LOST IN TRANSLATION

Frustrated that AI tools rarely understand their native languages, thousands of African volunteers are taking action *By* **Sandeep Ravindran**

Imagine joyfully announcing to your Facebook friends that your wife gave birth, and having Facebook automatically translate your words to "my prostitute gave birth." Shamsuddeen Hassan Muhammad, a computer science Ph.D. student at the University of Porto, says that's what happened to a friend when Facebook's English translation mangled the nativity news he shared in his native language, Hausa.

Such errors in artificial intelligence (AI) translation are common with African languages. AI may be increasingly ubiquitous, but if you're from the Global South, it probably doesn't speak your language.

That means Google Translate isn't much help, and speech recognition tools such as Siri or Alexa can't understand you. All of these services rely on a field of AI known as natural language processing (NLP), which allows AI to "understand" a language. The overwhelming majority of the world's 7000 or so languages lack data, tools, or techniques for NLP, making them "low-resourced," in contrast with a handful of "high-resourced" languages such as English, French, German, Spanish, and Chinese.

Hausa is the second most spoken African language, with an estimated 60 million to 80 million speakers, and it's just one of more than 2000 African languages that are mostly absent from AI research and products. The few products available don't work as well as those for English, notes Graham Neubig, an NLP researcher at Carnegie Mellon Univer-

ILLUSTRATION: A. MASTIN/SCIENCE

sity. “It’s not the people who speak the languages making the technology.” More often the technology simply doesn’t exist. “For example, now you cannot talk to Siri in Hausa, because there is no data set to train Siri,” Muhammad says.

He is trying to fill that gap with a project he co-founded called HausaNLP, one of several launched within the past few years to develop AI tools for African languages. Many projects have their roots in Masakhane, a pan-African volunteer effort led primarily by African researchers and coders determined to create translation products that would let ordinary Africans reap the benefits of the internet—and better cope with its pitfalls. Muhammad, for example, hopes to use these tools to help fight hate speech on social media and decolonize science by making research papers more accessible in African languages.

Similar projects have sprung up elsewhere across the Global South and among Indigenous communities in New Zealand and the Americas, aiming to use AI to preserve and revitalize languages discarded or disregarded because of colonialism. The work hasn’t yet produced the equivalent of a Siri or Google Translate, but these efforts are developing the data sets and software tools needed to build one. Jade Abbott, an NLP researcher and director at African startup Lelapa AI and co-founder of Masakhane, says the broader goal is to help more people in the Global South join the global economy. “The world of the internet is not a place for our languages yet, and it needs to be,” she says.

> “The world of the internet is not a place for our languages yet, and it needs to be.”
>
> **Jade Abbott**, Masakhane

BETTER AI for African languages could empower a huge number of people to access jobs and other opportunities that are now closed off to them, says Ignatius Ezeani, an NLP researcher at Lancaster University and a member of Masakhane, which has over 2000 volunteers from more than 30 countries. Ezeani says most Nigerians, including his parents, don’t speak English. As a result, they “struggle with their education, they struggle with the economy, with the agriculture, with the law, with health care, with disaster response,” Ezeani says.

Founded in 2019 by Abbott and her colleague Laura Martinus, Masakhane, which means “we build together” in isiZulu, hopes to help non-English speakers overcome such struggles. For example, African startups are using Masakhane’s data to build AI translation tools and chatbots to help people access financial services in their native languages. Such tools would also enable them to follow African news and government and legal communication—which in most countries currently exist primarily in English, French, or Arabic. “We need to try to use all these tools, if possible, to correct the errors of colonization and dehumanization of the last few centuries,” Ezeani says.

Abbott has a similar goal for science. Masakhane’s Decolonize Science project, which Muhammad is also involved in, aims to develop machine translations of African preprint research papers released on AfricArXiv. The preprints are often in English or European languages, but the project plans to translate them into six diverse African languages: isiZulu, Northern Sotho, Yoruba, Hausa, Luganda, and Amharic, together spoken by about 140 million people.

