



July 28, 2011

# A Chance to Record Some of the 112 Local Languages

By REYHAN HARMANCI

Standing on a BART platform or strolling down Market Street, it is easy to pick up snippets of languages other than English — rapid-fire Spanish, of course, or maybe animated Mandarin.

But that is only the beginning. Croatian, Navaho, Swahili, Czech, even Scottish Gaelic and more are spoken in the Bay Area.

With 112 languages in use locally, according to the census, the Bay Area is one of the country's most linguistically diverse metropolitan areas. (Los Angeles, the most diverse, has 137.)

On Saturday, residents will get a rare opportunity to hear — and speak in — as many tongues as possible, thanks to a collaboration between the [Internet Archive](#) and the [Long Now Foundation](#), nonprofits dedicated to archival work, and [Mighty Verse](#), a start-up that has created a database of phrase videos.

More than 100 educational and cultural organizations in the Bay Area have been asked to participate, which they can do remotely or in person at the Internet Archive's San Francisco office, where the event will be held.

The organizers are hoping to capture 50 different languages during this "record-a-thon" with webcams and cellphone cameras and in a studio. There will be a morning workshop, but the event is meant to be fairly free-form — a combination of "structured recordings," with texts provided, and opportunities to mingle and record casual conversations.

Like some other recent linguistic projects, namely recordings of the 2011 freshman class of the University of California, Berkeley, the record-a-thon uses digital tools to capture and analyze what used to be unmanageable amounts of data.

Laura Welcher, director of the Rosetta Project, the language-focused arm of the Long Now Foundation, said the Record-a-thon was in part inspired by StoryCorps. StoryCorps, an independent nonprofit oral history project whose media partner is National Public Radio, sets up traveling booths for people to record personal histories. “I thought, that’s cool, I wonder if we can use that for language documentation,” Dr. Welcher said.

This project reflects the larger mission of Rosetta, to document as much of the world’s linguistic offerings as possible before they disappear. According to Dr. Welcher, a language becomes extinct every two weeks, and expeditions into every remote area are financially impossible.

Currently, the Rosetta disk holds 14,000 images, viewable by microscope, of “parallel texts” — the same word or phrase in about a thousand different languages — in a form created to last thousands of years. In addition, the project has 100,000 pages of documents in more than 2,500 languages available to view or hear at the Internet Archive.

“We don’t have the hope of documenting all the languages before they disappear,” Dr. Welcher said. But “we have over 100 different languages in the Bay Area, and that’s a really good place to start.”

While Dr. Welcher said she would be thrilled if speakers of endangered languages showed up, her main goal, besides adding to the Rosetta library, is to lower the barrier to professional language documentation.

Paul Lundahl, co-founder of Mighty Verse, a company that is creating a database of videos of people speaking useful phrases in everything from Urdu to Japanese to Na’vi, the language from the film “Avatar,” agreed, saying he hoped the event would get people to “celebrate the idea of recording their language.”

Mr. Lundahl and his team will be on hand to capture people reciting the Universal Declaration of Human Rights, one of the parallel texts archived by Rosetta because it has been translated into more than 400 languages.

The recordings will be available on the Mighty Verse and Internet Archive Web sites.

Berkeley is also experimenting with language recordings, although the university is interested in a different kind of diversity. Keith Johnson, a linguistics professor, is spearheading an effort to collect and map the voices of all 5,800 incoming freshmen, the majority of whom come from California, for his long-term research on California dialects.

“Typically, the dialect map of English stops at the Mississippi River,” Dr. Johnson said, making it hard to determine how Bay Area speakers may differ from the rest of the state and country. But he did cite “hella” as a kind of shibboleth, revealing the speaker to be from the East Bay.

The digital era has brought setbacks as well as opportunities in the struggle to keep linguistic diversity intact. Dr. Welcher, for instance, cited research showing that the bulk of communication online is in only 10 languages.

But efforts like the record-a-thon and other large-scale digitalization projects, like Google Books and the Gutenberg Project, which aims to translate 10 million books in the public domain into 100 languages, could go far toward keeping rare languages alive.

For instance, Brewster Kahle, the Internet Archive’s founder, said his organization was helping scan Balinese literature. Since Balinese children are being taught Indonesian, the traditional Balinese language needs support, Mr. Kahle said. “Technology, in this case, can help.”

Besides, he finds the cultural differences fascinating. “Digitizing their literature is just so darn cool,” Mr. Kahle said. “They write text on palm leaves.”

*rharmanci@baycitizen.org*