TRANSLATORS
WITHOUT BORDERS

THE COMMON LANGUAGE INITIATIVE

Solutions for Underserved Languages

#languagematters

Abstract

Advancements in language automation are transforming the way the world communicates – but what about languages that lack a digital presence? The Common Language Initiative focuses on bringing underserved languages online to serve the world's most marginalized populations.

Aimee Ansari
Executive Director, Translators without Borders
Rebecca Petras
Deputy Director, Translators without Borders

Executive summary

Access to information in a language an individual understands is a human right[1]. From health care and life skills to education and economic empowerment, language understanding is fundamental to the human experience. Language connects us—to each other, to our communities and to opportunities in the world at large.

Today, rapid advancements in artificial intelligence and machine learning are transforming how we communicate. Machine-generated text or speech translation is faster and more accurate than ever, making it easier to do business in new markets, recruit diverse talent or travel abroad. Language, once a roadblock, is now a game-changer.

Yet even with the power and potential of this technology, hundreds of millions of the world's poorest, least educated, most vulnerable populations are being left behind. This stark reality reveals not just an investment bias towards the world's most commercially relevant languages—major Western and Asian languages such as Chinese, French, German and Japanese. Rather, this gap also reflects a massive data challenge: language automation requires human input—massive amounts of original and translated information, in a digital format. Such parallel data doesn't exist for many languages.

English as a 'global language' isn't the answer either. Less than ten percent of the world's language speakers are proficient in English, a number that is expected to decline to five percent by the year 2050[2]. Investing in native languages also preserves culture and history.

To bridge this digital divide, Translators without Borders (TWB), a non-profit organization that provides translation services for humanitarian purposes around the globe[3], is sponsoring the Common Language Initiative (CLI), which focuses on bringing language technology to bear for the world's most marginalized communities. Over the next decade, Translators without Borders and its coalition of partners will bring twenty or so new underserved languages online. The Common Language initiative will make resettlement, healthcare and skills-building information available in these languages, providing a sustainable and free asset to those who need it most.

This paper outlines the Common Language Initiative strategy, operating framework, and pilot projects. Translators without Borders welcomes your feedback and commentary at any time (info@translatorswithoutborders.org).

What is the Common Language Initiative?

The Common Language Initiative was introduced in March 2017 at the Humanitarian Information and Communications Technology Forum, in a session called *Climbing the Tower of Babel: Eliminating the Language Barrier in Response*, which was sponsored by the United Nations' High Commissioner of Refugees (UNHCR) and Office for the Coordination of Humanitarian Affairs (UNOCHA) and moderated by Translators without Borders.

In this session, Translators without Borders proposed the creation of a free repository of spoken and written data for underserved languages, an idea that grew out of Translators without Borders's long history of delivering translation and interpreting services, often under extraordinary circumstances. Rather than building new languages one at a time, Translators without Borders advocated for a scalable, repeatable and continuously improving model that could make rapid progress on multiple languages at once. In the months since, Translators without Borders's proposal has been endorsed by thought leaders across the technology, linguistic, native speaker and humanitarian assistance communities, including UNHCR, UNOCR, the International Federation of the Red Cross and Red Crescent Societies, Google, Microsoft, Facebook and many more.

[1] Universal Declaration of Human Rights (UDHR), Article 2.

[2] "English in Decline as a First Language," National Geographic News (Feb 2004).
February 26, 2004 See https://news.nationalgeographic.com/news/2004/02/0226_040226_language.html

[3] For more about Translators without Borders work, please see translatorswithoutborders.org.

The Common Language Initiative is designed to alleviate suffering, improve access to healthcare and support education and job skilling for hundreds of millions of the world's most marginalized populations including:

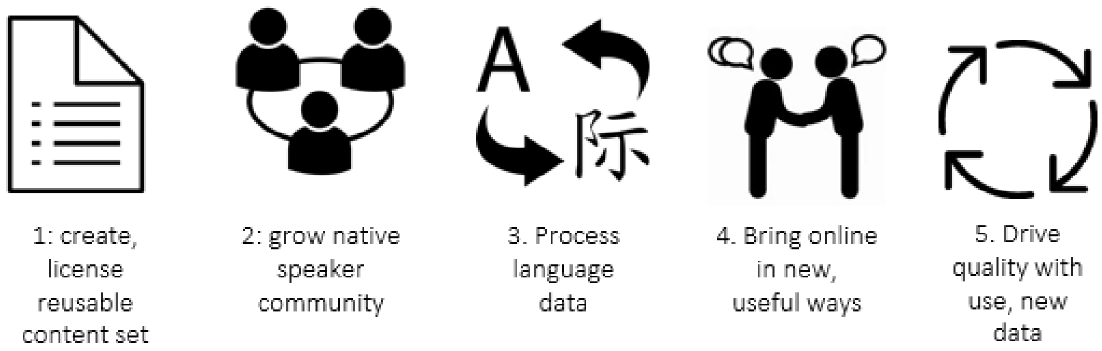
- Refugees from Syria, Myanmar and Somalia fleeing civil strife;
- Communities impacted by natural disasters and reliant on international aid workers with whom they don't share a language, such as been the case in Haiti, Nepal and Bangladesh;
- Local health care workers, racing to provide accurate, life-saving information in indigenous languages during deadly, fast-moving pandemics such as Ebola and Zika, or during outbreaks of cholera; and
- Immigrants dependent on government and aid agencies for help resettling in new communities, schools and jobs in Europe and the United States.

