The race to save indigenous languages using automatic speech recognition

October 11 2021, by Tanner Stening

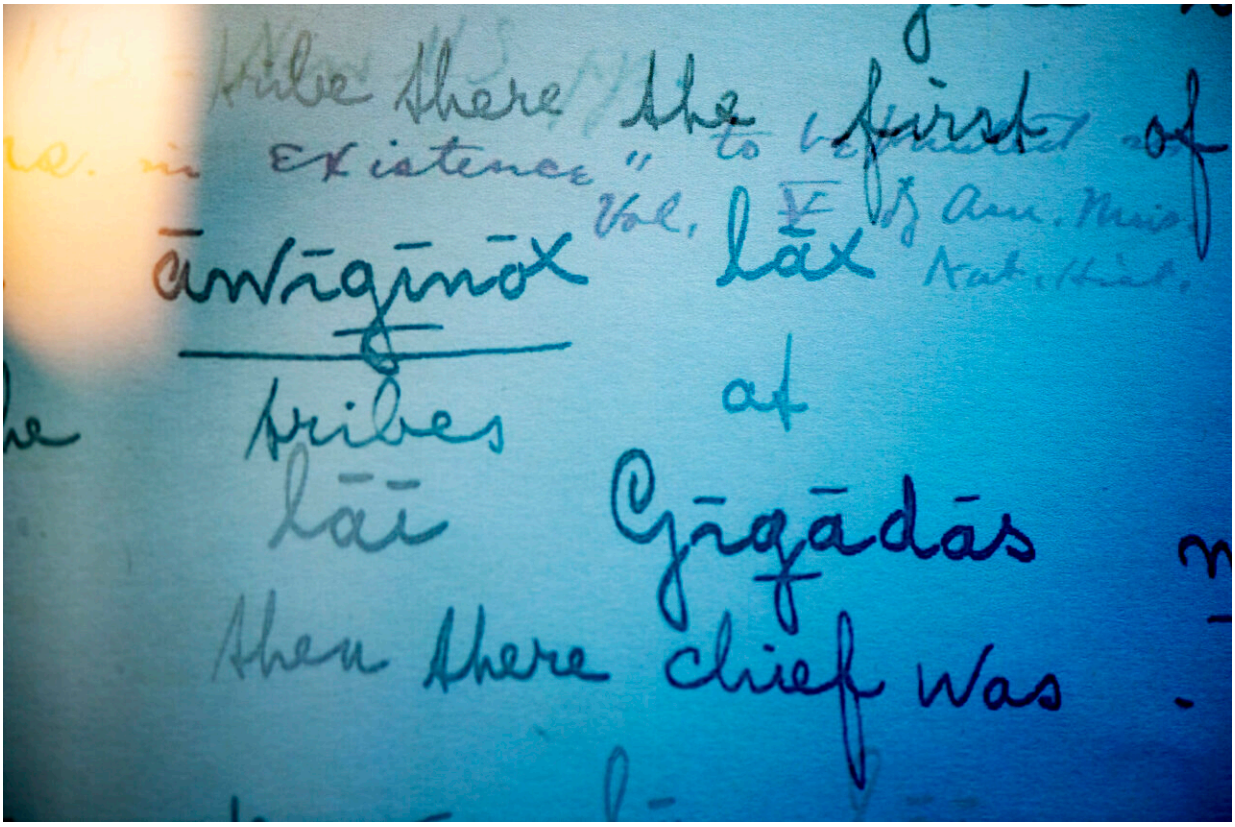

Photo illustration of Kwak'wala text written by Northeastern clinical instructor Michael Running Wolf. Credit: Alyssa Stone/Northeastern University

Michael Running Wolf still has that old TI-89 graphing calculator he used in high school that helped propel his interest in technology.

"Back then, my teachers saw I was really interested in it," says Running Wolf, clinical instructor of computer science at Northeastern University. "Actually a couple of them printed out hundreds of pages of instructions for me on how to code" the device so that it could play games.