To create these tools, Masakhane can’t just copy what Google or Meta does—throwing massive amounts of data and computing power at the complex task of understanding a language. NLP works by breaking the task down into many smaller steps that machine learning algorithms can solve individually, by recognizing patterns in the text. One algorithm might split a paragraph of text into separate sentences. Another would then deconstruct each sentence into individual words. Additional models try to analyze each word separately to figure out whether it’s a noun, verb, or some other part of speech, and how different words in the sentence relate to each other.

Many current AI models learn to do all this by training on immense amounts of text data. “Google has basically scanned virtually every piece of human literature in the world, and so they have this huge data set,” says Michael Running Wolf, a software engineer and AI ethicist who founded Indigenous in AI. For high-resource languages such as English, those data can come in large part from web crawler programs that vacuum up all the text on the internet. African languages, however, are virtually absent from the internet. “It’s not a purely technical problem, it’s a societal problem,” Abbott says. Under colonialism, Africans were heavily discouraged from using their native languages, particularly in writing. “People were taught to feel ashamed for their own language,” she says. That leaves little written text for AI translation models to train on, let alone annotated speech for speech-to-text or voice recognition.

The data scarcity isn’t necessarily an insurmountable problem, Abbott says. Masakhane lacks Silicon Valley’s vast computational resources and cutting-edge software tools. It has to make do with older models that run on simpler hardware. And, she points out, “If you don’t have as much data, there’s no point having that big a model anyway, it’s not going to give you any advantage.”

Masakhane and a project for the Māori language called Papa Reo have found that a bit of data can go a long way. Papa Reo, for example, created an AI model using 300 hours of audio it collected by holding a competition to encourage people across New Zealand to record themselves speaking specific phrases in the Māori language, te reo Māori.

Masakhane has also developed techniques to create more data-efficient language models. In a recent paper, David Adelani, a Masakhane member and NLP researcher at University College London, and colleagues showed that instead of the 100,000 or 1 million sentences typically used to train NLP systems for high resource languages, existing models trained on large data sets to work with multiple languages could be fine-tuned to work with just 2000 sentences. The examples were drawn from high-quality translations of African news in 16 languages, including eight the model had never been exposed to before. That’s a hopeful sign that existing models can be adapted for low-resource languages, Adelani says.

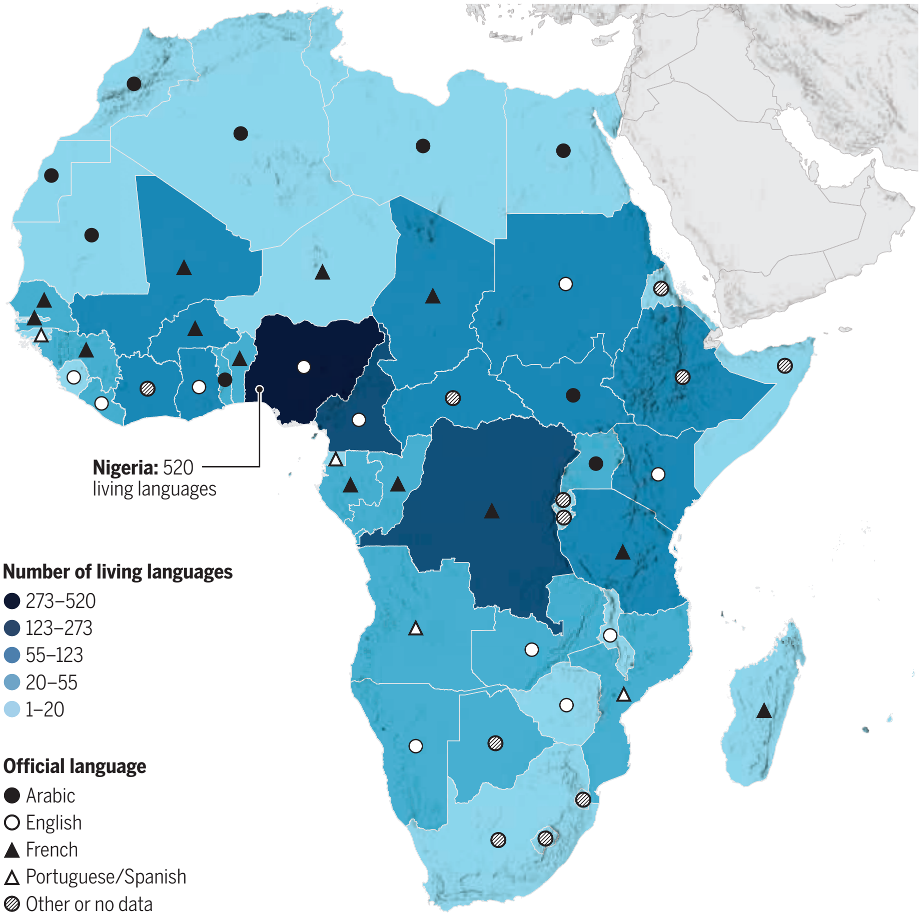
But even if these approaches require fewer data, Masakhane still has to collect those data from scratch. They’ve developed a participatory research process to create data sets based on community input, whether from news articles or from discussions with volunteers. “Most people are volunteering and doing that for the love of their language, for the survival of their language,” says Adelani, who has contributed to participatory research projects by Masakhane.

