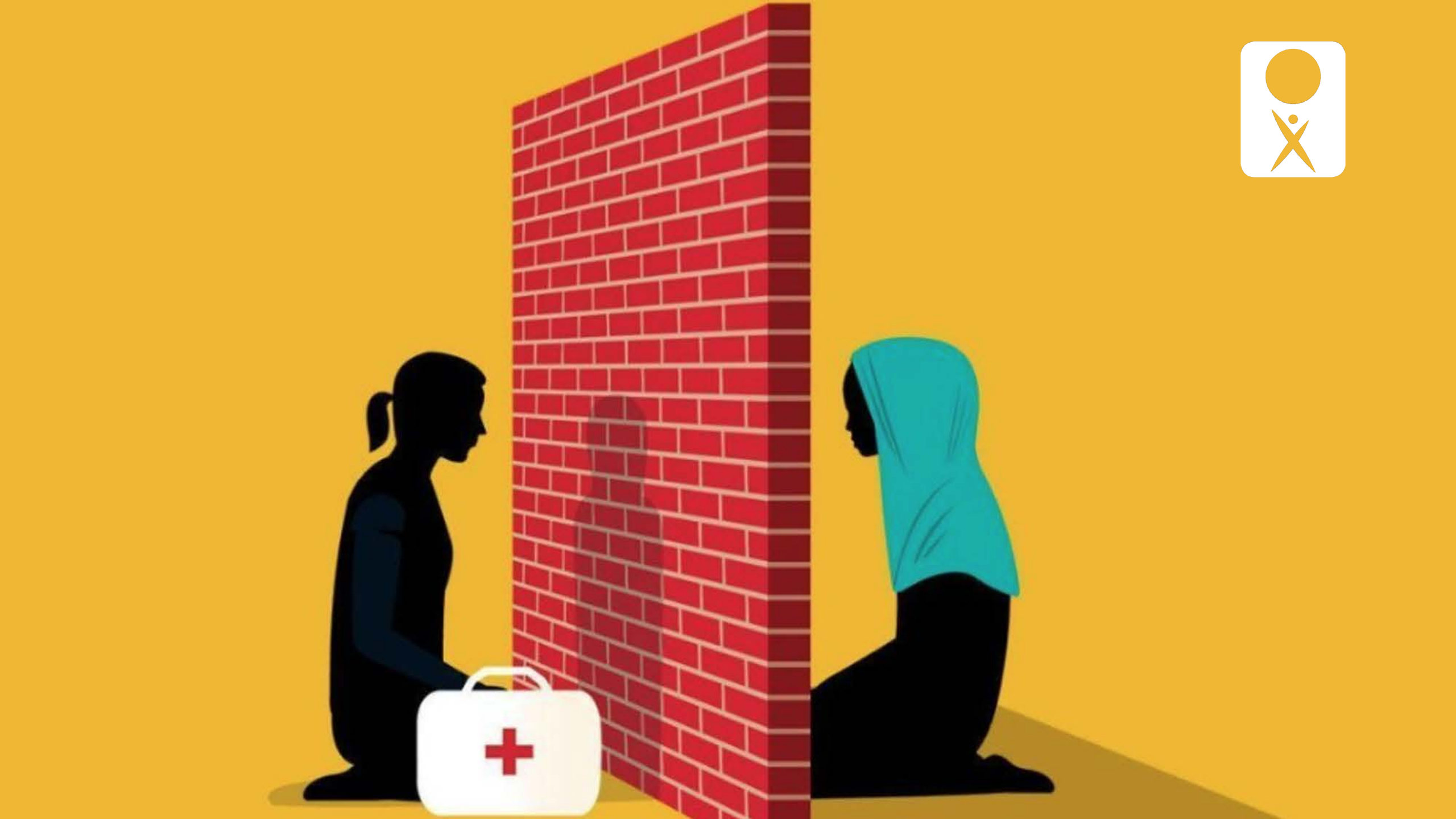
Translators without Borders: Using language technology to enable two-way communication in humanitarian assistance

Eric Paquin
Chief Technology Officer
Translators without Borders

Creating a world where knowledge knows no language barriers



In a life or death situation:

Air pollution is under your control. Put it on top of your head and tie it to your body. You can only ventilate the air when you leave the boat. *(Hausa - 44 million)*

Your life jackets are under your seat. This is putting your head to form on your own. Cotton. After leaving the plane on your life jackets in the air breathed. *(Burmese - 32 million)*

Your life jacket is under your seat. Pass it over your head and tie it around the body. Only swallow the life vest after leaving the plane. *(Turkish, 75.7 million)*

When there’s enough data, you get:

Your life jacket is under your seat. Put it on your head and attach your shoulder strap. Once you leave the plane, you should only inflate the vest. *(traditional Chinese, 1.2 billion)*

You will find a life jacket under your seat. Put it over your head and trim your straps. Inflate the vest only after leaving the aircraft. *(Lithuanian, 3.2 million)*

