CLEAR Global

Language technology for humanitarian action

How collaborating on language technology development could reduce the digital language gap

Up to four billion people globally face digital or language exclusion. Developing language technology for humanitarian and development use represents a significant opportunity to reduce exclusion across the aid sector and beyond:

- **For the many millions of marginalized language speakers globally,** it could help overcome the risks and obstacles of digital services being in inaccessible languages and formats, open opportunities and bridge the digital divide.
- **For humanitarian and other service providers,** it could ease tensions between the competing demands of inclusiveness and cost-effectiveness.

- **For governments and civil society,** it could bring many millions more citizens into civic conversations and decision-making and expand the reach of social and economic initiatives.
- **For the private sector,** including mobile network operators, it could mean providing better services to new and existing customers.

This brief proposes a collective approach to language technology development based on:

- Identifying where language barriers prevent people from accessing a service they want to use.
- Ensuring organizations understand concerns and risks related to a particular language, and hear from marginalized language speakers themselves on those issues.
- Ensuring users' consent is based on an understanding of what language technology is, how it could help them and how their language data will be used.
- Involving language communities in decisions on the applications their language data is used to develop.
- Identifying where language technology isn't an appropriate or desired solution.

This brief is part of wider research done by CLEAR Global and the GSMA's Mobile for Humanitarian Innovation program on the state of inclusion and exclusion for marginalized language speakers in digital humanitarian services. Read the full report here.

What is language technology? A few key concepts

Automated transcription or speech recognition: turns spoken input into text that can be processed electronically, including by machine translation, for instance enabling less literate, visually impaired and second-language users the option of navigating an interface using voice.

Machine translation: translates between a source and target language. Can require 'post-editing' by a human translator to ensure accuracy.

Natural language interface: users can ask questions and provide input in their own words.

Natural language understanding: allows machines to process the content of human speech. Applications include machine translation, responding to questions, and matching the 'intent' of a user's input to relevant information which can be provided in response.

Natural language generation: allows machines to generate understandable natural speech. Part of generative AI.

Speech synthesis or text-to-speech capability: generates understandable spoken language from text, for instance for users who can't read text in a given language.

The problem: language technology isn't available in many languages of crisis-affected people

Only a handful of the estimated 7,000 human languages have a strong online presence. Digital platforms are available in just a few hundred, and language technology of a high enough quality for general communication is still only available for around 50. These are broadly the dominant languages of the more economically and politically powerful nations, **not the first (sometimes only) languages of many people affected by conflict and crisis.**

Even in the 'right' languages, existing technology may not work for crisis-affected communities

Many of those most affected by humanitarian emergencies don't fit the profile most language technology was developed for:

- **Current language models are unlikely to prioritize relevant variants.** Language technology is widely available in Portuguese, for example, but speech recognition is typically geared to European and Brazilian Portuguese, and works less well for Portuguese speakers in countries like Angola and Mozambique.

- **Second-language speakers often mix in words from their first language or other local languages.** Language models trained exclusively on standard variants of these languages will not accurately recognize or reproduce this.
- **Models are still mostly trained using data from people with accents from industrialized countries**, and don't work as well for Nigerian English speakers or second-language French speakers in the Democratic Republic of Congo.
- **Less educated individuals - disproportionately women and girls, older adults, people with disabilities and marginalized minorities -** typically speak and understand less of the dominant or national language and have a smaller vocabulary.
- **People experiencing long-term displacement may have learned to speak the language of their host community, but not comfortably read or write it.** Others who could only go to school in a second language may be more comfortable reading that language, but prefer using audio or video in their first language.
- **Commercial applications of language technology tend to assume end users are literate, own a device, and know how to navigate digital spaces.** In the aid sector, many intended end users will be less familiar with technology, if at all.

Two broad data gaps limit the sector's ability to understand the extent of this issue and take practical action to address it:

1. The pervasive lack of information about the languages used by half the world's population - or 'language use data'.
2. The shortage of voice and text data ('language data') in the relevant languages to build the technology itself.

Both data gaps are costly to fill completely - but partial gains would have significant impact on expanding the reach of digital services and bringing down costs.

A solution: a collective approach to language technology development can reduce these challenges

A recent GSMA-CLEAR Global study points to the potential for collaboration between aid organizations, marginalized communities and technologists to bridge this gap:

- **Marginalized communities generate language data -** the raw material for language technology - when they communicate with aid organizations, and could benefit from digital services in languages they speak and understand.

