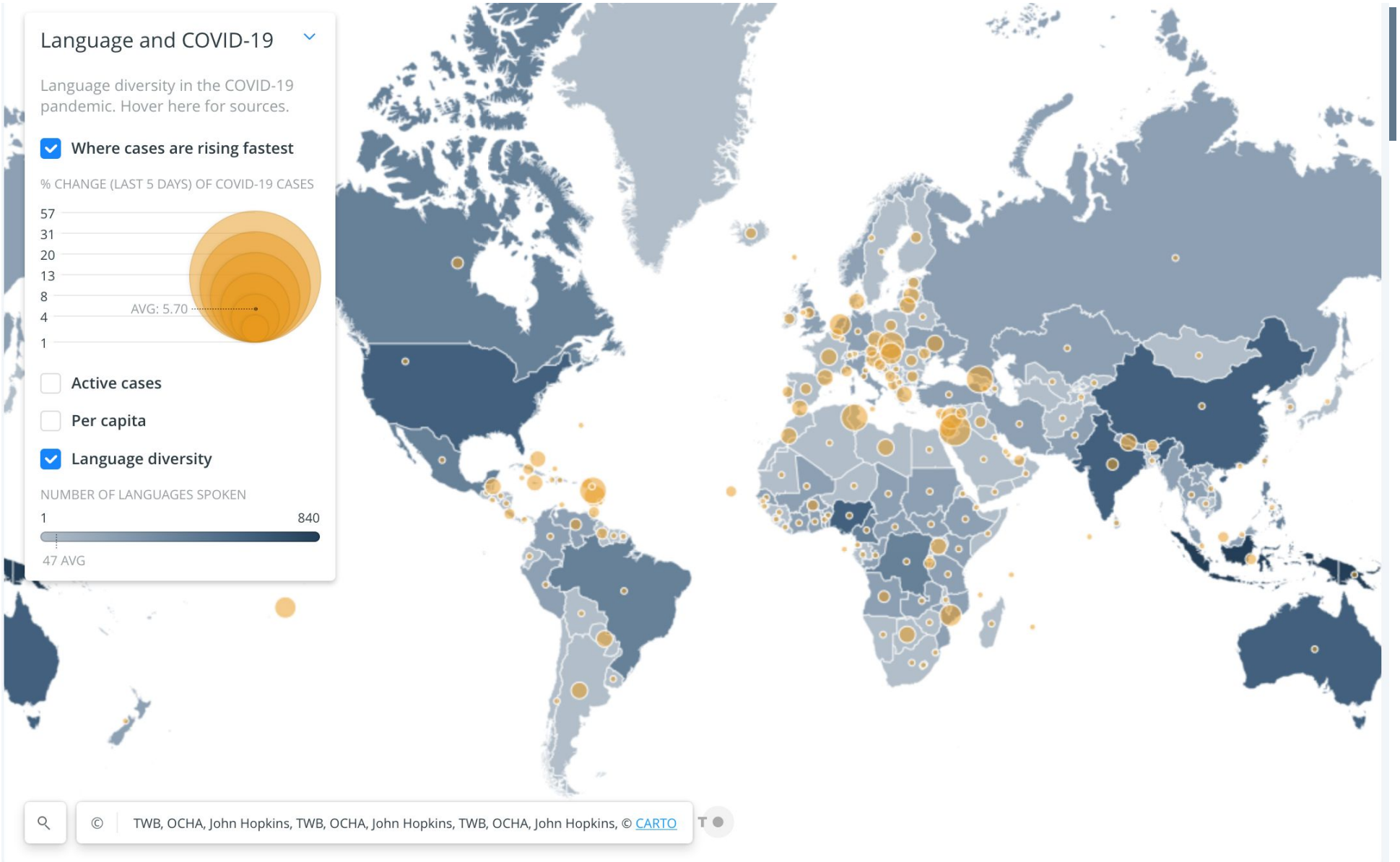
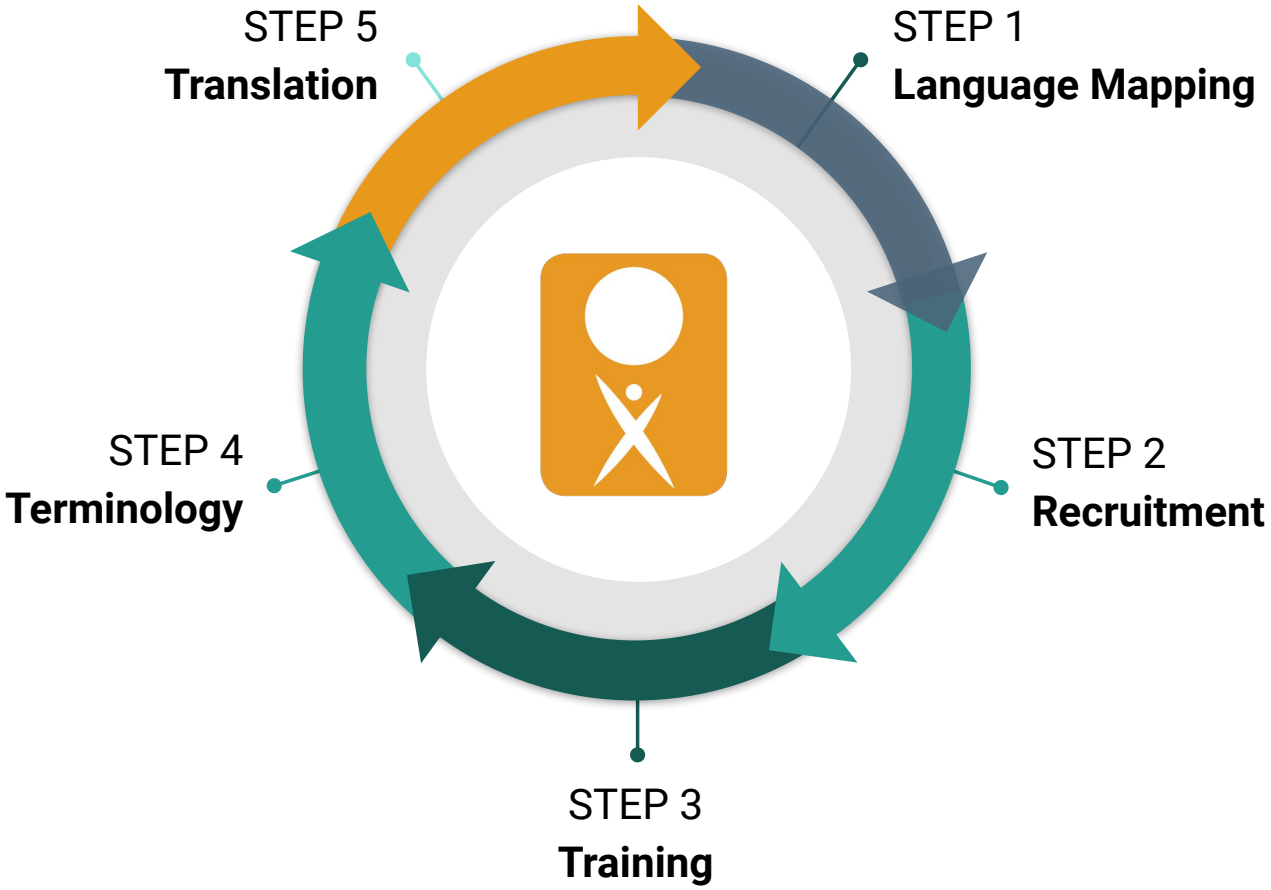
Empowering translators of marginalized languages through the use of language technology