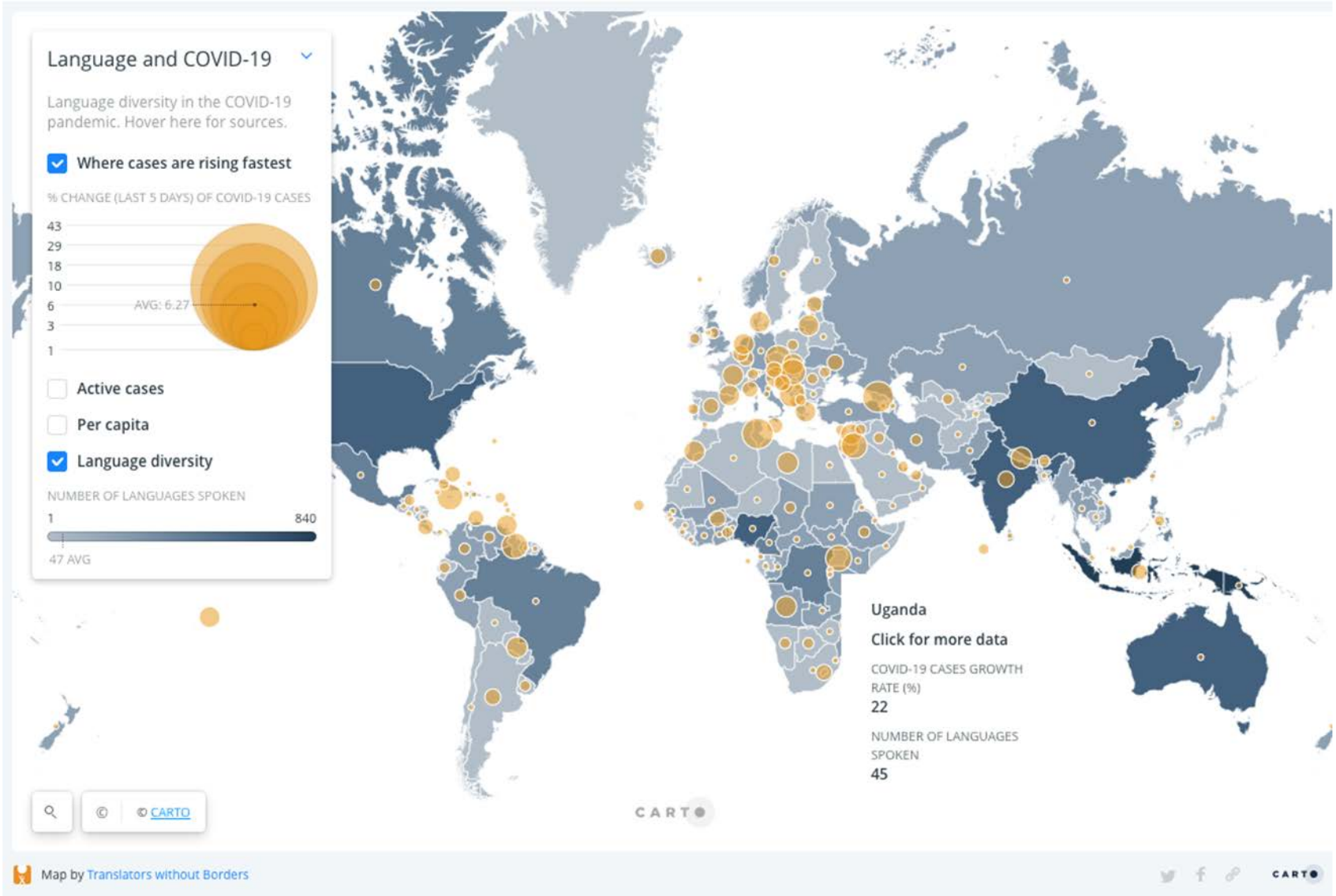
Your life jacket is under your seat. Place it over your head and tie the straps around you. Inflate your life jacket after leaving the aircraft. *(Greek, 13.4 million)*