For example, when working with Khoekhoegowab, a very low resource language from Namibia, the consortium held an 8-day workshop with native speakers. The researchers started discussions with community participants using seed words in relevant topics, and jointly workshopped them into a list of sentences that community members deemed to be a natural use of those words. These efforts gathered smaller quantities of high-quality data directly relevant to training translation models, in contrast to the immense quantities of low-quality data big tech companies harvest from the internet.

“It really kind of changes the entire nature of the way we see data,” Abbott says. “Instead of a thing that’s scraped and extracted, it’s this beautiful area of creation,” she says.

A challenge for machine translation

As of 2022, there were more than 2000 living languages in Africa. With 520 languages, Nigeria accounted for about one-fourth of the total. For most African languages, data and software tools for natural language processing are scarce.



GRAPHIC: D. AN-PHAM/*SCIENCE*

THE MASAKHANE RESEARCHERS are also creating databases that address more specific needs in NLP. There's probably nothing more personal than ensuring that an AI understands your name and where you live. And yet when it comes to African languages, most AI translation tools struggle with named entity recognition (NER), the process of identifying proper names—such as a person, location, or organization.

To rectify this, Adelani, Abbott, and their colleagues helped create MasakhaNER, the first large-scale African language data set for NER. They annotated thousands of sentences from local news articles in 20 languages, flagging proper names by hand, to create a data set to train AI models to detect and categorize named entities in those languages. "NER is actually far bigger than just being an NLP task, it is about technology understanding and acknowledging you as a person," Abbott says.

She and her colleagues are also building language data sets for accurate sentiment analysis, which allows AI to understand the emotions of a particular text. When Muhammad started his Ph.D. in 2018, NLP researchers relied on translations of English sentiment analysis data. "This data set does not represent people in our community in Nigeria, the culture, the values, the knowledge," Muhammad says. He adds that translation can alter sentiment, such as when "my wife gave birth" becomes "my prostitute gave birth."

Through HausaNLP, Muhammad created an African language data set for sentiment analysis for the four most widely spoken Nigerian languages—Hausa, Igbo, Nigerian-Pidgin, and Yoruba. Volunteers helped him manually annotate about 30,000 tweets in each language with their corresponding sentiment, creating a training data set for AI models to detect sentiments in these languages. Lately, he has been focusing on one particular sentiment—hate—and trying to use sentiment analysis to automatically detect African language hate speech on social media.

Twitter, for example, has the ability to automatically block offensive tweets in English, but it has no such function for African languages. "There isn't even a data set to train a model to be able to understand whether this is hate or not hate," Muhammad says. Hate speech in African languages has to be manually swatted down. That is nowhere as effective as automated blocking, and aggrieved users often have to actively retweet a hateful tweet in order for it to get flagged and taken down by Twitter. Muhammad hopes his data set can help make online spaces safer for African language speakers.