Alp Öktem, Manuel Locria, Eric Paquin, Grace Tang

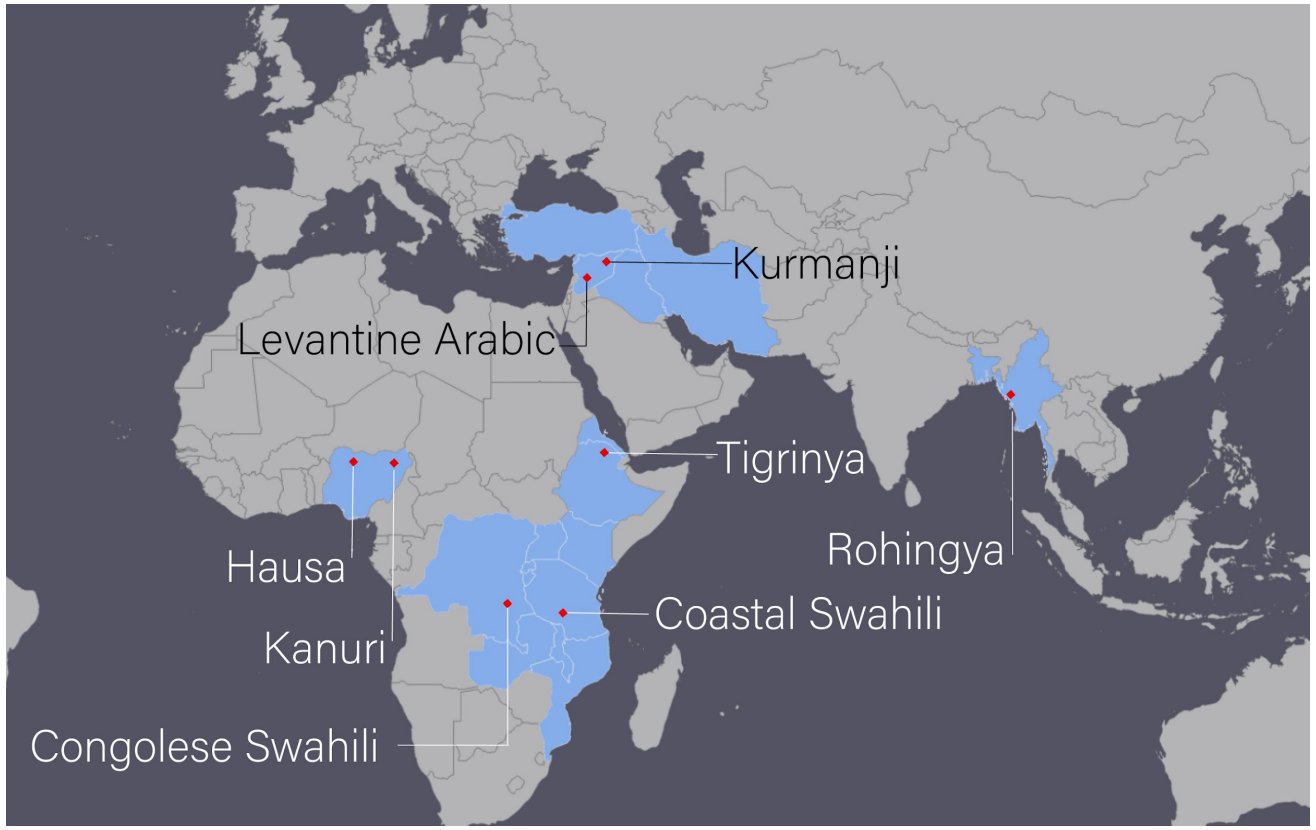
TRANSLATORS
WITHOUT BORDERS


Language and COVID-19
Language diversity in the COVID-19 pandemic. Hover here for sources.
Where cases are rising fastest
% CHANGE (LAST 5 DAYS) OF COVID-19 CASES
57
31
20
13
8
4
1
AVG: 5.70
Active cases
Per capita
Language diversity
NUMBER OF LANGUAGES SPOKEN
1
840
47 AVG
TWB, OCHA, John Hopkins, TWB, OCHA, John Hopkins, TWB, OCHA, John Hopkins, © CARTO

