TECHNOLOGY

The AI Revolution Is Crushing Thousands of Languages

English is the internet's primary tongue—a fact that may have unexpected consequences as generative AI becomes central to daily life.

By Matteo Wong

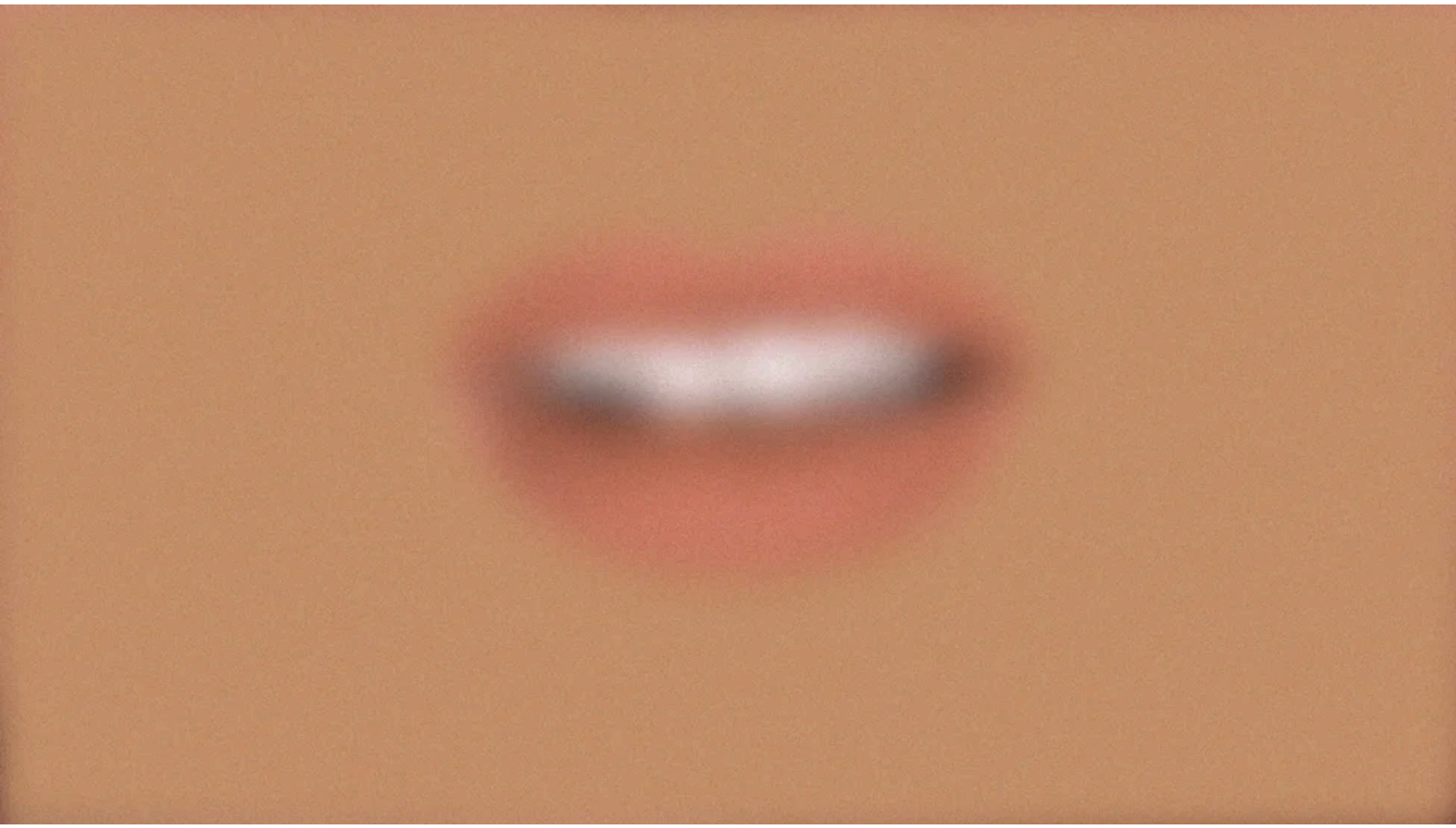

Illustration by Matteo Giuseppe Pani. Source: Getty.

APRIL 12, 2024

Recently, Bonaventure Dossou learned of an alarming tendency in a popular AI model. The program described Fon—a language spoken by Dossou's mother and millions of others in Benin and neighboring countries—as "a fictional language."

This result, which I replicated, is not unusual. Dossou is accustomed to the feeling that his culture is unseen by technology that so easily serves other people. He grew up with no Wikipedia pages in Fon, and no translation programs to help him communicate with his mother in French, in which he is more fluent. "When we have a technology that treats something as simple and fundamental as our name as an error, it robs us of our personhood," Dossou told me.