The Common Language Initiative model

For language automation to make a significant impact for underserved languages, the methodology, process and workflow must be efficient, cost-effective and—most critically—replicable. The Common Language Initiative has been designed with this repeatable model in mind while also taking into account the end-to-end challenges associated with gathering language data and making it useful.

To start, Translators without Borders and its editorial and content communities are creating a basic content set that may be used as the starting point for many languages. From this core, additional content sets may be added based on local needs – health care for a specific disease, for example, or vocabulary for refugees resettling in a new country. During the translation process, these content sets will also be localized – tailored to the specific language or region – to ensure accuracy and relevance.

Here's how it works:


A
际
1: create, license reusable content set
2: grow native speaker community
3. Process language data
4. Bring online in new, useful ways
5. Drive quality with use, new data

- Step 1: create/license a test set of simple and accurate content that can be reused for any/all languages.
- Step 2: recruit, train and incent communities of translators, able to translate and localize content for their community.
- Step 3: work with technology partners who host manage large-scale language data as well as creating new language engines.
- Step 4: partner with local developers and humanitarian agencies to bring each new language online in ways that are immediately useful.
- Step 5: with translation communities and content providers, drive continuous improvement in language accuracy and quality through every day use and the addition of new content sets driven by local needs.

Getting started: The Common Language Initiative pilots for 2018

Work is already underway to test the Common Language Initiative model with a set of pilots that will kick off early in 2018. During each phase, Translators without Borders staff will monitor and resolve issues, manage resources and schedules, and communicate progress against goals. Periodic reviews will ensure real-time learnings are continuously incorporated into each work stream as appropriate.

- Pre-pilot readiness: finalize goals, scope, timeline, partner organizations and content sources – IN PROCESS NOW.
- Phase one: kick off first two languages, Bengali and Hausa[4].
- Phase two: extend the Common Language Initiative model to Kiswahili[5] while also beginning to test early builds of Bengali and Hausa with partner organizations.
- Phase three: deliver the plan to extend the Common Language Initiative model to three or four new languages per year while also adding additional content sets for all languages underway.

Each phase of the CLI pilot is expected to last nine-to-twelve months.

The role of partner organizations

The Common Language Initiative's success relies on diverse partnerships from across the native speaker, linguistic, technology and humanitarian sectors. Our partners offer strong support for the Common Language Initiative vision and scope, and actively participate in the Common Language Initiative by providing engineering expertise, in-kind donations for content and tools and access to volunteers who fill many roles for the Common Language Initiative.

Please see Appendix A for more detail.

Funding

Translators without Borders is managing the process to secure funding for the pilot as outlined above. In addition, Translators without Borders and its advisors at UNHCR's Innovation Fund are working to develop a sustainable financial model, which Translators without Borders envisions will be fund by its diverse group of worldwide donors, partner organizations and a long-term international development grant.

Risks

The Common Language Initiative is an ambitious, multi-year effort that requires many kinds of support. As with any project, there are many dependencies that may impact the success of the Common Language Initiative including:

- Native speakers must agree to contribute data and expertise, and to receive appropriate compensation for their intellectual property.
- Recruiting, training and keeping translator communities engaged requires ongoing investment.
- Technology partners must agree to devote resources to processing the data and to building and training new engines.
- Local language authorities may need to be engaged as part of the language development process.
- Partner organizations must be prepared to use new automated languages as available, especially since use drives quality over time; and
- Once available, the bilingual corpora created by the Common Language Initiative must remain free for use by all.

Call to action

We ask for your feedback and support for the Common Language Initiative vision and plan. There are numerous ways to show your support:

[4] Bengali and Hausa were chosen as the pilot languages based on the following criteria: humanitarian need, number of speakers worldwide (combined, Bengali and Hausa have 245M speakers worldwide), mobile phone use and Translators without Borders's ability to develop and manage local community/partnerships.

[5] Baseline language automation (text-based machine translation) exists for Kiswahili and its 140M speakers worldwide. The Common Language Initiative model will be used to grow the quality and usefulness of Kiswahili text translation while also building new assets for speech recognition/translation.

- Join our partner coalition and be an active voice in the ongoing Common Language Initiative discussion.
- Contribute healthcare, crisis relief, education and/or job skilling content to grow the usefulness of the data set.
- Share parallel data in Bengali, Hausa and/or Kiswahili to help seed our work.
- Volunteer your expertise in localization, project management and/or engineering to help Translators without Borders keep the data fresh and relevant for the long-term.
- Follow our progress on Facebook , Instagram and Twitter.
- Donate to Translators without Borders to support the Common Language Initiative.

Thank you for your support.

Appendix A: partner organizations

This list reflects a partial list of the Common Language Initiative partner organizations.

Organization
Acrolinx
Amazon
American Red Cross
The Barefoot Guides
BBC Media Action
CDAC-Network members
The Cisco Foundation
CrowdFlower
Facebook
Google
impact.org
Language Connect
Media Action
Meedan
National Institutes of Health (NHI)
International Federation of the Red Cross and Red Crescent Societies
Microsoft Corporation
Shonjog
StoryWeaver
Translation Automation User Society (TAUS)
United Nations High Commissioner for Refugees (UNHCR) Innovation Fund
United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA)
World Health Organization (WHO)