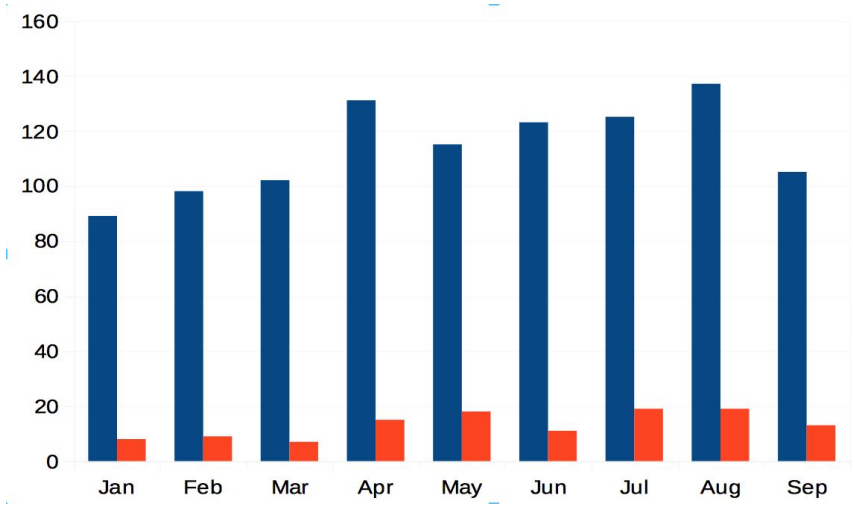
Linguistic crisis response



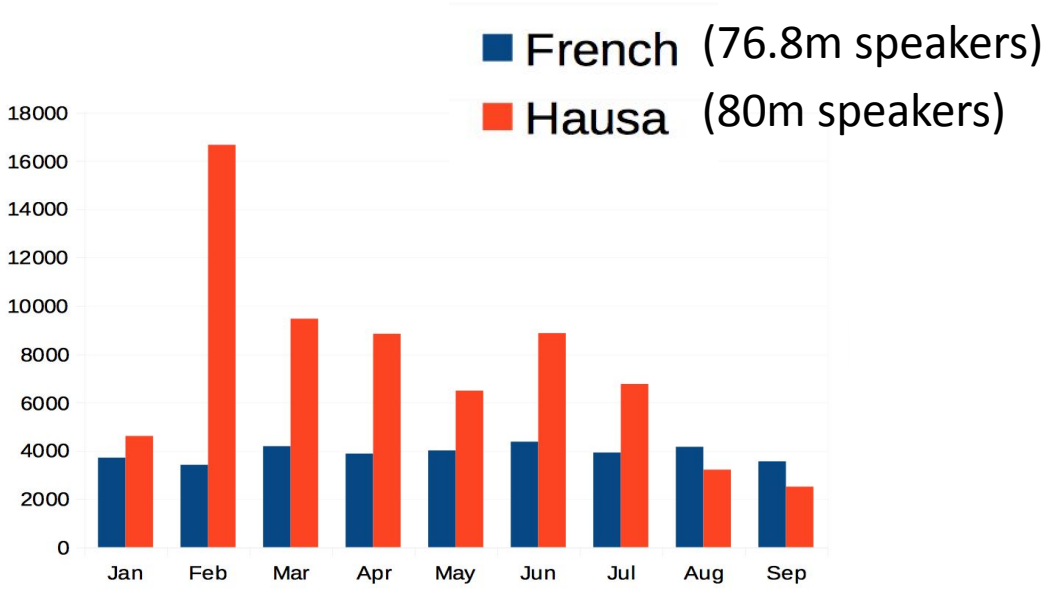
Linguistic crisis response



Hausa vs. French



Volunteer translators



Word count/translator

Data from Kato: TWB's translation platform during Covid-19 pandemic

How can
language technology help
to empower translators of **marginalized languages**?

Language data collection

parallel and audio data

Language data collection

parallel and audio data

MT model development

leveraging low-resource methodologies

Language data collection

parallel and audio data

MT model development

leveraging low-resource methodologies

Machine-assisted translation

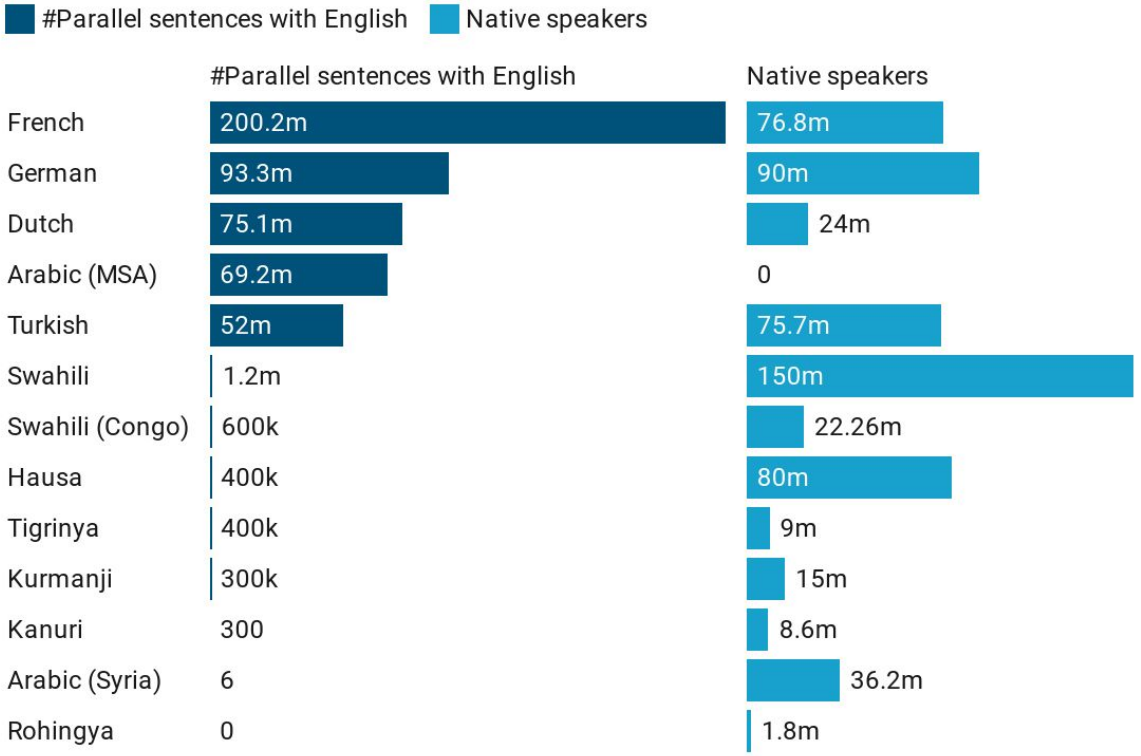
tailored for non-professional translators

Language data collection

NMT for humanitarian impact

Language Data Disparity

Data has been consolidated from the OPUS collection of publicly available parallel corpora paired with English.



Gamayun kits

- Starting point for developing audio and text corpora for languages without pre-existing data resources.
- Four dataset versions:
  - Mini-kit - 5,000 sentences
  - Small-kit - 10,000 sentences
  - Medium-kit - 15,000 sentences
  - Large-kit - 30,000 sentences.
- Source sentences in English, Spanish, French
- Freely available from https://gamayun.translatorswb.org/
  - Currently mini-kits in Hausa, Kanuri, Rohingya, Swahili, Nande

MT model development

MT model development

- Languages: Levantine Arabic, Tigrinya, Congolese Swahili
- Main techniques employed:
  - Domain adaptation
  - Dialect adaptation
  - Cross-lingual transfer learning
  - Back-translation