And enabling real-time two-way communication

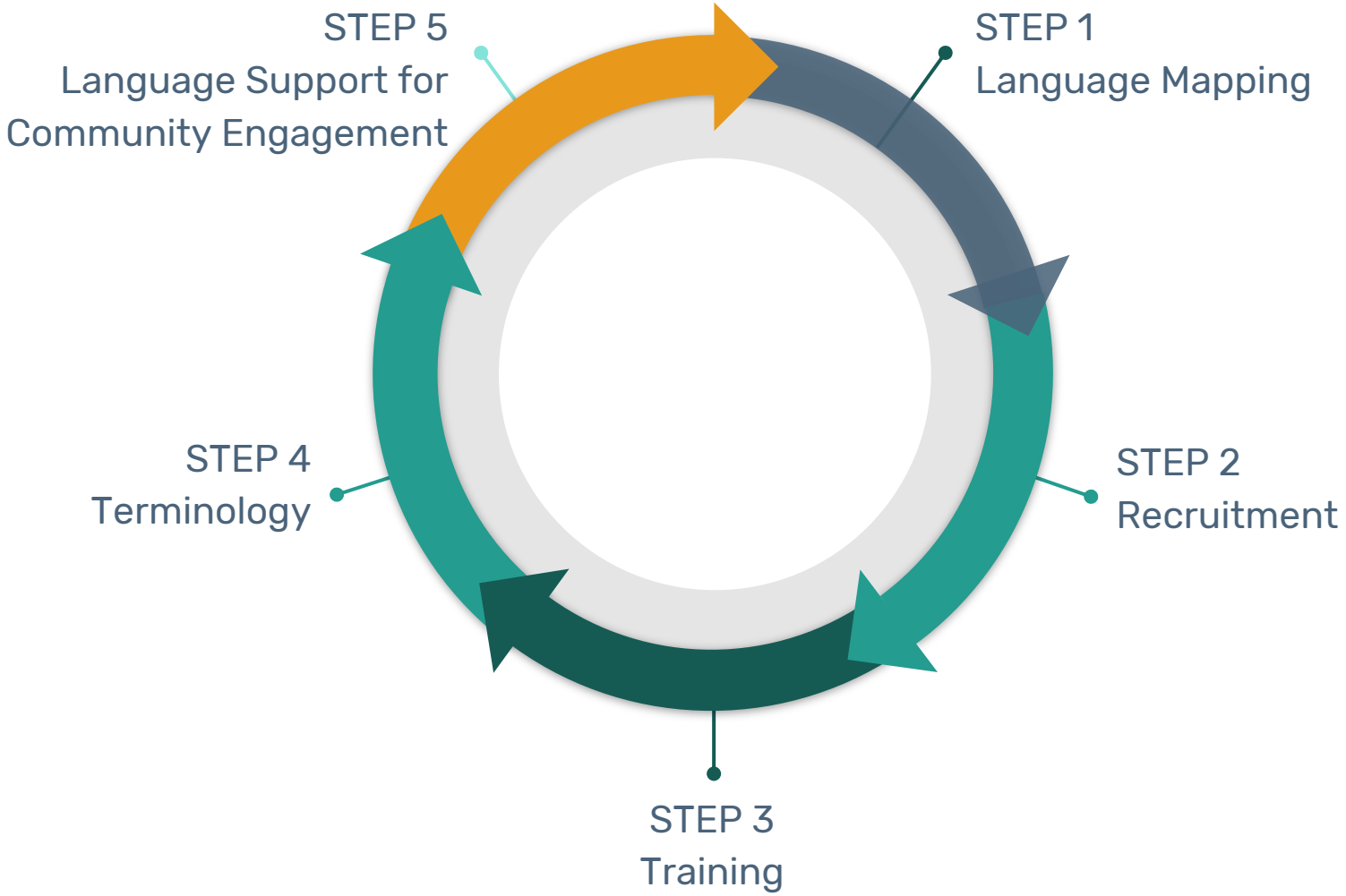
Today

- TWB and language technologies
- Towards a multilingual future
- Call to Action

Multilingual emergencies



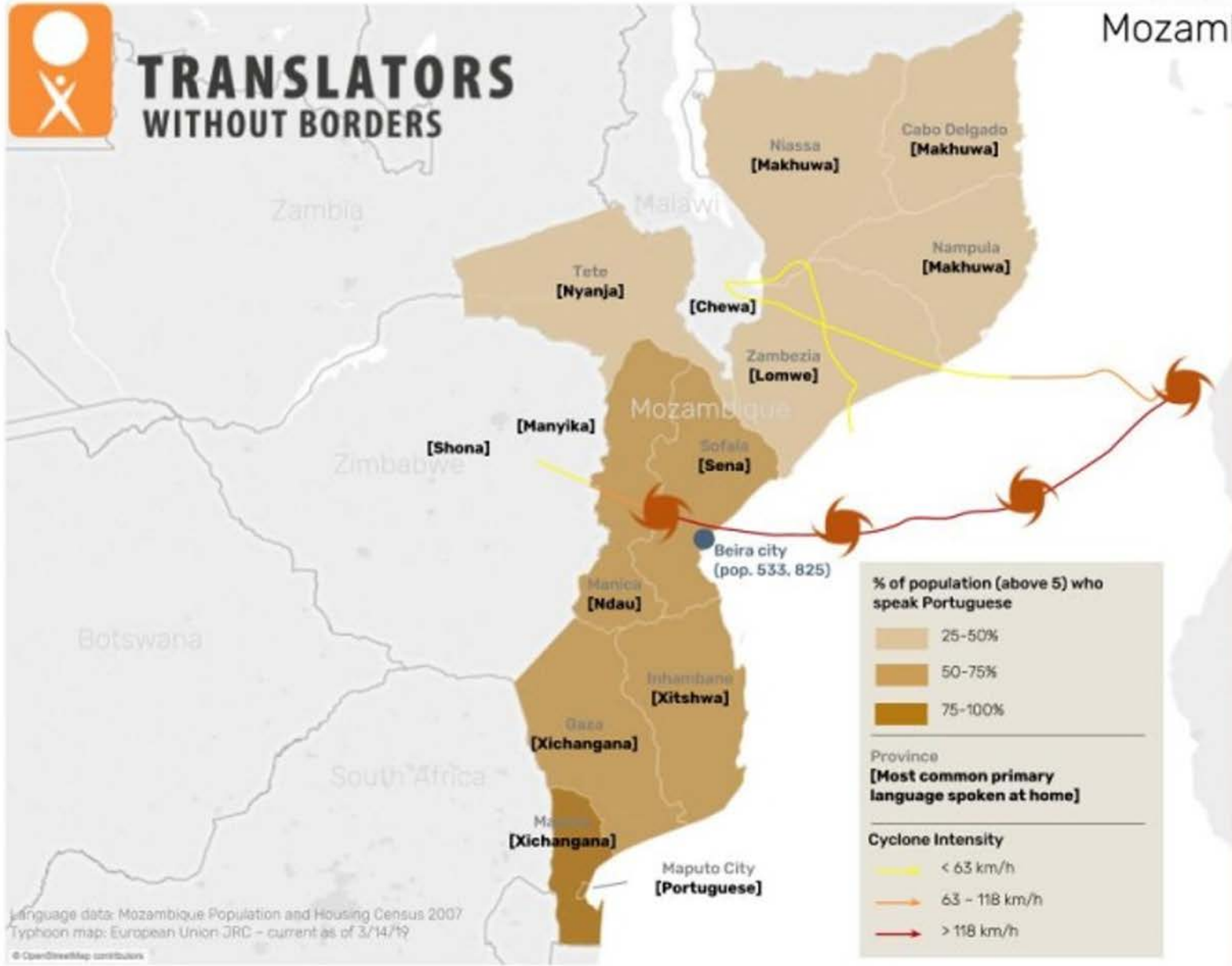
Crisis response



CRISIS LANGUAGE MAP

Mozambique – Cyclone Idai

17 March 2019



Over 40 languages are spoken in Mozambique.

Only half of the population can speak Portuguese, the official language of Mozambique.

Rates of comprehension are especially low in rural areas and amongst women.

Other languages spoken in the most affected provinces include:

- Sena (over 1.5 million speakers)
- Lomwe (4th largest language nationally)
- Xitshwa (a Tswa-family language)
- Ndau (similar to Shona in Zimbabwe)

Rates of illiteracy in Mozambique are high, especially amongst women:

- Sofala (23% men / 62% women)
- Zambezia (44% men / 79% women)
- Manica (24% men / 60% women)
- Tete (39% men / 72% women)

Responders should prioritize audio and pictorial communication, particularly when targeting women in rural areas.