The rise of the internet, alongside decades of American hegemony, made English into a common tongue for business, politics, science, and entertainment. More than half of all websites are in English, yet more than 80 percent of people in the world don't speak the language. Even basic aspects of digital life—searching with Google, talking to Siri, relying on autocorrect, simply typing on a smartphone—have long been closed off to much of the world. And now the generative-AI boom, despite promises to bridge languages and cultures, may only further entrench the dominance of English in life on and off the web.

Scale is central to this technology. Compared with previous generations, today's AI requires orders of magnitude more computing power and training data, all to create the humanlike language that has bedazzled so many users of ChatGPT and other programs. Much of the information that generative AI "learns" from is simply scraped from the open web. For that reason, the preponderance of English-language text online could mean that generative AI works best in English, cementing a cultural bias in a technology that has been marketed for its potential to "benefit humanity as a whole." Some other languages are also well positioned for the generative-AI age, but only a handful: Nearly 90 percent of websites are written in just 10 languages

(English, Russian, Spanish, German, French, Japanese, Turkish, Portuguese, Italian, and Persian).

Some 7,000 languages are spoken in the world. Google Translate supports 133 of them. Chatbots from OpenAI, Google, and Anthropic are still more constrained. "There's a sharp cliff in performance," Sara Hooker, a computer scientist and the head of Cohere for AI, a nonprofit research arm of the tech company Cohere, told me. "Most of the highest-performance [language] models serve eight to 10 languages. After that, there's almost a vacuum." As chatbots, translation devices, and voice assistants become a crucial way to navigate the web, that rising tide of generative AI could wash out thousands of Indigenous and low-resource languages such as Fon—languages that lack sufficient text with which to train AI models.

"Many people ignore those languages, both from a linguistic standpoint and from a computational standpoint," Ife Adebara, an AI researcher and a computational linguist at the University of British Columbia, told me. Younger generations will have less and less incentive to learn their forebears' tongues. And this is not just a matter of replicating existing issues with the web: If generative AI indeed becomes the portal through which the internet is accessed, then billions of people may in fact be worse off than they are today.

Adebara and Dossou, who is now a computer scientist at Canada's McGill University, work with Masakhane, a collective of researchers building AI tools for African languages. Masakhane, in turn, is part of a growing, global effort racing against the clock to create software for, and hopefully save, languages that are poorly represented on the web. In recent decades, "there has been enormous progress in modeling low-resource languages," Alexandra Birch, a machine-translation researcher at the University of Edinburgh, told me.

In a promising development that speaks to generative AI's capacity to surprise, computer scientists have discovered that some AI programs can pinpoint

aspects of communication that transcend a specific language. Perhaps the technology could be used to make the web *more* aware of less common tongues. A program trained on languages for which a decent amount of data are available—English, French, or Russian, say—will then perform better in a lower-resourced language, such as Fon or Punjabi. "Every language is going to have something like a subject or a verb," Antonios Anastasopoulos, a computer scientist at George Mason University, told me. "So even if these manifest themselves in very different ways, you can learn something from all of the other languages." Birch likened this to how a child who grows up speaking English and German can move seamlessly between the two, even if they haven't studied direct translations between the languages—not moving from word to word, but grasping something more fundamental about communication.

Read: The end of foreign-language education

But this discovery alone may not be enough to turn the tide. Building AI models for low-resource languages is painstaking and time-intensive. Cohere recently released a large language model that has state-of-the-art performance for 101 languages, of which more than half are low-resource. That leaves about 6,900 languages to go, and this effort alone required 3,000 people working across 119 countries. To create training data, researchers frequently work with native speakers who answer questions, transcribe recordings, or annotate existing text, which can be slow and expensive. Adebara spent years curating a 42-gigabyte training data set for 517 African languages, the largest and most comprehensive to date. Her data set is *0.4 percent* of the size of the largest publicly available English training data set. OpenAI's proprietary databases—the ones used to train products such as ChatGPT—are likely far larger.

Much of the limited text readily available in low-resource languages is of poor quality—itself badly translated—or limited use. For years, the main sources of text for many such low-resource languages in Africa were translations of the Bible or missionary websites, such as those from Jehovah's Witnesses. And

crucial examples for fine-tuning AI, which has to be intentionally created and curated—data used to make a chatbot helpful, human-sounding, not racist, and so on—are even rarer. Funding, computing resources, and language-specific expertise are frequently just as hard to come by. Language models can struggle to comprehend non-Latin scripts or, because of limited training examples, to properly separate words in low-resource-language sentences—not to mention those without a writing system.

The trouble is that, while developing tools for these languages is slow going, generative AI is rapidly overtaking the web. Synthetic content is flooding search engines and social media like a kind of gray goo, all in hopes of making a quick buck.

