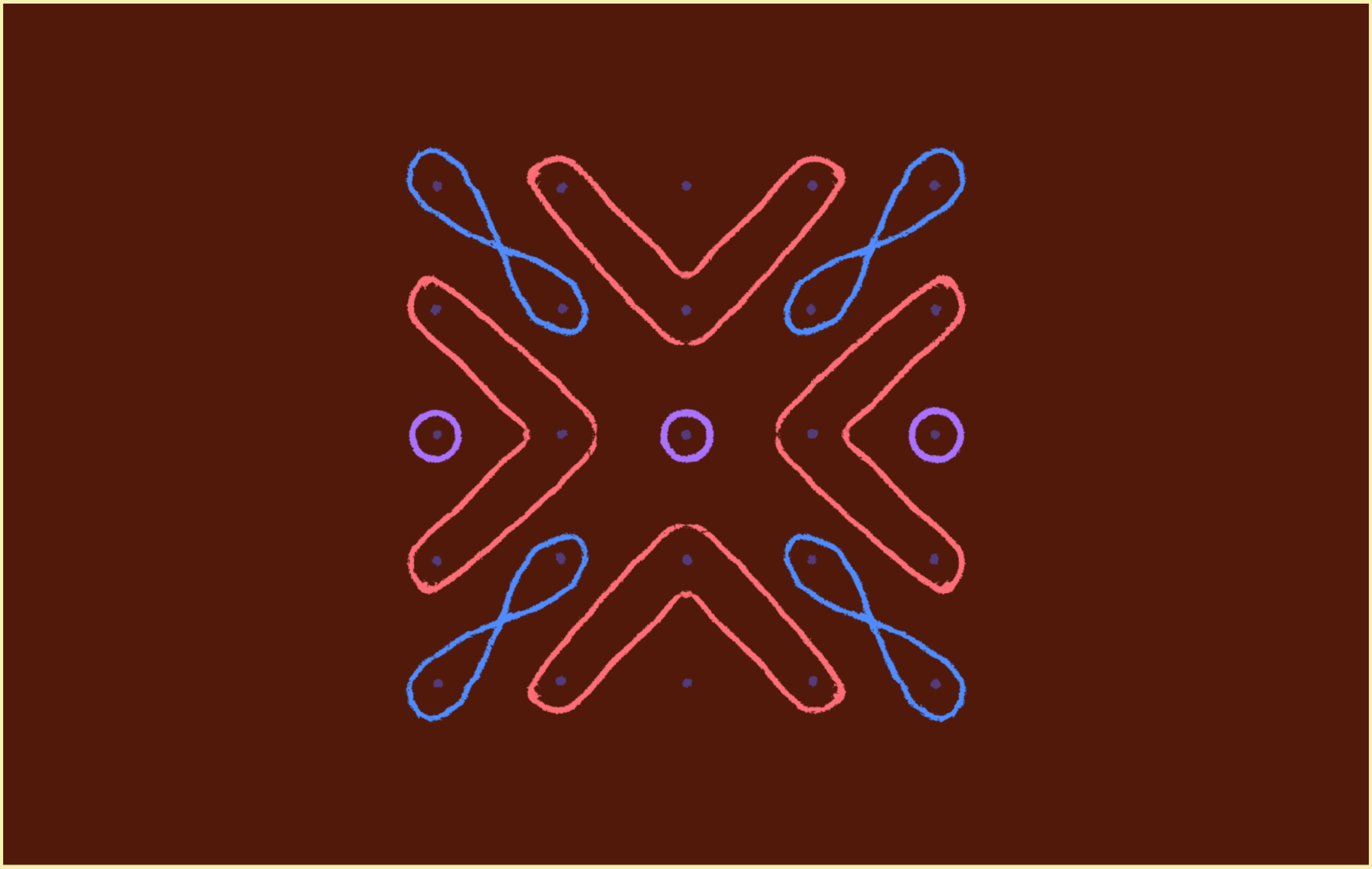
Reframing Impact:
AI Summit 2026

Linguistic Diversity

Chenai Chair

February 2026



This piece is part of Reframing Impact, a collaboration between AI Now Institute, Aapti Institute, and The Maybe. In this series we bring together a wide network of advocates, builders, and thinkers from around the world to draw attention to the limitations of the current discourse around AI, and to forge the conversations we want to have.