For language support:
crisis@translatorswithoutborders.org

2019 Achievements

23,000,000 words translated

37,000 linguists around the world

8,000 projects

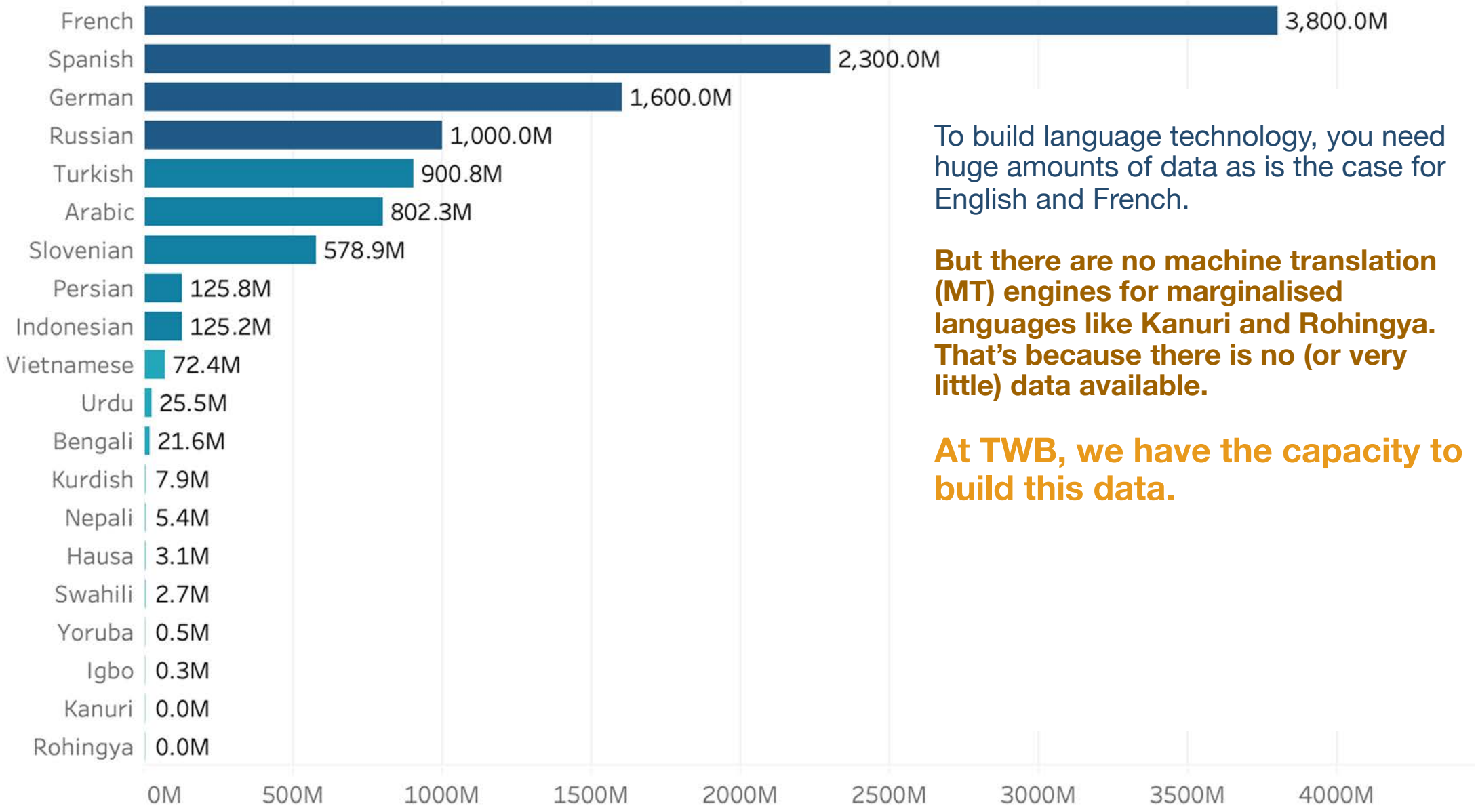
200 language combinations

160 NGO partners

TWB and Language technologies

Rohingya crisis

Size of publicly available language datasets

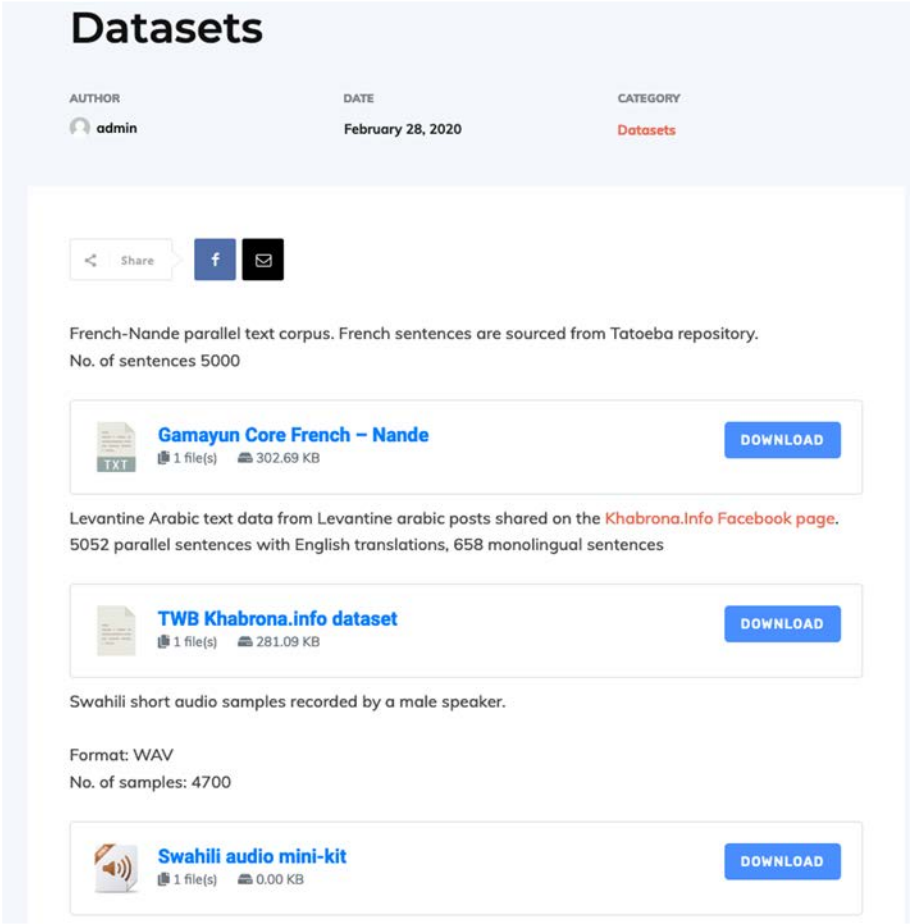


To build language technology, you need huge amounts of data as is the case for English and French.

But there are no machine translation (MT) engines for marginalised languages like Kanuri and Rohingya. That's because there is no (or very little) data available.

At TWB, we have the capacity to build this data.