What Running Wolf, who grew up in a remote Cheyenne village in Birney, Montana, didn't realize at the time, poring over the stack of printouts at home by the light of kerosene lamps, was that he was actually teaching himself basic programming.

"I thought I was just learning how to put computer games on my calculator," Running Wolf says with a laugh.

But it hadn't been his first encounter with technology. Growing up in the windy plains near the Northern Cheyenne Indian Reservation, Running Wolf says that although his family—which is part Cheyenne, part Lakota—didn't have daily access to running water or electricity, sometimes, when the winds died down, the power would flicker on, and he'd plug in his Atari console and play games with his sisters.

These early experiences would spur forward a lifelong interest in computers, artificial intelligence, and software engineering that Running Wolf is now harnessing to help reawaken endangered indigenous languages in North and South America, some of which are so critically at risk of extinction that their tallies of living native speakers have dwindled into the single digits.

Running Wolf's goal is to develop methods for documenting and maintaining these early languages through automatic speech recognition software, helping to keep them "alive" and well-documented. It would be a process, he says, that tribal and indigenous communities could use to supplement their own language reclamation efforts, which have intensified in recent years amid the threats facing languages.

"The grandiose plan, the far-off dream, is we can create technology to not only preserve, but reclaim languages," says Running Wolf, who teaches computer science at Northeastern's Vancouver campus. "Preservation isn't what we want. That's like taking something and embalming it and putting it in a museum. Languages are living things."

The better thing to say is that they've "gone to sleep," Running Wolf says.

And the threats to indigenous languages are real. Of the roughly 6,700 languages spoken in the world, about 40 percent are in danger of atrophying out of existence forever, according to UNESCO Atlas of Languages in Danger. The loss of these languages also represents the loss of whole systems of knowledge unique to a culture, and the ability to transmit that knowledge across generations.

While the situation appears dire—and is, in many cases—Running Wolf says nearly every Native American tribe is engaged in language reclamation efforts. In New England, one notable tribe doing so is the Mashpee Wampanoag Tribe, whose native tongue is now being taught in public schools on Cape Cod, Massachusetts.

But the problem, he says, is that in the ever-evolving field of computational linguistics, little research has been devoted to Native American languages. This is partially due to a lack of linguistic data, but it is also because many native languages are "polysynthetic," meaning they contain words that comprise many morphemes, which are the smallest units of meaning in language, Running Wolf says.

Polysynthetic languages often have very long words—words that can mean an entire sentence, or denote a sentence's worth of meaning.

Further complicating the effort is the fact that many Native American

languages don't have an orthography, or an alphabet, he says. In terms of what languages need to keep them afloat, Running Wolf maintains that orthographies are not vital. Many indigenous languages have survived through a strong oral tradition in lieu of a robust written one.

But for scholars looking to build databases and transcription methods, like Running Wolf, written texts are important to filling in the gaps. What's holding researchers back from building automatic speech recognition for indigenous languages is precisely that there is a lack of audio and textual data available to them.

Using hundreds of hours of audio from various tribes, Running Wolf has managed to produce some rudimentary results. So far, the automatic speech recognition software he and his team have developed can recognize single, simple words from some of the indigenous languages they have data for.

"Right now, we're building a corpus of audio and texts to start showing early results," Running Wolf says.

Importantly, he says, "I think we have an approach that's scientifically sound."

Eventually, Running Wolf says he hopes to create a way for tribes to provide their youth with tools to learn these ancient languages by way of technological immersion—through things like augmented or virtual reality, he says.

Some of these technologies are already under development by Running Wolf and his team, made up of a linguist, a data scientist, a machine learning engineer, and his wife, who used to be a program manager, among others. All of the ongoing research and development is being done in consultation with numerous tribal communities, Running Wolf

says.

"It's all coming from the people," he says. "They want to work with us, and we're doing the best to respect their knowledge systems."

Provided by Northeastern University

Citation: The race to save indigenous languages using automatic speech recognition (2021, October 11) retrieved 28 April 2024 from https://techxplore.com/news/2021-10-indigenous-languages-automatic-speech-recognition.html