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 IP—massive amounts of parallel data[2] in a useable (digital) format, which, for many languages, simply doesn't exist.

To bridge this digital divide, Translators without Borders (TWB), a non-profit organization that provides humanitarian translation services around the globe, is sponsoring the Common Language Initiative (CLI), a partnership effort that, for the first time, is bringing language technology to bear for the world's most marginalized communities. Over the next decade, TWB and its coalition of partners will bring some twenty new underserved languages online, creating a useful, sustainable and free asset to alleviate suffering and create opportunity for those who need it most.

This paper outlines the CLI strategy, operating framework, and pilot projects. TWB welcomes your feedback and commentary by 15 January 2018.

What is the CLI?

The CLI was introduced in March 2017 at the Humanitarian ICT 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 TWB.

In this session, TWB proposed the creation of a free repository of spoken and written data for underserved languages, an idea that grew out of TWB's long history of delivering translation and interpreting services, often under extraordinary circumstances. Rather than building new languages one at a time, TWB advocated for a scalable, repeatable and continuously improving model that could make rapid progress on multiple languages at once. In the months since, TWB'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.

The CLI 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;

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

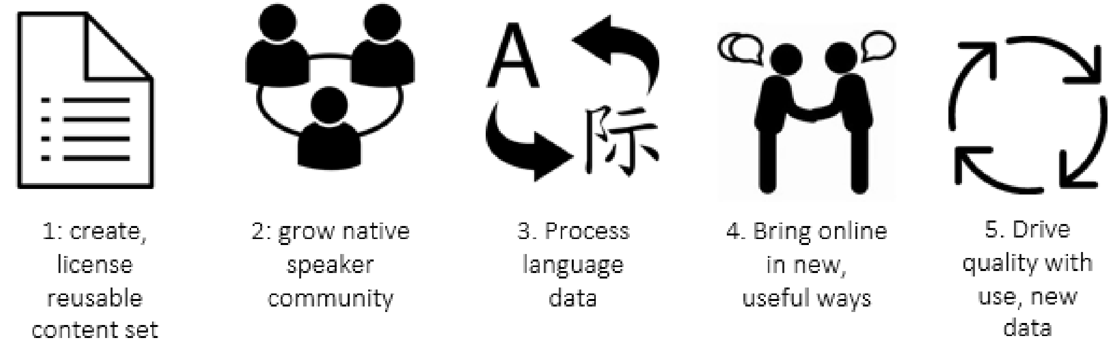
[2] Parallel data is the same exact information in both source and target languages. Parallel data is the foundation for building spoken and written automated languages.

- 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 CLI 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 CLI 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.

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.

Getting started: the CLI pilots for 2018

Work is already underway to test the CLI model with a set of pilots that will kick off early in 2018. During each phase, TWB 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[3].
- Phase two: extend the CLI model to Kiswahili[4] while also beginning to test early builds of Bengali and Hausa with partner organizations.

[3] 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 TWB's ability to develop and manage local community/partnerships.

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

- Phase three: deliver the plan to extend the CLI model to three or four new languages per year while also adding additional content sets for all languages underway.

All phases of the CLI pilot are expected to last nine-to-twelve months.

The role of partner organizations

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

Please see Appendix A for more detail.

Funding

TWB is managing the process to secure funding for the pilot as outlined above. In addition, TWB and its advisors at UNHCR's Innovation Fund are working to develop a sustainable financial model, which TWB envisions will be fund by its diverse group of worldwide donors, partner organizations and a long-term international development grant.

Risks

The CLI 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 CLI 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, each automated language must remain free for use by all.

Call to action

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

- Join our partner coalition and be an active voice in the ongoing CLI 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 TWB keep the data fresh and relevant for the long-term.
- Follow our progress on Facebook , Instagram and Twitter.
- Donate unrestricted funding to Translators without Borders to support the CLI.

Thank you for your support.

Appendix A: partner organizations

This list reflects a partial list of the CLI partner organizations.

Organization	Role	Status
Acrolinx	Content tools	
Amazon	Language technology	
American Red Cross	Crisis relief, healthcare content	
The Barefoot Guides	Content	
BBC Media Action	News, information content	
CDAC-Network members		
The Cisco Foundation	Pilot funding	
Facebook		
Google	Language technology	
impact.org	Cloud development, hosting	
Media Action	Content tools	
Meedan	Community building	
National Institutes of Health (NHI)	Healthcare content	
International Federation of the Red Cross and Red Crescent Societies	Crisis relief, healthcare content	
Microsoft	Language technology	
Shonjog	Community building	
StoryWeaver	Content tools	
Translation Automation User Society (TAUS)	Cloud development, hosting	
UNHCR Innovation Fund	Cloud development, hosting	
UNOCHA	Needs assessment	
World Health Organization	Healthcare content	