Most websites make money through advertisements and subscriptions, which rely on attracting clicks and attention. Already, an enormous portion of the web consists of content with limited literary or informational merit—an endless ocean of junk that exists only because it might be clicked on. What better way to expand one's audience than to translate content into another language with whatever AI program comes up on a Google search?

Read: Prepare for the textpocalypse

Those translation programs, already of sometimes questionable accuracy, are especially bad with low-resourced languages. Sure enough, researchers published preliminary findings earlier this year that online content in such languages was more likely to have been (poorly) translated from another source, and that the original material was itself more likely to be geared toward maximizing clicks, compared with websites in English or other higher-resource languages. Training on large amounts of this flawed material will make products such as ChatGPT, Gemini, and Claude even worse for low-resource languages, akin to asking someone to prepare a fresh salad with nothing more than a pound of ground beef. "You are already training the

model on incorrect data, and the model itself tends to produce even more incorrect data," Mehak Dhaliwal, a computer scientist at UC Santa Barbara and one of the study's authors, told me—potentially exposing speakers of low-resource languages to misinformation. And those outputs, spewed across the web and likely used to train future language models, could create a feedback loop of degrading performance for thousands of languages.

Imagine "you want to do a task, and you want a machine to do it for you," David Adelani, a DeepMind research fellow at University College London, told me. "If you express this in your own language and the technology doesn't understand, you will not be able to do this. A lot of things that simplify lives for people in economically rich countries, you will not be able to do." All of the web's existing linguistic barriers will rise: You won't be able to use AI to tutor your child, draft work memos, summarize books, conduct research, manage a calendar, book a vacation, fill out tax forms, surf the web, and so on. Even when AI models are able to process low-resource languages, the programs require more memory and computational power to do so, and thus become significantly more expensive to run—meaning worse results at higher costs.

AI models might also be void of cultural nuance and context, no matter how grammatically adept they become. Such programs long translated "good morning" to a variation of "someone has died" in Yoruba, Adelani said, because the same Yoruba phrase can convey either meaning. Text translated from English has been used to generate training data for Indonesian, Vietnamese, and other languages spoken by hundreds of millions of people in Southeast Asia. As Holy Lovenia, a researcher at AI Singapore, the country's program for AI research, told me, the resulting models know much more about hamburgers and Big Ben than local cuisines and landmarks.

It may already be too late to save some languages. As AI and the internet make English and other higher-resource languages more and more convenient for

young people, Indigenous and less widely spoken tongues could vanish. If you are reading this, there is a good chance that much of your life is already lived online; that will become true for more people around the world as time goes on and technology spreads. For the machine to function, the user must speak its language.

By default, less common languages may simply seem irrelevant to AI, the web, and, in turn, everyday people—eventually leading to abandonment. "If nothing is done about this, it could take a couple of years before many languages go into extinction," Adebara said. She is already witnessing languages she studied as an undergraduate dwindle in their usage. "When people see that their languages have no orthography, no books, no technology, it gives them the impression that their languages are not valuable."

Read: AI is exposing who really has power in Silicon Valley

Her own work, including a language model that can read and write in hundreds of African languages, aims to change that. When she shows speakers of African languages her software, they tell her, "'I saw my language in the technology you built; I wasn't expecting to see it there,'" Adebara said. "'I didn't know that some technology would be able to understand some part of my language,' and they feel really excited. That makes me also feel excited."

Several experts told me that the path forward for AI and low-resource languages lies not only in technical innovation, but in just these sorts of conversations: not indiscriminately telling the world it needs ChatGPT, but asking native speakers what the technology can do for them. They might benefit from better voice recognition in a local dialect, or a program that can read and digitize non-Roman script, rather than the all-powerful chatbots being sold by tech titans. Rather than relying on Meta or OpenAI, Dossou told me, he hopes to build "a platform that is appropriate and proper to African languages and Africans, not trying to generalize as Big Tech does." Such efforts could help give low-resource languages a presence on the internet

where there was almost none before, for future generations to use and learn from.

Today, there is a Fon Wikipedia, although its 1,300 or so articles are about two-thousandths of the total on its English counterpart. Dossou has worked on AI software that does recognize names in African languages. He translated hundreds of proverbs between French and Fon manually, then created a survey for people to tell him common Fon sentences and phrases. The resulting French-Fon translator he built has helped him better communicate with his mother—and his mother's feedback on those translations has helped improve the AI program. "I would have needed a machine-translation tool to be able to communicate with her," he said. Now he is beginning to understand her without machine assistance. A person and their community, rather than the internet or a piece of software, should decide their native language—and Dossou is realizing that his is Fon, rather than French.

ABOUT THE AUTHOR

Matteo Wong

Follow

Matteo Wong is a staff writer at *The Atlantic*.