Emerging Tech

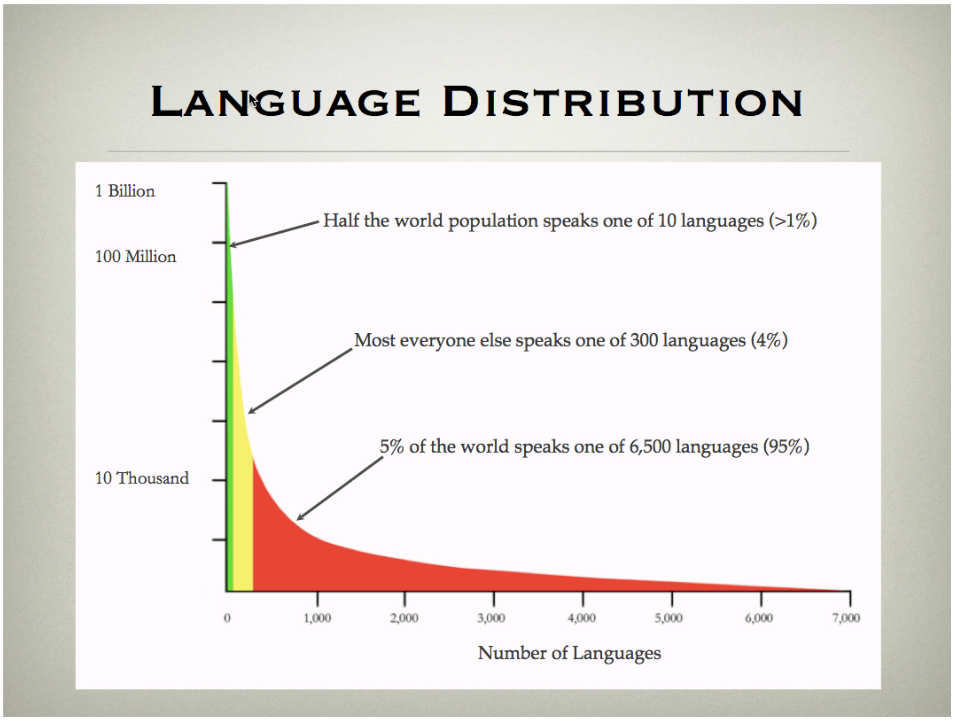
Apple Siri, Google Voice could help save the world's languages

By Chris Jablonski | November 14, 2011, 6:59pm PST

Summary: 80% of all web communication is in ten languages, yet 95% of humanity speaks roughly 300 languages. As digital services and devices move to voice control, the commercial opportunity could help close the digital linguistic divide, says the Long Now Foundation.

The majority of the world's languages have only a few thousand speakers each, therefore, provide no commercial incentives to preserve or to enable on the web.

If you look to the left of the long tail, however, said Dr. Laura Welcher, Director of Operations for the Rosetta Project at the Long Now Foundation, there are about **300 widely spoken languages that do provide motivation for providers of digital services and devices because this group accounts for 95% of all people on earth**. (See the yellow colored band in the image).



Credit: Dr. Laura Welcher, Long Now Foundation

In a recent talk given at UC Berkeley's Language Center, Welcher described her organization's goal of creating an open public digital collection of all human language as well as an analog backup– the Rosetta Disk– a solid nickel surface with 13,000 microetched pages of language documentation that can last for thousands of years.

Experts say that we lose a language every two weeks and up to 90% of roughly 7,000 languages will go extinct in 100 years. To counter the trend, the Long Now Foundation is leading a herculean effort to preserve thousands of endangered languages around the world.

In her talk, Welcher applauded Google's plan to sample 300 languages from around the world to help improve its Voice Search product, saying that ideally the data collected would find its way into the public domain such as Language Commons or Rosetta Language Base on Freebase (an open platform owned by Google).

Welcher said that the long tail of roughly 6,500 languages could benefit from development of the 300 (and vice versa) if we build better algorithms that can work with less data. Long tail languages can also be helped through philanthropic efforts.

"As companies make corpora, if it is open then linguists can access it and help build a platform to help endangered languages of the world," she asserted.

Welcher did not cover Apple's Siri voice controlled personal assistant technology. But it currently supports three languages (English, French, German) and in 2012 will include most of the top ten used languages on the web, namely Chinese, Japanese and Spanish. As Siri grows in both linguistic diversity and capability, any second-tier languages may take less resources to support, giving Apple the green light to contribute to open resources on human languages.

If there is anything that the Rosetta Project needs to fulfill its objective, it's help. The current collection contains 100,000 pages of scanned material documenting over 2,500 languages, as well as a growing library of crowd-sourced audio and video recordings. But that's just a scratch on the surface. There is substantial machine readable corpora for only about 20-30 of the world's languages. Welcher expects to add only 500 more into the digital domain over the next 10 years unless she can substantially scale the effort.

Programs like the 300 Languages Project and "Record-a-thon" are helping to close the gap, but it will take more to reach her goal of documenting at least 5,000 languages before they disappear. Welcher asked: "How do we get the isocode for all human languages and develop a universal corpus with reliable machine translation?"

Welcher ended her talk with a vision of a free and open encyclopedia of human languages that could model Wikipedia and the encyclopedia of life.

Further reading:

Internet Archive: The Rosetta Project
The DVD-Sized Rosetta Disk Will Preserve Human Language For Eternity
Found in Translation: The blog of the Berkeley Language Center

Related:

A ’stone-like’ optical disc that lasts for millennia
The Long Now Foundation’s 10,000 year clock

Web, Disk, Language, Chris Jablonski