Gamayun Corpora



Datasets

AUTHOR: admin
DATE: February 28, 2020
CATEGORY: Datasets

Share

French-Nande parallel text corpus. French sentences are sourced from Tatoeba repository.
No. of sentences 5000

Gamayun Core French – Nande
1 file(s) 302.69 KB
DOWNLOAD

Levantine Arabic text data from Levantine arabic posts shared on the Khabrona.Info Facebook page.
5052 parallel sentences with English translations, 658 monolingual sentences

TWB Khabrona.info dataset
1 file(s) 281.09 KB
DOWNLOAD

Swahili short audio samples recorded by a male speaker.

Format: WAV
No. of samples: 4700

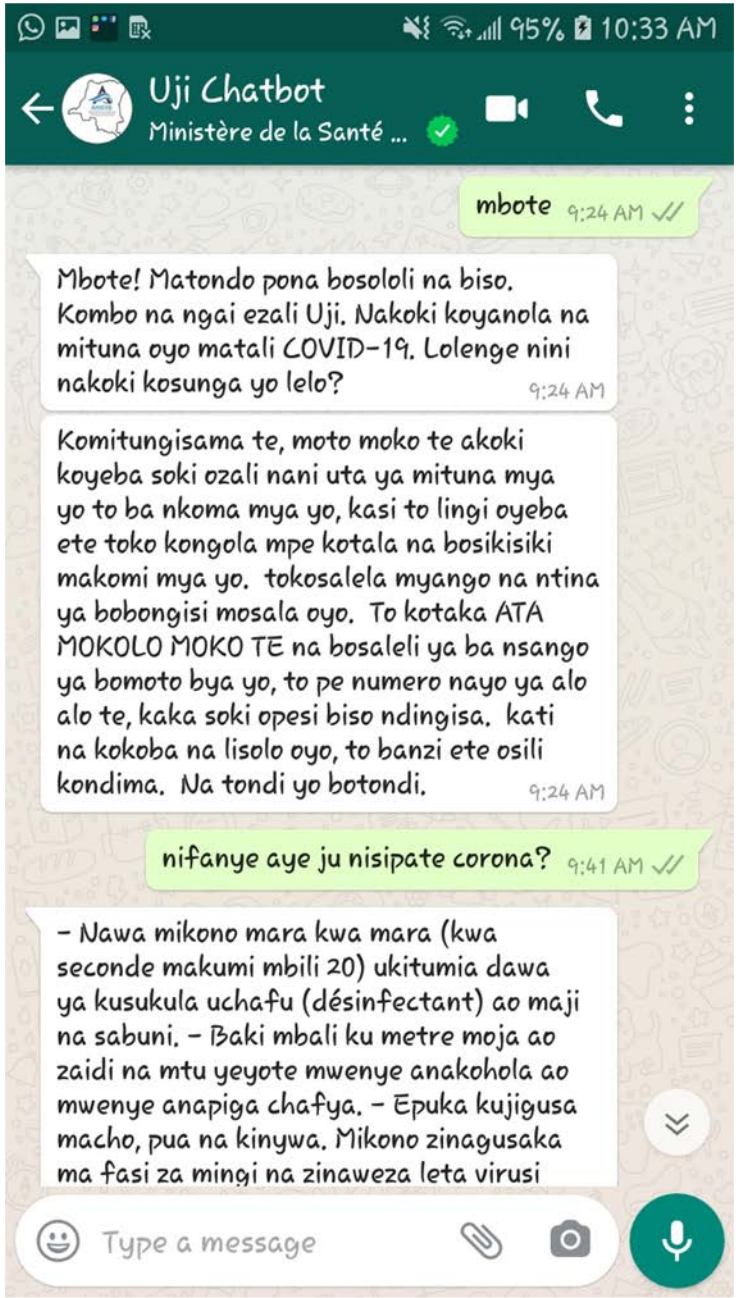
Swahili audio mini-kit
1 file(s) 0.00 KB
DOWNLOAD