- **Aid organizations receive and generate language data** through digital communication with affected people; better language technology for marginalized languages could expand their reach and reduce costs.
- **Language technologists have the capability to develop** that language technology but lack the language data and real-world use cases needed to develop workable language solutions for marginalized communities.

Collaboration on language technology for less well-served languages is already happening. A series of platforms and initiatives are expanding access to both text and speech data for those languages and to the tools needed to build language models. These provide a starting point for approaches to language technology development in the aid sector.

Examples of collaborative language technology development

Common Voice: a crowdsourcing project started by Mozilla to collect open-source data that can be used to develop speech recognition software

Hugging Face Hub: a platform hosted by for-profit Hugging Face where machine learning developers can share open-source datasets and language and speech models

Karya: a data cooperative developing training datasets for AI and machine learning and supporting livelihoods through paid collaboration with marginalized language communities in India

Keyman: an open-source platform developed by SIL International to enable the development of keyboard layouts for previously unsupported languages

Lanfrica: a platform that catalogs and connects African language resources to make them more easily discoverable for purposes including the development of language technology

Start by knowing who wants and needs which digital services, in which languages, and refine existing technology on that basis

Focused, evidence-based language technology development for specific populations and use cases could open up access for many millions of first- and second-language speakers. But existing language technology for major contact languages (lingua francas) like Swahili, Hausa or Spanish

doesn't take account of how people receiving aid are likely to use those languages; as a result, it doesn't reach many millions of second-language speakers.

Humanitarians also often overestimate how many people can confidently use a second language, or for what purposes. For example, an intended user might be able to use a second language for basic functions like checking a mobile money balance, but not for more complex services like using a telemedicine chatbot or reporting a complaint about aid. And digital platforms typically only collect data on users, not on those who can't use them in the languages provided, so organizations don't know how many of their target users are excluded through language.

Improving existing language technology in contact languages can support inclusion efforts for several communities at the same time, potentially across several crises, where first-language technology development isn't yet feasible. That better-quality general language technology can then be the starting point for targeted humanitarian applications. Technologists can mine data on local second-language comprehension and terminology use on a given topic to develop functioning language technology applications for specific applications and target populations. For instance they could develop applications for sexual and reproductive health information for women and girls in North Kivu, DRC (see box below) without having to start from scratch.

Information on affected people's languages and communication preferences would help organizations allocate resources to languages that would have the greatest impact. But this data is not yet routinely collected or shared, and communication needs assessments don't routinely happen in marginalized languages and so don't necessarily capture risks of language exclusion. Addressing this gap is necessary for building effective language technology for affected people - and significant data gains don't have to be costly or time-consuming.

The potential of use-case focused language data collection

To ensure a voice-activated digital service works for both first- and second-language Swahili speakers in eastern DRC (more than 22 million people in total), the speech model would need to be trained with data that reflects language use in the relevant communities:

- Include voice recordings of natural speech by community members in localized Congolese Swahili on topics related to the service.
- Include speakers of different ages and genders to minimize bias in whose speech is recognized, and identify where different groups use different terms.
- Vary the audio quality and background noise to reflect real-life conditions.

The greater the volume of this kind of data, the better the resulting tool will correctly identify the meaning of a range of callers in different situations.

Maximize the potential of language data that aid organizations already collect

Humanitarian digital services - and the wider social impact sector - already generate significant amounts of language data. Using this could significantly reduce the cost and time needed to generate language data in order to train new language models, especially in languages less served by the commercial sector. A humanitarian communication ecosystem in which every non-sensitive voice message, transcription or translation was safely used - with users' consent - to improve the accuracy and speed of multilingual communication for all involved would have far-reaching benefits for affected people.

But to work on the scale needed to deliver real change, it's not individual service providers feeding their data to technology developers that is required, but an ecosystem in which hundreds, thousands of service providers pool their data. And all those organizations, big and small, would need to be making use of the resulting applications, as the technology improves through use.

Building this ecosystem requires new incentives for collective action, requiring donors and organizations to approach ideas of scale differently. For instance, it implies looking at value for money over longer time frames and beyond specific emergencies, and considering impact for

marginalized groups beyond the numerical reach of a service. Ultimately it will require coordinated effort by donors, organizations and technologists to put the foundations in place, including establishing standards and tools for informed consent, data compatibility, data storage, and access for all.

