There are 7,000 languages. Wait, no, 7,117. No, 7,868. What's going on?

Written by: Martin Benjamin
Published by: Kamusi Project International, http://kamu.si/7000-languages, 28 May 2020

When knowledgeable people discuss the number of languages that exist in the world, they necessarily land on an imperfect number.

The most common figure bandied about is 7,000. This is really just a convenient round number that is certainly an undercount.

The 23rd edition of Ethnologue, produced by SIL International, lists 7,117 "living" languages. This number is 6 higher than the year before – 12 languages were newly identified or changed from "extinct" or "unattested", while 6 languages were determined to have gone extinct or never have been unique languages at all. Further efforts to view a language profile on Ethnologue lead to a paywall that mentions 7,465 language profiles. The number of languages in Ethnologue will continue to shift around as the editors piece together new information.

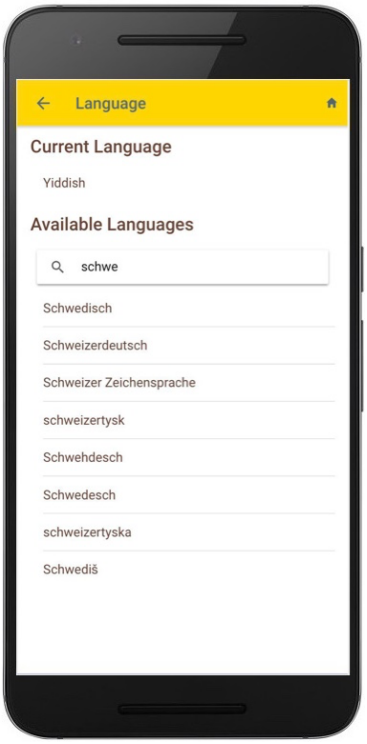
At the same time, SIL also lists 7,868 languages in its capacity as the agent for the International Standards Organization in determining ISO-639-3 codes. The big difference between ISO and Ethnologue is that ISO includes four other types of languages, classified as extinct, ancient, historic, and constructed. The difference between "extinct" and "ancient" is a bit fuzzy. "Historic" languages are versions with written records that have evolved into distinctly different contemporary languages, such as Old English. "Constructed" languages are supposed to be "complete" languages for human communication that have been passed to a second generation; Esperanto clearly falls into this category, but somehow Klingon also gets an ISO code, while other borderline constructed languages do not.

These numbers seem authoritative, but they are not the last word. The big problem is the distinction between a «language» and a "macrolanguage" on one side, and a "dialect" on the other. A lot of that comes down to politics. For example, Kirundi is the language of Burundi, and Kinyarwanda is the language of Rwanda, each with their own ISO code, their own army, and their own seat at the United Nations. However, not even a river marks much of the border between the two countries. The main differences between the two "languages" are certain decisions about spelling. On the other hand, several languages in Mali and Niger that are not mutually intelligible are lumped within the Songhay macrolanguage. And Chinese is a mess: a lot of aurally unintelligible languages use the same writing system, Taiwan and the mainland use different writing systems for essentially the same language, and calling something a separate language instead of a dialect can sometimes be interpreted as political separatism with dangerous consequences.

Are Bena and Hehe and Kinga and Sangu, spoken in southern Tanzania and owners of their own ISO codes, distinct languages, or dialects of something larger? A Bena lass could marry a Hehe lad, and their children would barely notice a difference between mom and dad. Things called "Fula" are spoken by some people in almost every country in western Africa, with no borders between "dialects", but no chance that a speaker in Senegal will understand someone in the Central African Republic, 4000 km away.

The Kamusi Project tries to keep outside of this debate. We must inevitably confront it, though, as we roll out games that could in principle be played to elicit data in any language we list. Our policy is to list every language with an ISO code, and some variants without. For example, Portuguese comes in a few flavors, including Brazilian, Mozambican, and Original, all served with the ISO code "por". We have two main strategies for dealing with these variants. First, we sometimes hardwire distinctions into the language code, such as "por-br" and "por-pt" for the two major branches of Portuguese. Second, yet to be coded, will be the geotagging of lexical items, so that you can see the words people use in a locality without having to declare artificial dialect or language boundaries within a linguistic continuum. When we have comparative data, researchers will be able to compare how languages like Bena and Hehe and Kinga and Sangu speak of thousands of similar concepts, and make declarations about language boundaries based on hard data.

Historical languages are important to us as well, for two reasons. One, we are coding a way to show the relation of words to their ancestors (their etymologies) such that users can drill forward and backward through time, and laterally across related languages. Two, we aspire to create dictionaries of past languages in their own right, when sufficient data exists. Therefore, we include these languages in our list, even though we do not yet have data for them, and will never have native speakers to contribute terms or take advantage of the resource.



Our language list has all of the names available to us in English, with whatever labels came from the original source. In addition, we have merged in all of the names that languages have been given in other languages (for example, the Chinese word for "English"), to the extent that the info is available via the CLDR. We have certainly missed some names that should be in our identification system; Ethnologue lists many alternate names for included languages, but not as open data. We periodically look for new sources to augment what we've got. That's behind the scenes. Up front, we're not going to try to count. For convenience sake, we'll just say "over 7,000", and get busy collecting data from as many languages and variants as we can, and putting that data to work.