In the run-up to the 2026 India AI Impact Summit, each piece addresses a field-defining topic in AI and governance. Composed of interview excerpts, the pieces are organized around a frame (analysis and critique of dominant narratives) and a reframe (provocations toward alternative, people-centered futures).

Chenai Chair leads the Masakhane African Languages Hub. She founded My Data Rights (Africa) and has led initiatives at Mozilla Foundation (Africa Innovation Mradi, Common Voice: African Languages), the World Wide Web Foundation's Gender and Digital Rights flagship, and Research ICT Africa's youth and gender research. She brings feminist perspectives to data governance and AI ethics.

In this conversation, Chair addresses the current discourse around linguistic diversity in AI. She argues that the push to build new, varied linguistic datasets is being promoted by players with vested interests to access new markets in the Majority World. But this approach comes with political risks and sidelines long-standing actors in the ecosystem. In its place, Chair calls for a collaborative, bottom-up approach that centers communities. Drawing lessons from Masakhane's experiences, Chair offers insights on how to include communities in such endeavors—and what to do when they refuse. The way forward, for Chair, lies in building on existing initiatives and sharing resources.

Following is a lightly edited transcript of the conversation.

FRAME: We are in the midst of a rush for linguistic diversity, driven by players with vested interests trying to access new markets. In the absence of safeguards and governance, this scramble presents political and social risks while sidelining long-standing efforts.

A lot of people who do not speak the most recognized languages, like European and popular American languages, have always wanted their own languages recognized. People fought to create datasets of their own languages. When we were talking about AI pre-Covid, we were laughed out of rooms, particularly on the African continent, because people were like, “We have a digital access issue.”

Our argument five years ago was that tech companies have the data, they care for markets with gains. Now, in 2026, all of a sudden linguistic diversity is a hot topic. Now the market is opening up. We’re seeing the reaction of “no person left behind” as we saw with the Internet for All and the Mobile Phones for Development movements. With each new technological development and hot topic, particularly in the development ecosystem, I always say, “Follow the money.” If you follow the money, you will see why there is this investment.

Without safeguards and governance in place, digitizing languages can have severe repercussions.

We are moving towards another level of data which is language—and language is also personal identity. To what extent are we thinking about the safeguards involved with curating this data? What’s the point of value extraction for the people who are providing these language datasets? Are they the ones actually building the technologies?

The rush to build these datasets may heighten already existing political tensions in countries where certain languages are prioritized over the others. There may be increased harms, such as increased surveillance, because now states and companies can better understand how people are talking in their languages. There is also a question of regulatory safeguards: We’re in contexts where some countries have not refined their data protection or their access to information laws. So people are having this data collected, but they are not sure what it's going to be used for.

Today's Big Tech-led AI lacks cultural nuance and social embeddedness for less-resourced languages.

The language datasets created by the people who are hoovering up all of this data always miss the cultural nuances. So you have hallucinations. You're being told, for example, "This name sounds African" and you're like, "What language is it?" You realize there's no language. It was just a bunch of syllables put together. This was an experience I recently had making use of a genAI platform. I laughed so hard because it sounded familiar in my language. And then when I asked where it's from, it was just like, "It sounded like it's African."

The communities that have been working on linguistic data for a long time—and are best positioned to govern it—now risk being excluded from global governance.

We've seen the success of small community entities actually digitizing their own languages, because maybe governments wouldn't have invested in them. Those communities now have oversight in the decision-making process and governance. Yet they may not have access to certain rooms because they may not be able to get a visa to come in time. They may not be able to afford the flight. Who actually is being included? Who are the people who've always been doing the work, and who is now taking up center stage?

