Keeping language rights at the heart of Common Voice

foundation.mozilla.org/en/blog/keeping-language-rights-at-the-heart-of-common-voice

October 27, 2021

With almost 80 language communities represented, Common Voice is today the biggest and most diverse multilingual voice dataset in the world, with contributors on every inhabited continent and a huge range of language groups. We support communities of all sizes – from Votic, with just 25 speakers, to English, with hundreds of millions.

We want every community to have equal opportunity to participate on the platform - and for the platform to perform equally well for every community. The final result should be a world in which everyone can interact with technology in any language they choose.

List of some of the launched langues on Common Voice. You can view the full list here: https://commonvoice.mozilla.org/languages

This mission obviously comes with a lot of challenges. When designing software, issues of multilinguality rarely come to the fore, and when they do it is usually taking into account either specific target languages or languages with which the team is familiar. For example, an improvement in font support may work for one language but not another, or a new guideline may improve the performance of the dataset for one language but may reduce it drastically for another.

So the process of supporting language diversity often becomes **reactive**, with patches happening after the change has been made, if and when a bug report comes in.

In order to try and make the process more **proactive**, we adopted the methodology of the impact assessment. This is a procedure that determines the impact of development decisions on language diversity on the platform.

A concrete example: For some time there have been requests to give better contribution criteria for validation and make them available through the site. We recently designed and deployed these in English and made them available for word-by-word translation into other languages. Whilst the guidance added a lot of value for some, in other contexts it didn't make any sense. For example, one of the criteria was that there shouldn't be an *-s* missing at the end of a word (*dinosaur* instead of *dinosaurs*). Although this criterion makes sense in English, it does not make sense in many of the other 82 languages - for example the final *-s* in French is most often unpronounced (e.g. *dinosaure* and *dinosaures* are pronounced identically). Had the language diversity assessment been in place during this feature development process, then perhaps we might have taken a different approach to localising the guidelines.

Language diversity impact assessment

How does it work?

The language diversity impact assessment is a two step process that can be run before any decision on a new proposal. The first step is that we ask ourselves a series of questions related to issues of language diversity to try and determine if there are any gaps in the proposal or any potential issues. A short report is written based on this survey to unpack the impact on different kinds of languages. Based on this report a final assessment is made.

I led the design of this survey based on a review of all GitHub issues and posts on Discourse for Common Voice, along with a series of interviews with members of the community, which includes language rights activists, machine learning experts, and software engineers.

What does it look like?

There are four axes to the survey, each of which contains several questions designed to probe the impact of potential changes.

1. **Localisation**: Everything related to the translation and adaptation of the Common Voice interface(s) for a local audience, that is for specific language communities.
2. **Resources**: Different languages have different amounts of resources available in terms of text, linguistic processing tools and available community expertise. A proposal should outline requirements clearly and propose mitigation strategies for languages that do not meet the requirements.
3. **Visibility and opportunity**: Visibility is about ensuring that all languages are equally visible and about respecting user choice in terms of people's preferred language, in other words, giving the same level of opportunity to all languages.
4. **Testing diversity**: When choosing a set of languages to develop a prototype for, languages in the selection should exhibit diversity in terms of geographic area, language typology, amount of resources and technical requirements like writing system, direction of text etc.

The final recommendation on each of these categories follows a *traffic-light system*, where each axis is assigned an evaluation of red, amber or green. Going back to our previous example, the English-centric design would have received a red traffic light for both the **localisation** and **testing diversity** axes. This would have given us a chance to reflect on the impact of the design before implementation and deployment.

Where next? We have started to roll out the impact assessment internally on recent changes, and received feedback from communities at the Common Voice Contribute-a-thon sessions. We encourage anyone interested to participate and share their thoughts about their process by emailing commonvoice@mozilla.org !