# How human input can help NLP research meet real-world needs in emergencies: a practical view of Rohingya

**Ellie Kemp** **Alice Castillejo** **Alp Öktem** **Milena Haykowska**

**CLEAR Global**
Garden City, Idaho, USA
<first-name>.<surname>@clearglobal.org

## Abstract

Voice technology could enable governments and humanitarian responders to reach, and hear from, people currently excluded by language or literacy — among the most vulnerable in any emergency. Whether it does so depends on human input on questions no automated benchmark can answer: whether a system targets the right language and variety, whether it works for women as well as men, which script it should transcribe into, and whether its output is good enough for human communication rather than merely scoring well against automated metrics. Drawing on CLEAR Global's work with community linguists in crisis settings, and using Rohingya as an extended example, we show what is lost when assumption and generalization displace local knowledge. We then set out five research questions on the integration of human input into voice technology development: where a community's limited time is best spent, who decides which variety a system serves, where an orthographic standard comes from when none exists, what counts as a correct transcription when spellings legitimately differ, and who decides which errors matter most. Each question is open, and each needs the participation of the communities the technology is meant to serve.

## 1 Introduction

For low-resource languages, human understanding is a critical resource for NLP development, especially in crisis preparedness and response. NLP research has a role to play in helping to integrate that understanding into voice technology.

Language technologies have great potential to support crisis-affected speakers of low-resource languages. Voice technology especially could enable governments and responders to reach and hear from many of the most vulnerable in emergencies who are currently excluded by language or literacy.

Yet research in this area often leaves the crucial question unanswered: how well does the technology work for people at risk from or directly affected by conflict, disaster and disease? Only humans can assess the fit between voice technology development and the real world.

CLEAR Global's research and experience with thousands of community linguists working on language exclusion in crisis preparedness and response underscores just how much is lost when assumption and generalization obscure local human understanding. The growth of AI risks amplifying that loss, so that language technology worsens the exclusion of marginalized language speakers instead of overcoming it.

The solution is to make better and more consistent use of the knowledge of the language communities it aims to assist. NLP research can support that by showing where human input is most needed and how best to maximize its benefit.

## 2 Mind the gap: the human knowledge that voice technology needs

The evidence to date indicates knowledge gaps that can prevent the outputs of NLP research from benefiting at-risk and crisis-affected speakers of low-resource languages.

Key questions which can only be answered with human insight include:

- Is this the right language and variant?
- What is needed for voice technology to work for women as well as men?
- How best to address the use of multiple scripts?
- How can the technology be deployed in ways that are relevant and accessible for excluded groups including older adults, people with disabilities, women and rural residents?

And the fundamental question:

- Is the output good enough for human communication, as distinct from scoring well against automated metrics?

## 3 The example of Rohingya

To illustrate these gaps, let's consider each question in relation to the Rohingya language, spoken by between 1.8 and 2.8 million people, most of them currently in Myanmar and Bangladesh. This is just one example among many: similar issues apply to Saraiki with its 40 million speakers in Pakistan and Congolese Swahili with its 35 million in the Democratic Republic of Congo.

Language and communication barriers help keep many Rohingya speakers isolated, insecure and without adequate access to either services or life and livelihood choices. Longstanding denial of mother-tongue education has resulted in generally low literacy levels and patchy knowledge of other languages. Women and girls are especially affected and also face additional barriers: gendered exclusion from public spaces leaves them less able to access and share information outside the home. In this context, voice technology could offer valuable solutions – improving the reach of emergency warnings, health information and training opportunities and helping people to make their voices heard.

### 3.1 Is this the right language and variant?

Getting the language right may seem self-evident, but most humanitarians supporting Rohingya refugees in Cox's Bazar in 2019 thought the refugees spoke Burmese (Myanmar) and that the local Chittagonian language was close enough to Rohingya for effective communication (TWB, 2019). In fact only 12.5 percent of women spoke Burmese. More than one-third of refugees surveyed couldn't understand simple messages in spoken Chittagonian. And the Rohingya and Chittagonian words for concepts like 'danger', 'rescue' and 'safe' are sufficiently different that emergency warnings broadcast only in Chittagonian could be fatally misunderstood (TWB, 2018).