Domain/dialect adaptation

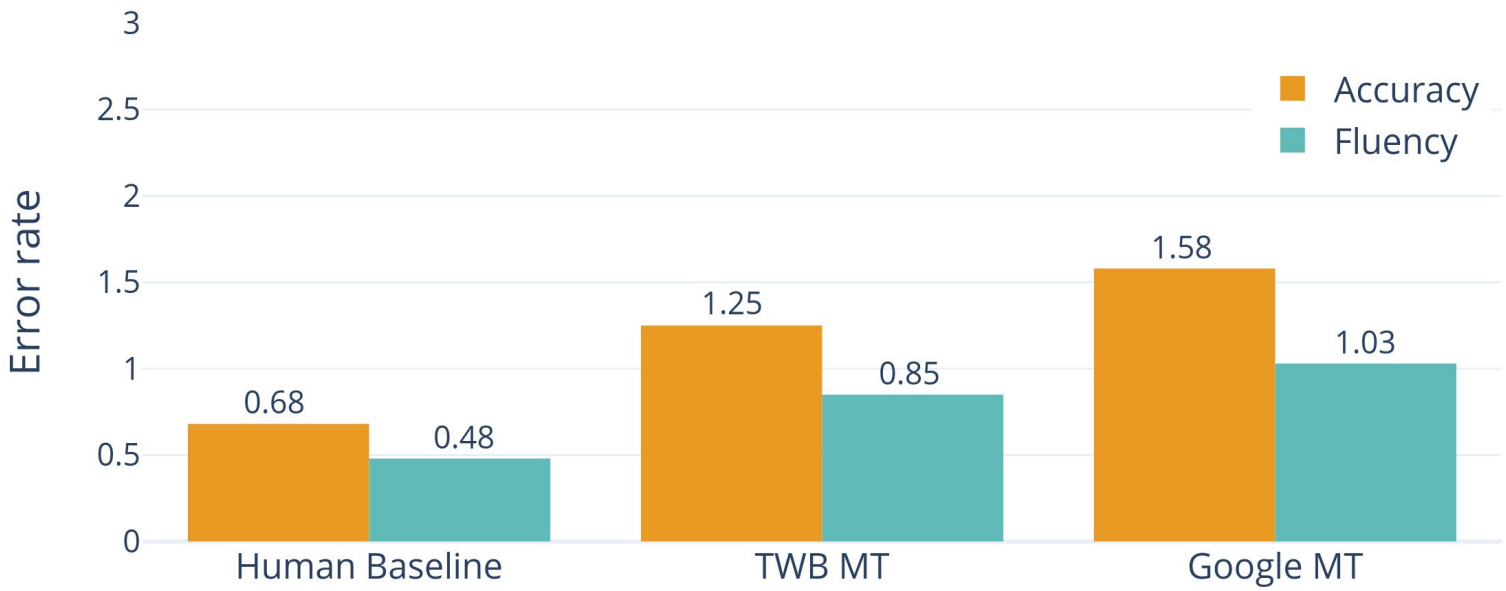
- Levantine Arabic to English machine translation
- For social media content by Syrian refugees in Jordan
- Small in-domain data (5200 sentences)
- Modern Standard Arabic as base model