https://gamayun.translatorswb.org

Chatbots

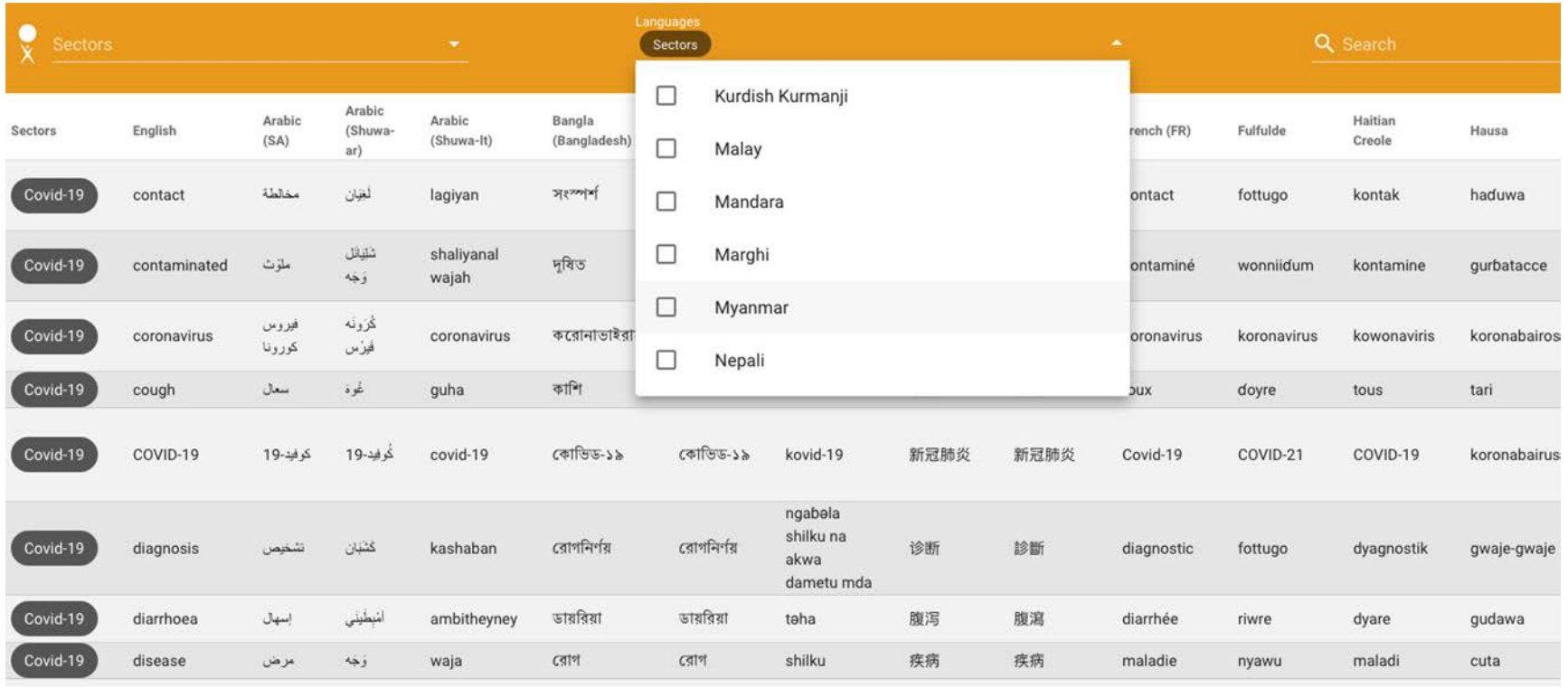
With COVID-19 movement restrictions, we need new ways for people to access information about the disease, in their own language

Meet Uji: TWB's **multilingual chatbots** using natural language processing

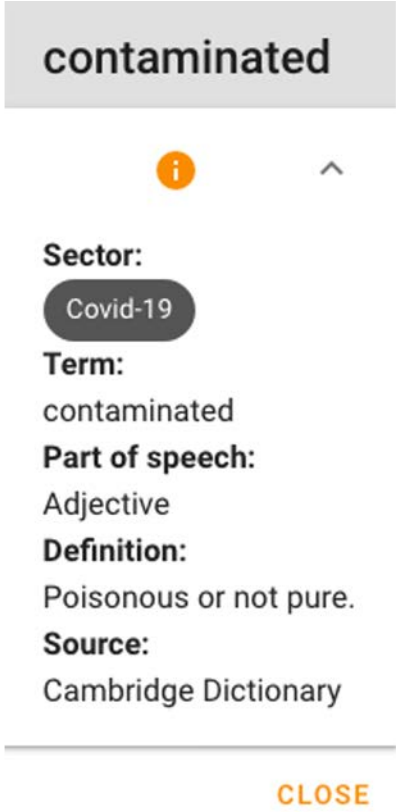


MT/INMT Models

Multilingual and augmented glossaries



Sectors	English	Arabic (SA)	Arabic (Shuwa-ar)	Arabic (Shuwa-lt)	Bangla (Bangladesh)					French (FR)	Fulfulde	Haitian Creole	Hausa
Covid-19	contact	مخالطة	لِغيَان	lagiyan	সংস্পর্শ					ontact	fottugo	kontak	haduwa
Covid-19	contaminated	ملوّث	شَلِيَانَل وَجَه	shaliyanal wajah	দূষিত					ontaminé	wonniidum	kontamine	gurbatacce
Covid-19	coronavirus	فيروس كورونا	كُرُونَه فِيرُس	coronavirus	করোনাভাইরা					oronavirus	koronavirus	kowonaviris	koronabairos
Covid-19	cough	سعال	غُوة	guha	কাশি					oux	ɗoyre	tous	tari
Covid-19	COVID-19	كوفيد-19	كُوفِيد-19	covid-19	কোভিড-১৯	কোভিড-১৯	kovid-19	新冠肺炎	新冠肺炎	Covid-19	COVID-21	COVID-19	koronabairus
Covid-19	diagnosis	تشخيص	كَشْبَان	kashaban	রোগনির্ণয়	রোগনির্ণয়	ngabəla shilku na akwa dametu mda	诊断	診斷	diagnostic	fottugo	dyagnostik	gwaje-gwaje
Covid-19	diarrhoea	إسهال	اَمْبِطِيْنِي	ambitheyney	ডায়রিয়া	ডায়রিয়া	təha	腹泻	腹瀉	diarrhée	riwre	dyare	gudawa
Covid-19	disease	مرض	وَجَه	waja	রোগ	রোগ	shilku	疾病	疾病	maladie	nyawu	maladi	cuta



TICO-19

Translation Initiative for COVID-19

Providing machine-readable translation data related to the COVID-19 pandemic

Translation Memories for the Translation Community

We have combined the terminologies and other translation data to create translation memories in .tmx format for the majority of the language pairs.