REFRAME: The scale and scope of languages on the African continent requires community-led collaborative approaches that build on shared strengths and respect social norms.

The Masakhane Africa Languages Hub has been developing a community-led approach to creating linguistic datasets.

Masakhane African Languages Hub comes from the wider Masakhane community. We were doing things by the bootstraps because we were rejected in rooms where people said, "What do you mean African languages have some form of intelligence and they can be written about?" That was the call to action. The response was Global Majority people who are often excluded with the tenacity to say, "I want to hack the system and I want to see my language represented." The result is a community-led grassroots approach to creating these datasets. There is value in people voluntarily with consent participating in the design of these systems, because they can communicate with their devices in the language of choice.

Only a collaborative approach can address the scale and social embeddedness of language.

There are over two thousand languages on the African continent and they are evolving. This work cannot be done by one entity. There's a need to have multiple players in the room. We have to think about linguists, sociologists, the actual community of speakers, and entities interested in education.

We already have use cases in play, but there need to be multiple players who are contributing. We need government positions that allow for the creation of access to datasets. We need private-sector players that are also African contributing to ensure that the value chain is homegrown and is invested into creating the use cases for people, so that people are not just dataset collectors, but they're actually seeing tools and resources coming out.

A lot of developmental interventions are trying to address the gendered digital divide. Increasing access to information by having tools in languages and in voices that are not predominantly identified as male is important. But that creates a social issue where sometimes when people go into communities to collect this data, there may be patriarchal norms that prevent women from interacting with these platforms or in these spaces. There's a need to have everybody in the room to figure out how to get buy-in. It's not just the linguist that you have to have in the room. You have to go talk to the chief or the village headman who signs off on this exercise.

We have to recognize that there's a diversity in the way that people speak. University of Ghana partnered with Google to collect datasets for people with differences in speech, not standard speech, as a way to have robust technologies that can understand the difference in the way that people speak. You also have issues around accessibility, which requires working with people who already work on disability issues. It cannot just be the technologists who are creating datasets for standard speech. We must take into account building for difference more than building for what is considered "normal".

Accounting for social norms means asking difficult questions about the ownership of languages—and respecting communities' refusal.

I have seen examples of some people where they didn't want to have the language digitized. That brings out the nuance of intergenerational struggles and cultural issues that needs to be developed. We actually think about social norms. You may find that, for example, younger people are interested in collecting their languages because it's a way to connect to family. And you may have family members, elder members of the language community, actually saying "absolutely not" because they may be afraid that it's personal.

In those instances, if there's a community that says they don't want their data to be digitized, that has to be recorded and it has to be clear. But it creates a vacuum of governance. You'll have the push for the language to be digitized and then you have people saying they'd rather not: "It's also a means of protection of our community and of self."

I don't have the answer. This is something we're constantly in conversation about. We often ask ourselves who actually owns the language, who gives us permission. There's a need to have state-recognized language councils as part of the conversations. And we need to go into communities and figure out the accepted means of getting people's buy-in and doing the work, explaining why it's of value. If people are not interested, you have to acknowledge that it's not going to happen.

People have already been doing the work. Instead of doing everything from scratch, we must build on existing efforts and pool resources.

Intergovernmental organizations under the United Nations are trying to get governments to start committing to investing in and building up infrastructure to collect languages. Masakhane is an example of people who are already doing the work, so that there's none of this idea that "there's no one who's doing it, we don't have capacity." In conversations with government institutions who ask how to build up language data, we're able to show that there is an existing community from Masakhane who is already in your country doing this work.

Before Masakhane, there was the Lacuna Fund that worked to build datasets across the Global Majority. We have continued work with the Lacuna Fund to figure out where those datasets exist and how to make sure they're accessible to people. If a project sunsets, how do you ensure it continues to exist even after the funding period has ended?

We don't have to redesign everything. We need to figure out what resources are currently out there and then how do we share it with the wider community. How do we create collaborations, how do we spotlight each other, and how do we pool resources? In our thinking around institutional capacity building, our saying is, "We go far together."