In Africa, rescuing the languages that Western tech ignores

December 23 2021, by Matt O'brien and Chinedu Asadu

In this photo taken Wednesday Nov 24, 2021, Kola Tubosun, is photograph in his house in Lagos, Nigeria. Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages. But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech

economy. Credit: AP Photo/Sunday Alamba

Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages.

But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech economy.

"We are getting to the point where if a machine doesn't understand your language it will be like it never existed," said Vukosi Marivate, chief of data science at the University of Pretoria in South Africa, in a call to action before a December virtual gathering of the world's artificial intelligence researchers.

American tech giants don't have a great track record of making their language technology work well outside the wealthiest markets, a problem that's also made it harder for them to detect dangerous misinformation on their platforms.

Marivate is part of a coalition of African researchers who have been trying to change that. Among their projects is one that found machine translation tools failed to properly translate online COVID-19 surveys from English into several African languages.

In this photo taken Wednesday Nov 24, 2021, Kola Tubosun, is photograph in his house in Lagos, Nigeria. Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages. But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech economy. Credit: AP Photo/Sunday Alamba

"Most people want to be able to interact with the rest of the information highway in their local language," Marivate said in an interview. He's a founding member of Masakhane, a pan-African research project to improve how dozens of languages are represented in the branch of AI

known as natural language processing. It's the biggest of a number of grassroots language technology projects that have popped up from the Andes to Sri Lanka.

Tech giants offer their products in numerous languages, but they don't always pay attention to the nuances necessary for those apps work in the real world. Part of the problem is that there's just not enough online data in those languages—including scientific and medical terms—for the AI systems to effectively learn how to get better at understanding them.

Google, for instance, offended members of the Yoruba community several years ago when its language app mistranslated Esu, a benevolent trickster god, as the devil. Facebook's language misunderstandings have been tied to political strife around the world and its inability to tamp down harmful misinformation about COVID-19 vaccines. More mundane translation glitches have been turned into joking online memes.

Omolewa Adedipe has grown frustrated trying to share her thoughts on Twitter in the Yoruba language because her automatically translated tweets usually end up with different meanings.

In this photo taken Wednesday Nov 24, 2021, Kola Tubosun, is photograph in his house in Lagos, Nigeria. Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages. But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech economy. Credit: AP Photo/Sunday Alamba

One time, the 25-year-old content designer tweeted, "T'Ílù ò bà dùn, T'Ílù ò bà t'òrò. Èyin l'ęmò bí ę şe şé," which means, "If the land (or country, in this context) is not peaceful, or merry, you're responsible for it." Twitter, however, managed to end up with the translation: "If you are

not happy, if you are not happy."

For complex Nigerian languages like Yoruba, those accent marks—often associated with tones—make all the difference in communication. 'Ogun', for instance, is a Yoruba word that means war, but it can also mean a state in Nigeria (Ògùn), god of iron (Ògún), stab (Ógún), twenty or property (Ogún).

"Some of the bias is deliberate given our history," said Marivate, who has devoted some of his AI research to the southern African languages of Xitsonga and Setswana spoken by his family members, as well as to the common conversational practice of "code-switching" between languages.

"The history of the African continent and in general in colonized countries, is that when language had to be translated, it was translated in a very narrow way," he said. "You were not allowed to write a general text in any language because the colonizing country might be worried that people communicate and write books about insurrections or revolutions. But they would allow religious texts."

In this photo taken Wednesday Nov 24, 2021, Kola Tubosun, is photograph in his house in Lagos, Nigeria. Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages. But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech economy. Credit: AP Photo/Sunday Alamba

Google and Microsoft are among the companies that say they are trying to improve technology for so-called "low-resource" languages that AI systems don't have enough data for. Computer scientists at Meta, the company formerly known as Facebook, announced in November a

breakthrough on the path to a "universal translator" that could translate multiple languages at once and work better with lower-resourced languages such as Icelandic or Hausa.

That's an important step, but at the moment, only large tech companies and big AI labs in developed countries can build these models, said David Ifeoluwa Adelani. He's a researcher at Saarland University in Germany and another member of Masakhane, which has a mission to strengthen and spur African-led research to address technology "that does not understand our names, our cultures, our places, our history."

Improving the systems requires not just more data but careful human review from native speakers who are underrepresented in the global tech workforce. It also requires a level of computing power that can be hard for independent researchers to access.

Writer and linguist Kola Tubosun created a multimedia dictionary for the Yoruba language and also created a text-to-speech machine for the language. He is now working on similar speech recognition technologies for Nigeria's two other major languages, Hausa and Igbo, to help people who want to write short sentences and passages.

In this photo taken Wednesday Nov 24, 2021, Kola Tubosun, is photograph in his house in Lagos, Nigeria. Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages. But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech economy. Credit: AP Photo/Sunday Alamba

"We are funding ourselves," he said. "The aim is to show these things can be profitable."

Tubosun led the team that created Google's "Nigerian English" voice and

accent used in tools like maps. But he said it remains difficult to raise the money needed to build technology that might allow a farmer to use a voice-based tool to follow market or weather trends.

In Rwanda, software engineer Remy Muhire is helping to build a new open-source speech dataset for the Kinyarwanda language that involves a lot of volunteers recording themselves reading Kinyarwanda newspaper articles and other texts.

"They are native speakers. They understand the language," said Muhire, a fellow at Mozilla, maker of the Firefox internet browser. Part of the project involves a collaboration with a government-supported smartphone app that answers questions about COVID-19. To improve the AI systems in various African languages, Masakhane researchers are also tapping into news sources across the continent, including Voice of America's Hausa service and the BBC broadcast in Igbo.

In this photo taken Wednesday Nov 24, 2021, Kola Tubosun, is photograph in his house in Lagos, Nigeria. Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages. But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech economy. Credit: AP Photo/Sunday Alamba

In this photo taken Wednesday Nov 24, 2021, Kola Tubosun, is photograph in his house in Lagos, Nigeria. Computers have become amazingly precise at translating spoken words to text messages and scouring huge troves of information for answers to complex questions. At least, that is, so long as you speak English or another of the world's dominant languages. But try talking to your phone in Yoruba, Igbo or any number of widely spoken African languages and you'll find glitches that can hinder access to information, trade, personal communications, customer service and other benefits of the global tech economy. Credit: AP Photo/Sunday Alamba

Increasingly, people are banding together to develop their own language approaches instead of waiting for elite institutions to solve problems, said Damián Blasi, who researches linguistic diversity at the Harvard Data Science Initiative.

Blasi co-authored a recent study that analyzed the uneven development of language technology across the world's more than 6,000 languages. For instance, it found that while Dutch and Swahili both have tens of millions of speakers, there are hundreds of scientific reports on natural language processing in the Western European language and only about 20 in the East African one.