KHABRONA.INFO خبرونا

WFP World Food Programme

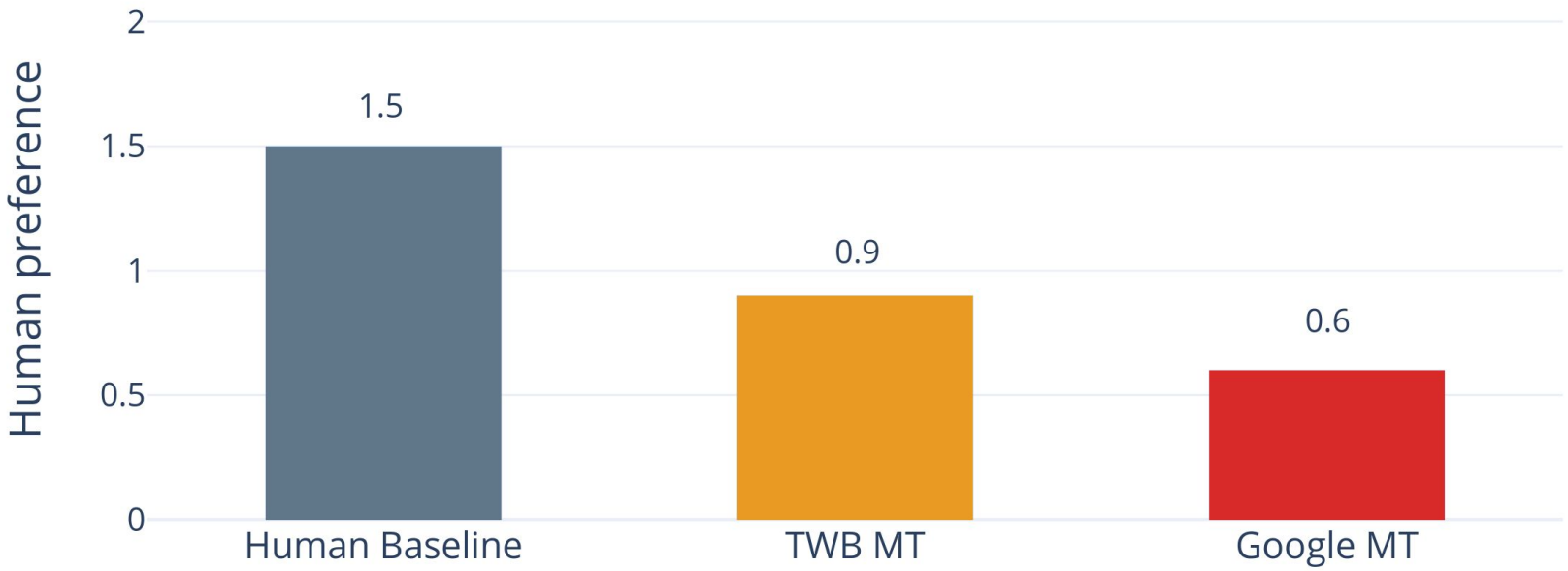
Data | **MT** | Application

Domain/dialect adaptation



Manual evaluation of TWB's Levantine Arabic MT for usability in social media monitoring

Domain/dialect adaptation



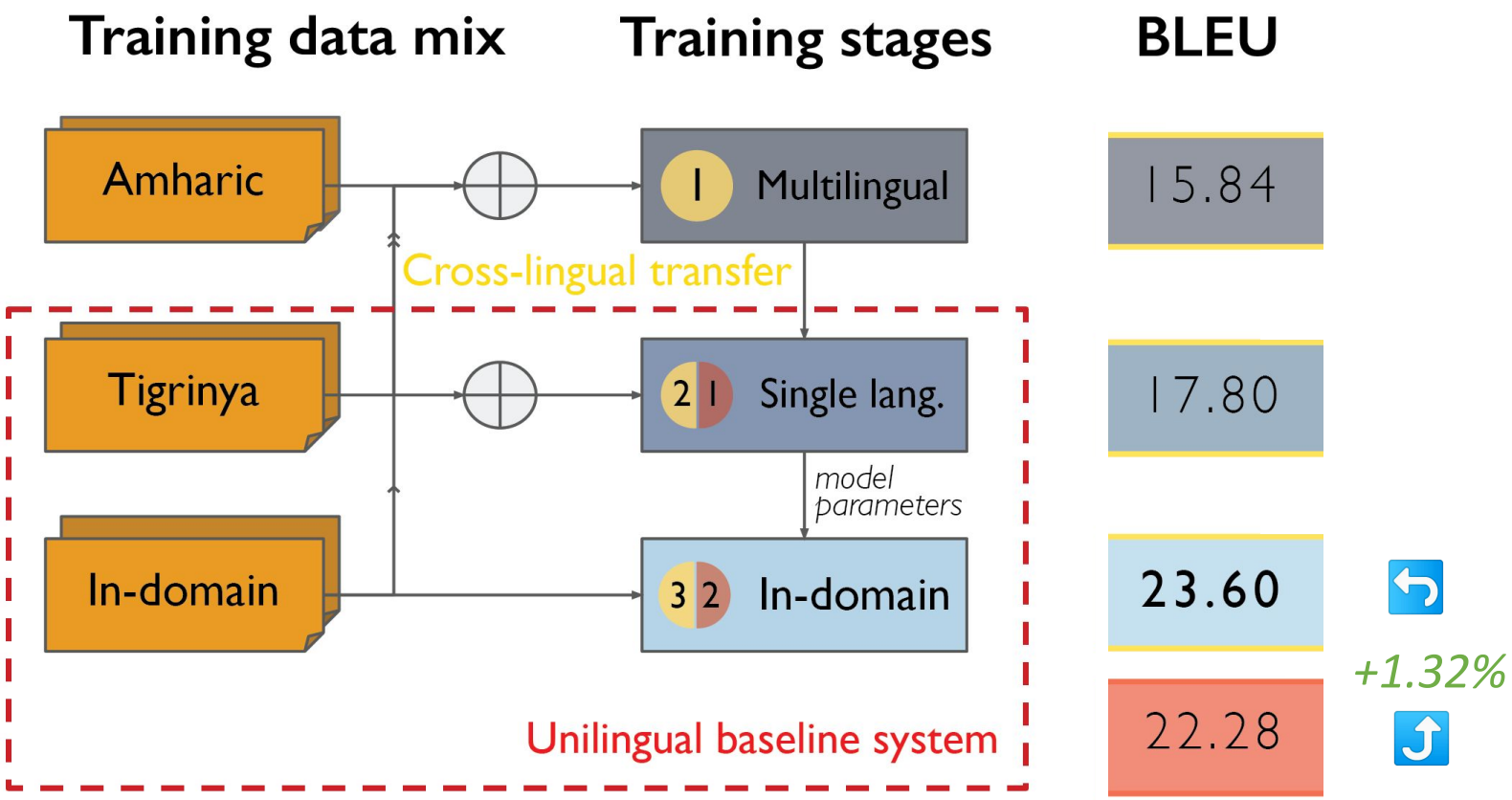
Tigrinya NMT

- Semitic language with estimate # speakers of 7.9 million
- Refugee language in Europe and USA
- Hard-to-resource for translation
  - 3 active translators
  - %81 claimed in 2020
  - 72-day average delay
- Transfer learning from Amharic