BUILDING AFRICAN language data sets has been essential, but it's only one step toward making NLP work for African languages. Masakhane has also had to develop bespoke NLP tools.

Back in 2019, most NLP tools were built for English and a few other languages that are structured very differently from African languages. For example, tools that "tokenize" or separate English sentences into individual words don't work well for many African languages. That's particularly true for African languages such as isiZulu that are agglutinative—their words are made by combining shorter words in a way that's hard for English-trained AI to parse. Many African languages also include diacritics—marks such as a dot or an accent that guide pronunciation—making it harder to adapt English-trained AI to understand them. Some European languages do share these features, but so far only preliminary efforts have been made to adapt, say, German-trained AI to African languages.

Masakhane has made steady progress in developing tools to understand African languages, and in learning how best to apply models trained on other languages. For example, Adelani found that models trained on English work poorly on Yoruba compared with models transferred from certain other African languages. "The better the language similarity, the better is the transfer for any task," he says.

Adelani hopes to identify a small number of African languages that could be used to train versatile AI models that can work with many additional languages. "If you're able to identify these good donor languages that are easy to transfer to others, then basically even though we have 2000 languages, we might be able to do great things with maybe 20 languages," he says.

AI RESEARCHERS in the Global South and among Indigenous communities are wary of big tech companies harvesting their data to train proprietary AI models. "Data is the new oil," Running Wolf says. "And so there's sort of this very colonial perspective of, this is a land grab," he says.

As a result, many Indigenous communities are crafting protective licensing rules for the data they collect and the AI tools they develop. The New Zealand–based Papa Reo project uses a data license stipulating that any projects that use Māori data must respect Māori values and pass on any benefits to them. Similarly, the CARE Principles for Indigenous Data Governance developed by the Global Indigenous Data Alliance are aimed at ensuring that Indigenous communities worldwide maintain sovereignty over their data and ensure that they are used according to their principles and for their benefit.

But Masakhane, like some other projects, has so far kept its data sets and models open source. Some project leaders say it hasn't been an easy decision and remains a topic of discussion. Given the long history of exploitation of Indigenous and Global South communities and the continuing power imbalances between North and South, the potential misuse of data is a real concern. But for now, Masakhane has decided that the benefits of data sharing—such as making it easier for big tech companies to work on their native languages—outweigh the risks.

Several African startups—among them GhanaNLP, Lesan AI, and one founded by Abbott called Lelapa AI—have begun to develop consumer tools from Masakhane's data, such as apps and websites for text translation and speech recognition and transcription. "Ultimately what I'd love to see is ownership, in that these tools are owned by the communities that speak the languages rather than by the West," Abbott says. She envisions AI tools for native languages as a way to keep these languages alive. "Often you find people from the African continent who maybe left to go study abroad who now can't even speak to their mom because they don't speak the same language," she says.

Thanks to Masakhane's open-source policy, its data are also spurring efforts by Google and Meta to tailor their tools for African languages. "Data is a bottleneck in a lot of these NLP projects," Neubig says. On their own, big tech companies have little incentive to work on low-resource languages, but providing the data, as Masakhane and other projects have done, can act as a catalyst, he says. Abbott agrees. "Google Translate has managed to get some [African languages] up to reasonable performance with a lot of effort and push from people who are actually part of Masakhane," she says. A Google spokesperson acknowledged that limited data has slowed the company's efforts to develop tools for African languages, but said: "As our systems evolve and more data becomes available, we will continue to improve access and support for these languages in the future."

Rather than focusing on individual languages, Google and Meta have built large multilingual models for many hundreds of languages. For example, Google's model released in May 2022 supports more than 1000 languages and helped add 24 underresourced languages—including 10 African languages—to Google Translate. Meta's machine translation model released in July 2022 supports 200 languages, including more than 50 different African languages. And Meta's more recent models from May can recognize and produce speech for more than 1000 languages.