Collective action also helps develop the technology faster. For example, pooling language data for a translation memory in a given language would improve translation accuracy and reduce the time needed for human moderation. Human capacity could be redirected to more complex translation needs. The more a language model is trained, the more effective it gets and the cheaper it becomes per hour of data to build and refine for new use cases. Parallel increases in connectivity and device access are making it easier to generate the language data needed more cheaply, and in ways that more actively involve marginalized language speakers themselves.

A language ecosystem for social impact will take time to establish fully. But **many of the pieces are already in place to start creating context- and use-case specific technologies in strategic languages of affected communities in the short term.**

Address concerns around informed consent and language exploitation

Language data collection can reproduce exploitative dynamics. Some language communities reject the development of language technology as turning their language and culture into a commodity that can be 'sold' back to them as commercial digital services. In contexts where language intersects with conflict or persecution, collecting people's voice and text data carries risks. It's also hard to be sure someone's consent is informed when it may be perceived as a condition of aid, and when the future uses of data by artificial intelligence (AI) are unknown. These issues need to be addressed openly, possibly drawing on the experiences of Karya's model of individual ownership of language data, and on international standards for obtaining the consent of Indigenous communities.

Language barriers exacerbate these challenges when people are not consulted in their first language and where the concepts involved in discussions of data protection have no local equivalents or are not in mainstream use. People may not have enough information in their language to know what their language data would be used for, who would get to use it, or what the implications are for them. For instance, while personally identifiable information can be removed from text, voice data can identify an individual speaker. Since voice data is needed to build speech

recognition and other non-text communication tools, the data protection challenge particularly affects technology development for less literate individuals and speakers of oral languages - who are already likely to be more vulnerable.

Given this background, resistance from marginalized language communities is valid and understandable. **But digitalization of the aid sector isn't slowing down.** To do no harm and reach those most left behind, the aid sector needs to develop governance of its technology use on the understanding that language matters for people's safety, well-being and identity.

Support smaller organizations to integrate language technology

Consultation with nonprofit organizations in low- and middle-income countries suggests many that see a benefit in language technology lack the in-house expertise to incorporate it. And even organizations already using language technology would need support to feed into language technology development. This includes:

- Resources and expertise for removing sensitive and other personally identifiable information from voice and text data and formatting data for upload to a data sharing hub.
- Promoting awareness of the current and potential capacities of language technology and the mechanics of using it.

Existing resources offer a starting point, and common formats and workflows would reduce the demands on individual organizations. Capacity for the remaining in-house tasks could become a standard feature of program budgets.

Harnessing the potential of language technology with and for marginalized language speakers is no small endeavor. But the benefits for crisis-affected people, and the organizations supporting them, would be profound. This is a conversation worth having.

CLEAR Global can help. We promote and support the development of open-access language technology for social impact in marginalized languages by:

- Developing processes, frameworks and tools to help social impact organizations safely create and/or curate language data and make it available on open-access platforms
- Mobilizing local language communities to collect language data in under-served resource languages, focusing on voice, building on our established processes and our core community of over 100,000 linguists working in hundreds of languages
- Working with social impact organizations, language technology developers and interested donors to develop the conditions for a more collaborative and sustainable user-centered development of language technology
- Where other options are unavailable, building sustainable tools for marginalized languages

UK International Development
Partnership | Progress | Prosperity

This research was supported by the GSMA and funded by UK International Development from the UK government. The views expressed do not necessarily reflect those of the GSMA, nor do they reflect the UK government's official policies.

GSMA™

About the GSMA

The GSMA is a global organization unifying the mobile ecosystem to discover, develop and deliver innovation foundational to positive business environments and societal change. Our vision is to unlock the full power of connectivity so that people, industry, and society thrive. Representing mobile operators and organizations across the mobile ecosystem and adjacent industries, the GSMA delivers for its members across three broad pillars: Connectivity for Good, Industry Services and Solutions, and Outreach. This activity includes advancing policy, tackling today's biggest societal challenges, underpinning the technology and interoperability that make mobile work, and providing the world's largest platform to convene the mobile ecosystem at the MWC and M360 series of events.

We invite you to find out more at gsma.com