ትግርኛ

Data | **MT** | Application

Tigrinya NMT



Cross-lingual transfer learning and domain adaptation

Tigrinya NMT

- Bidirectionality challenge:
  - Tigrinya-to-English: 23.60 BLEU
  - English-to-Tigrinya: 9.92 BLEU
- More details on paper:
  - A. Öktem, M. Plitt, G. Tang. *Tigrinya neural machine translation with transfer learning for humanitarian response.* AfricaNLP Workshop organized within ICLR, Addis Ababa, Ethiopia, April 2020.

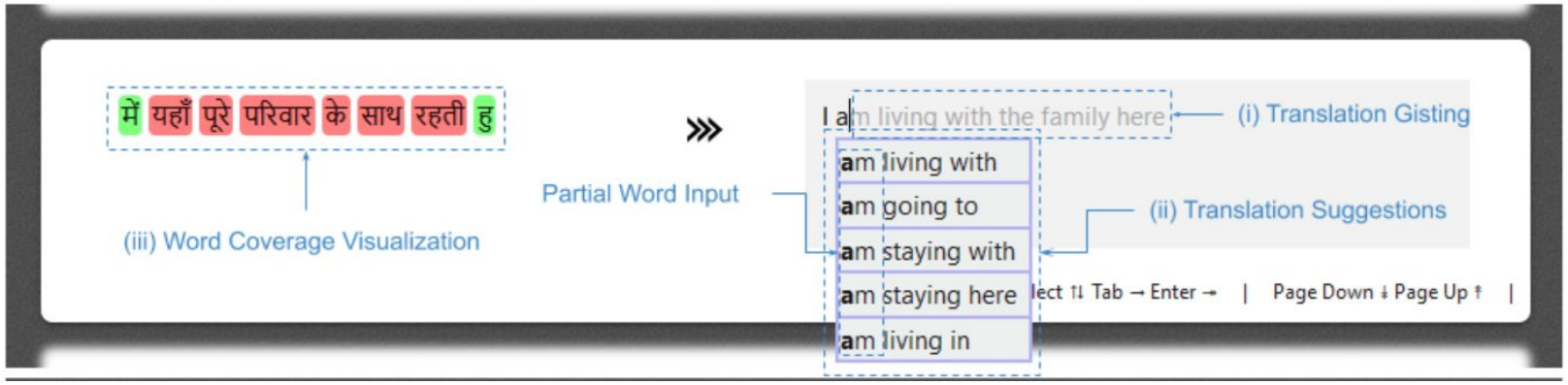
https://gamayun.translatorswb.org/

Data → **MT** → Application

Machine-assisted translation

Interactive Machine Translation

- Proof-of-concept by Microsoft Research India
- Assisted translation through:
  - on-the-fly hints
  - suggestions
- Alternative to post-editing



Santy, Dandapat, Choudhury, Bali. “INMT: Interactive Neural Machine Translation Prediction”. EMNLP 2019

Interactive Machine Translation

- Faster turnaround of document translations
  - compared to manual, and post-edited

Word Coverage and Translation Gisting	Suggestions	Keystrokes
उसी प्रकार मानसिक स्वास्थ्य के लिए ज्ञान की प्राप्ति आवश्यक है Similarly , knowledge for mental health is necessary .	Similarly , In the The knowledge Thus , So the	↓ ↓ Enter ↩
उसी प्रकार मानसिक स्वास्थ्य के लिए ज्ञान की प्राप्ति आवश्यक है **In the** same way , knowledge of knowledge is essential for mental health	same way	Tab Tab Tab Tab
उसी प्रकार मानसिक स्वास्थ्य के लिए ज्ञान की प्राप्ति आवश्यक है **In the same way , knowledge** of knowledge is essential for mental health	of knowledge is essential is necessary for mental	i
उसी प्रकार मानसिक स्वास्थ्य के लिए ज्ञान की प्राप्ति आवश्यक है **In the same way , knowledge i**s essential for mental health	**i**s essential for **i**s necessary for **i**s required to	Enter ↩
उसी प्रकार मानसिक स्वास्थ्य के लिए ज्ञान की प्राप्ति आवश्यक है **In the same way , knowledge is essential for mental health**		Page ↓

Santy, Dandapat, Choudhury, Bali. “INMT: Interactive Neural Machine Translation Prediction”. EMNLP 2019

Interactive Machine Translation

- Faster turnaround of document translations
  - compared to manual, and post-edited
- Human-machine collaboration to best leverage low-resource models