- Additional details and data download here.

Translated Terminologies

Translations of covid19-related terms in dozens of languages and locales, provided by Facebook and Google.

- Download all terminologies here (zip file, 0.8MB).
- Complete list and details here.

TICO-19 Translation Benchmark

The benchmark will include 30 documents (3071 sentences, 69.7k words) translated from English into 36 languages: Amharic, Arabic (Modern Standard), Bengali, Chinese (Simplified), Dari, Dinka, Farsi, French (European), Hausa, Hindi, Indonesian, Kanuri, Khmer (Central), Kinyarwanda, Kurdish Kurmanji, Kurdish Sorani, Lingala, Luganda, Malay, Marathi, Myanmar, Nepali, Nigerian Fulfulde, Nuer, Oromo, Pashto, Portuguese (Brazilian), Russian, Somali, Spanish (Latin American), Swahili, Congolese Swahili, Tagalog, Tamil, Tigrinya, Urdu, Zulu.

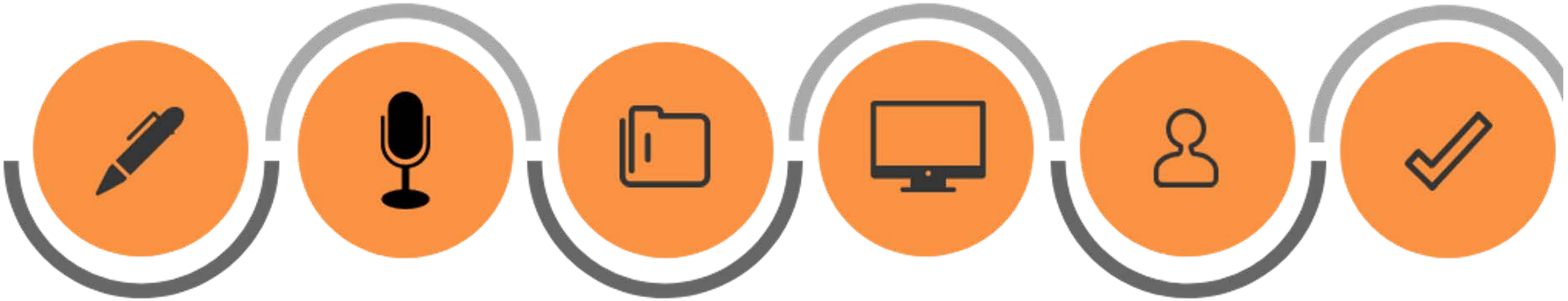
- Download the evaluation set here (zip file, 12MB).
- Paper describing the creation of the dataset here.
- Paper, Results, and more details here.

Voice & MT

Towards a multilingual future

Tiles

Building our future



Translate basic content
(Gamayun core corpus)

Record audio content

Collect parallel data
(TWB, other sources, etc)

Develop and test model

Develop interface

Integration
Value added

Voice MT and Phonetic MT

Call to action

Opportunities

Data sharing

Collaboration: private and aid sector

2-way communication

Broader usage beyond humanitarian

Core to TWB's vision for the future is Gamayun: the language equality initiative

A collaborative effort with tech partners, native-speaker communities, NGOs, and content creators and owners, Gamayun uses text and speech data to automate translation in under-served languages.

Using a repeatable, scalable model, TWB and its coalition of partners are developing state-of-the-art voice and text tools in languages such as Congolese Swahili, Tigrinya, Kanuri and Rohingya

It's a revolutionary way to use cutting-edge technology for crisis communication, closing the language equality gap and enabling two-way communications.

With the help of our partners, we're making our vision come true.

cisco Foundation

Microsoft

twilio

KoBoToolbox

KINSHASA DIGITAL

Call to action

- Share low resource data
- Collaborate with us
- Research, workshop
- Donate

TWB's support makes an impact

In 2017, **23%** of Rohingya refugees said they had enough information to make good decisions for themselves and their families

TWB's support makes an impact

In 2017, **23%** of Rohingya refugees said they had enough information to make good decisions for themselves and their families

By early 2019, thanks in part to TWB, **84%** felt they had info on available assistance

“When I speak my own language,
I am free.

When I hear someone else speaking Rohingya,
I feel like I am home.”

– Rohingya person

Thank you!