But Abbott says some of the big tech efforts have fallen short. Google's own researchers reported that some of their 1000-language model's translations of low-resource languages were rated very poorly by native speakers. And Meta's model for 200 languages performed poorly on some of the African languages it claimed to translate, Abbott says. She worries the publicity around Meta's model could hurt funding for the handful of small African AI startups that might do better. "To the rest of the world, it sounds like this is a solved problem because Facebook's gone and created this big model," Abbott says. A Meta spokesperson declined to comment.

Some Masakhane researchers give big tech companies credit for helping the homegrown efforts by sharing their models and data sets and by funding some community-led NLP projects. For instance, multiple Masakhane projects and similar projects for underserved languages have received funding from the Lacuna Fund, which began as a collaboration between the Rockefeller Foundation, Google, and Canada's International Development Research Centre and has since expanded. And some Google researchers have been part of the Masakhane community on Slack and helped mentor volunteers, including Muhammad.

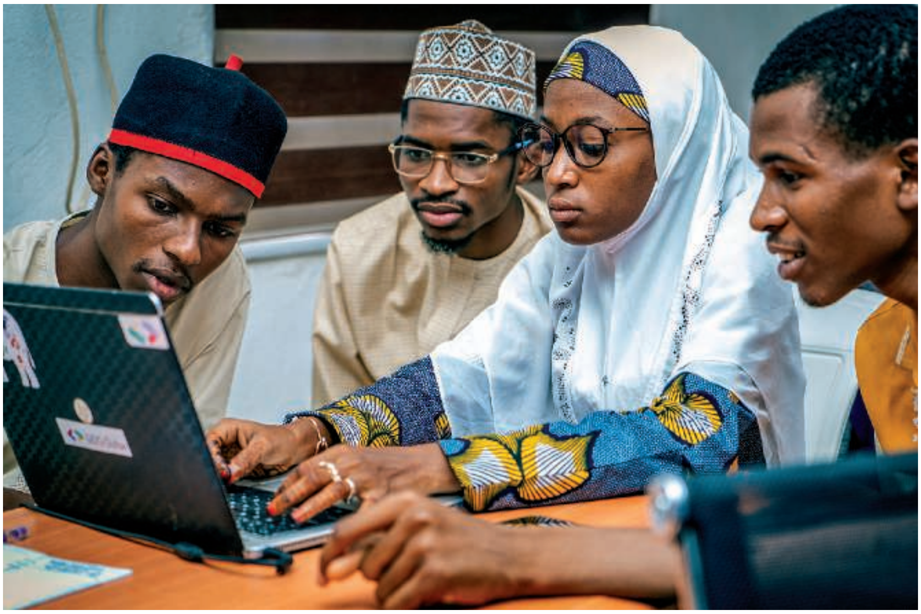


PHOTO: AREWA DATA SCIENCE

In May, fellows of the Arewa Data Science Academy, a free training program for Nigerian youth who want to learn data science and machine learning, participated in an artificial intelligence hackathon in Nigeria.

MASAKHANE'S MOST lasting legacy may be its people. For many volunteers, Masakhane has been a stepping stone toward pursuing academic research on AI, and the project has created a pipeline of AI researchers who are native speakers of African languages. "If the people who train the model understand the language, they can pick up issues in data sets," Abbott says.

Volunteers have collectively published hundreds of scientific papers, including translation models for at least 38 African languages, and presented several workshops at major AI conferences. That's a change from 2019, when Abbott says she was one of just three researchers from the entire continent of Africa among the thousands of attendees at a major computational linguistics conference. "The fact that it got so many people started and built this community of thousands of people is the real success story," Neubig says.

No one thinks AI will suddenly undo the ravages of hundreds of years of colonialism. "The damage has been done, and it took a long time to do the damage," Ezeani says. But Masakhane and similar projects are a positive step toward reducing the dominance of a handful of mostly European languages in AI research. "Many machine translation models that we developed were the first of their kind," Abbott says. "They exist because a person cared about that language." ■

Sandeep Ravindran is a science journalist near Washington, D.C.

Lost in translation

Sandeep Ravindran

Science, **381** (6655), .
DOI: 10.1126/science.adj8519

Science (ISSN 1095-9203) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.