	Data Size	0%	10%	20%	40%
bn-en	1.1M	25.31	27.54	35.68	54.03
hi-en	1.5M	40.64	42.06	47.90	62.18
ml-en	897K	19.76	21.95	29.84	49.88
ta-en	428K	18.71	20.90	27.05	44.55
te-en	104K	11.92	14.57	21.17	41.98

Table 2: Multi-BLEU Score with x% of partial input

Santy, Dandapat, Choudhury, Bali. “INMT: Interactive Neural Machine Translation Prediction”. EMNLP 2019

Interactive Machine Translation

- Faster turnaround of document translations
  - compared to manual, and post-edited
- Human-machine collaboration to best leverage low-resource models
- Boost for hard-to-source languages
  - for translation by non-experts
  - for crowdsourced data collection

Language data collection

parallel and audio data

MT model development

leveraging low-resource methodologies

Machine-assisted translation

tailored for non-professional translators

#LanguageTechnologyMatters

alp@translatorswb.org
https://translatorswithoutborders.org/

TRANSLATORS
WITHOUT BORDERS


TRANSLATORS
WITHOUT BORDERS

NMT for humanitarian impact

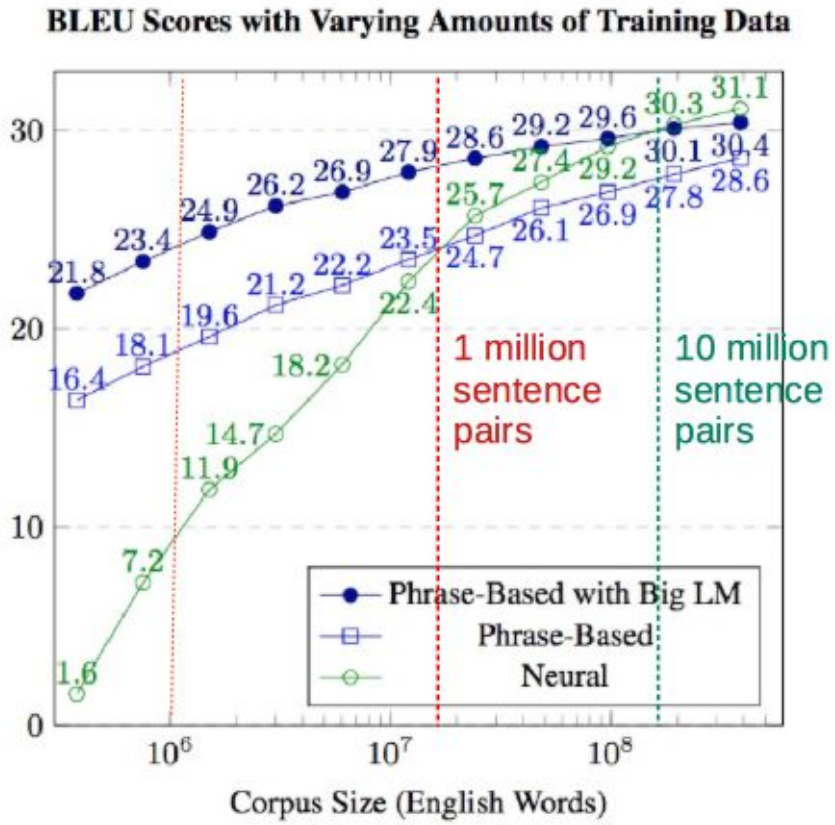


Diagram edited from Koehn and Knowles (2017)

Tigrinya NMT

SMT on 7 Ethiopian languages (Teferra Abate et al., 2018)

Parallel corpus of 300+ languages from *jw.org* (Agić and Vulić, 2019)

Available on OPUS repository (Tiedemann, 2012)

TWB's translation memories

	Ethiopian corpus	JW300	Bible-uedin	Global voices	GNOME	Tanzil	TWB	TOTAL
Amharic	66K	722K	61K	1.6K	57K	94K	-	1M
Ge'ez	11K	-	-	-	-	-	-	11K
Tigrinya	36K	400K	-	-	-	-	2.5K	439K

Dataset sizes (#sentences) for Ge'ez scripted languages

Data → **MT** → Application

Gamayun kits

Language	kit-5k	Audio	Language tech development goals
Hausa	✔	⚙	Machine-assisted data collection
Kanuri	✔	⚙	Machine-assisted data collection
Kurmanji Kurdish		⚙	Machine-assisted survey transcription
Rohingya	✔	✔	Glossary with voice search
Coastal Swahili	✔	✔	MT and audio keyword detection
Congolese Swahili	✔		Interactive neural machine translation
Tigrinya	⚙		Interactive neural machine translation

Interactive Machine Translation

How?

- Constrained decoding on top of *OpenNMT* models
- Latest development: BPE integration
- Work-in-progress: Evaluation with our volunteer translators

Demo

- https://microsoft.github.io/inmt/