How we're making Common Voice even more linguistically inclusive

foundation.mozilla.org/en/blog/how-we-are-making-common-voice-even-more-linguistically-inclusive

December 2, 2021

Languages and Common Voice

People want to be heard.

To be ignored, misunderstood and misinterpreted can strike at our core sense of self. For some of us, this only happens infrequently, and not in profound ways. For others, an entire life can be spent fighting to be understood on one's own terms.

A huge part of this is language. Language is more than sounds - it can be a sense of home, of belonging, the way we express our emotions and move others. In today's world, languages are not all treated the same. Many - even most - are ignored, threatened, exploited or degraded.

Common Voice wants to change all that, by unlocking the power of communities everywhere to share their voices, so we can all be part of building something new. Today, almost half a million people have shared sentences, read them aloud, and validated other people's recordings to create the largest, most diverse open source voice dataset in the world. Common Voice is structured around language - it's made up of 80+ communities, and it's growing all the time.

How Common Voice works now

There are two categories for speech on Common Voice; language, or accent. Communities can localise the site and begin collecting voice data for a whole new language dataset, or they can suggest accents to an existing language via github.

We spoke to language community members, contributors, dataset consumers and linguists and identified several issues in how this is working right now;

1. New communities who don't have a lot of context weren't always sure how to categorise themselves, and once they'd made their decision it is hard to change their minds.
2. Diverse communities - for example those with large diaspora populations - may feel they need to split up entirely and set up a whole new language, fragmenting the dataset and confusing contributors.
3. Some language communities and contributors make use of accent tags, but can feel marginalised and undermined by this. Talking about language is talking about power, and some people want to have the ability to identify their speech beyond 'accent', in ways which respect and represent them.
4. Some communities felt there was no suitable arrangement for them, as their spoken language had multiple writing systems, and currently MCV assumes a 1:1 relationship between spoken word and written word.

Where we propose to go next

1. **We want to support communities to better understand their options from the start**

In consultation with the community, linguistic advisors and machine learning practitioners, we are releasing a set of high level definitions and guidelines that communities can use to understand their choices. We intend that these are as broad and flexible as possible, giving communities scaffolding, rather than constraints.

Ultimately, we feel that communities are the best judge of how they'd like to structure themselves. Our hope is just that they have all the information they need to make a choice that works for them long term, and also takes account of the trade offs for the wider community of people who contribute to, and use, the dataset.

1. **We would like to enable a new category on the platform - Variant**. This is to help communities systematically differentiate within languages, and especially to support large languages with a diverse range of speakers.

Where possible, we will use BCP-47 codes to tag these. BCP 47 is a flexible system that allows communities to pull out key information such as region, dialect and orthography.

For example, the Kiswahili community might like to differentiate between Congolese Swahili and Chimwiini. At present on the platform, this would be framed as an 'accent' difference - despite the fact that the variants have different vocabulary, grammar and

would not be easily mutually intelligible. Communities will be free to choose whether and how they make use of the variant tag.

We plan to roll this out in the platform experience in phases, starting by allowing people to set their preferences during profile creation, and thus tag their voice clips with the information. Next year, we will evaluate other locations for its inclusion, for example sentence collection, and the Speak interface.

1. **We will also make it easier, and more flexible, to identify your accent**

As of December 2021, people will also be able to easily add multiple accents in their profile page.

They won't be limited by what else people have chosen, or need to submit changes via Github. This is designed to make it easier for contributors to quickly identify their speech in a way that feels natural to them.

1. **We plan to build out support for multiple orthographies for a single language**

Following some exciting hiring in 2021 to build out our technical team, in 2022 we will break apart the tight, inflexible relationship between browser locale, written word and spoken word.

This means that for languages with multiple writing systems - such as Serbian or Konkani - the platform will give the option to contribute to a shared dataset, but with multiple orthography options.

We thank the dozens of community members who have contributed to this discovery process - either formally or informally - by sharing their experiences in online events, github issues, 1:1 interviews or in survey form. We also thank the experts, researchers and fellows who have provided strategic input.

Frequently asked questions

What are the new definitions around language, variant and accent?

These are living definitions that will evolve with our community, and can be found in our public folder. Currently we propose;

Language: A shared system of spoken, signed, or written symbols through which a group of people can understand one another. For the purposes of Common Voice and ASR, languages would be driven for the most part by **mutual intelligibility**.

Variant: A **specific form** of a language or language cluster associated with a group of speakers, for example those living in a shared region or country, or who have a shared culture or heritage, and thus experience **vocabulary, grammar and norms** that

differentiate their speech from others. *Note that we use the word variants rather than dialects because the latter has sometimes been used in derogatory ways, however variants are not limited to geographies.

Accent: Pronunciation differences associated with a group or an individual speaker. This may take account of specific, complex or nuanced factors, such as tribal affiliation, other languages spoken or long periods of time spent in different places.

Why does it matter if language communities create lots of smaller and smaller datasets?

For one, it requires the platform to be re-localised repeatedly, which can be frustrating for volunteers. For another, it can confuse contributors, who are then unsure where they should be contributing.

Moreover for dataset consumers who are building ASR, in general, the more data the better. Models are trained to predict characters (or symbols) from frames (segments) of audio, and if we want machines to understand a greater diversity of speakers, whether first language or language learners - then data diversity is crucial.

But what if some language communities do have 'mutual intelligibility', but because of history, politics and culture, feel their languages are best categorised separately?

We completely understand this reality, and will always listen to the community and take account of social and historical experiences. Our intention is to give communities more options, not to force them along a particular path.

What if my language has fewer speakers and we don't feel we have variants?

That's fine! Again, the variant option is there to give communities more options - it's not a requirement.

How will variants be selected?

Communities will decide for themselves whether they need them, and which they would like. Using the new variants feature will be optional. We will provide guidance in the coming days on how they might go about this, but they will be the ultimate deciders. We encourage them to engage with their wider constituents, as well as linguistic experts, to make their decision.

To make sure we have a manageable solution in place, we will for now limit languages to 10 variants. Other information can be added through the accent tag.

What if we want to change our existing structure, for example by making a language a variant, or an accent a variant?

Our Community Manager and Linguistic Advisor will be reaching out to those language communities who might be interested in transitioning to be a variant in the coming weeks to help them understand their options, and support them to decide.

For every language that may wish to start creating new variants for their language, a clear mechanism for adding them will be shared via Discourse in the coming days. Again, the Community Manager and Product Team will be on-hand for anything communities need.

How should we use the accent tag in the meantime?

The accent tag will now be available for freeform additions in your profile. You may add as much information as you like. If you suspect that one of your accents is in fact a variant, you might wish to flag it to your wider community, but there is no need to remove it. We will be creating variants afresh, rather than migrating accents within the database.

The world needs your voice!