baseline results*. https://arxiv.org/abs/2310.15612

Gineikytė-Kanclerė, V., Eggert, M., & Skiotytė, G. (2025). *European software and cyber dependencies*. European Parliament. https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/780413/ECTI_ATA(2025)780413_EN.pdf

Hao, K. (2025). *Empire of AI*. Penguin.

Harris, R. (1980). *The language-makers*. Duckworth.

Kornai, A. (2013). Digital Language Death. *PLoS ONE*, 8(10), e77056. https://doi.org/10.1371/journal.pone.0077056

Kovacs, A., Fuller, J., Moyo, M., Schroeder, Z., & Omar, N. (2025). *Embodied Data and AI: a Collection of Essays*. Research ICT Africa.

Makoni, S., & Pennycook, A. (2007). *Disinventing and Reconstituting Languages*. Multilingual Matters.

Malik, Z., & Malik, N. (2025). *Episode 13: Responsible AI and participatory design: Discussing the Canadian AI regulatory landscape with Natasha Malik* (No. 13) [Podcast]. The Intersections Podcast by McMaster University. https://theintersections.ca/podcast/#

Masssly, & Abdulai, M. S. (2026, March 6). *Why Dagbanli Needs a Dictionary (And How We Built One with Wikidata)*. Wikimedia Foundation. Diff. https://diff.wikimedia.org/2026/03/06/why-dagbanli-needs-a-dictionary-and-how-we-built-one-with-wikidata/

Mbembe, A. (2003). Necropolitics. *Public Culture*, 15(1), 11–40.

Ndlovu-Gatsheni, S. J. (2018). *Epistemic freedom in Africa : Deprovincialization and Decolonization*. Routledge, Taylor and Francis Group.

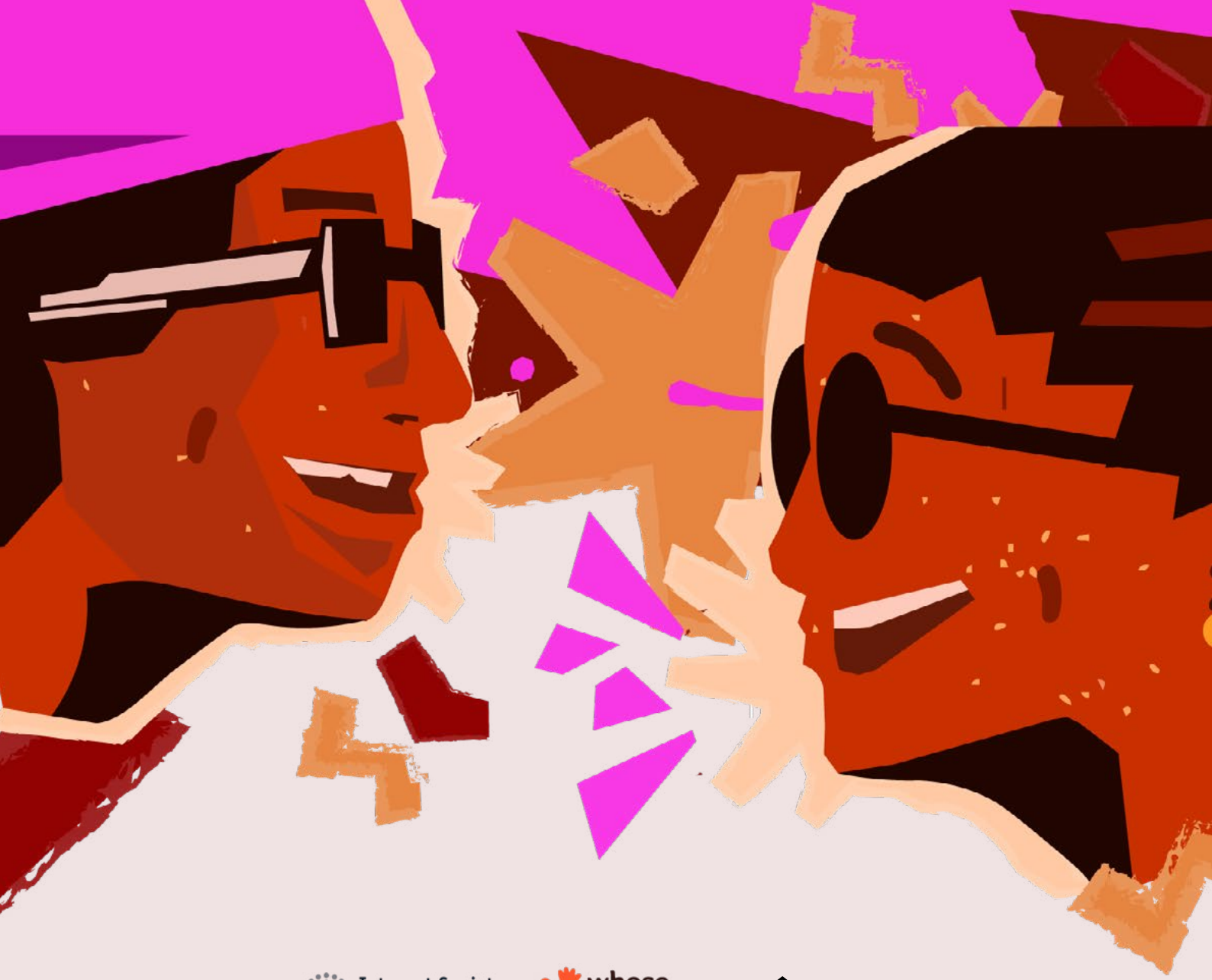
Quijano, A. (2024). *Foundational Essays on the Coloniality of Power*. Duke University Press.

Shadel, J. (2025, March 16). *How to hype-bust "AI" bullshit: A curious reader's guide with the pulitzer center's Karen Hao*. ESC KEY .CO. https://www.esckey.co/how-to-hype-bust-ai-bullshit-a-curious-readers-guide-with-the-pulitzer-centers-karen-hao/

Smitherman, G. (1986). *Talkin and testifyin : The Language of Black America*. Wayne State Univers. Press.

Thiong'o, N. W. (1986). *Decolonising the Mind: the Politics of Language in African Literature*. James Currey.

Whose Knowledge?, Oxford Internet Institute, The Centre for Internet and Society, & Community Members. (2022). The State of the Internet's Languages. In *https://internetlanguages.org/en/*



Internet Society Foundation

whose knowledge?

DAIR