Getting the variant right is also critical – especially in a context where forced displacement has increased variation through contact with other languages, cultures and structures. And where a longstanding lack of formal language development has prevented the emergence of standardized Rohingya terminology around topics like medicine, law, technology, education, and administration (TWB, 2021). ASR applications need to recognize the speech of likely users in the given context, and TTS applications need the target audience to understand them if they are to build trust and adoption.

Within Myanmar, there is sufficient variation between the Rohingya spoken in urban and rural areas that speakers from different regions can struggle to understand one another. The language of refugees in Bangladesh meanwhile depends partly on their education: Arabic and Urdu loanwords for those educated in *madrasas*, Burmese and English for those schooled back in Myanmar. It also varies with their migration journey: longstanding residents of Cox's Bazar borrow from Bangla or Chittagonian, newcomers from Burmese or Rakhine, and those who have spent time in Malaysia from Malay. This linguistic evolution was visible in the camps as early as 2018, with newly arrived and established refugees using different words to say the same thing (TWB, 2018a).

Given this variation, human insight is needed to determine the vocabulary, syntax and pronunciation that are most widely used and understood. NLP research can then focus on either the most relevant variant for a given use case, or a representation of Rohingya that recognizes and is understood by speakers of a range of variants.

### 3.2 What is needed for the technology to work for women as well as men?

Technology that is not designed by or for women is unlikely to work well for them. If NLP is to avoid that pitfall, it needs input from women. That includes factoring in both where systems of knowledge and vocabulary exclude women, and where women's own knowledge and vocabulary differ from men's. Both these issues are relevant for Rohingya speakers.

Traditional gender norms restrict girls' access to schooling and women's access to public spaces and discussions. As a result, many Rohingya women are unable to read or write in any language (TWB, 2021a) and may lack the vocabulary and frame of reference to discuss the same issues as their male relatives.

As in many societies, Rohingya women use a different vocabulary from men for concepts viewed as being within a woman's sphere, such as child care, or too sensitive to be discussed with men, such as sexual and reproductive health. So while

men recognize the formal word *haiz* – a borrowing from Arabic – as meaning menstruation, Rohingya women typically prefer the euphemism *gusol*, meaning 'shower'. Women might refer to their breasts as their 'chest' to avoid embarrassment; a male healthcare professional or interpreter may misunderstand (TWB, 2019). Rohingya women and girls also use euphemisms like 'dishonor' and 'oppression' for the sensitive topic of sexual abuse, which require human understanding of culture and nuance to decode (TWB, 2021b). Any NLP application relating to sexual and reproductive health would need to identify those terms correctly when they are used, and use similarly indirect language in response.

### 3.3 How best to address the use of multiple scripts?

Rohingya does not have a single standardized writing system; instead, several scripts are in use. These include the Arabic-based Fonna script, a Latin-based system commonly referred to as 'Rohingyalish', and Hanifi, a script developed specifically for the language in the 20th century. In the refugee camps in Cox's Bazar, learning centers teach children to read and write Rohingya using both Hanifi and Rohingyalish, and humanitarian organizations tend to use whichever script their staff is most familiar with.

Low literacy levels make voice-based interfaces more relevant than text-based interaction, but the choice of script for transcription must be matched to the intended users. An ASR model trained on voice data transcribed using Rohingyalish will produce automated transcriptions in that script; it will not automatically support an organization whose staff uses the Hanifi script.

One solution is to provide training data transcribed in both scripts so that ASR models can support different user preferences. This is technically feasible, but requires additional annotation effort and increases the cost of developing the training data and model. Another solution may be to convert transcription automatically from one script into another. Research with language informants is needed to determine whether this can reliably be done between Hanifi and Rohingyalish.

### 3.4 How can the technology be deployed in ways that are relevant and accessible for those who are most excluded?

In Rohingya society as in most human societies, some groups face disproportionately greater structural barriers to accessing information and making themselves heard. These include older adults, people with disabilities, women, children and rural residents. Targeted development of voice technology can help overcome two of these barriers: language and literacy. But it can only benefit those groups if its deployment addresses the other blocks they face, including social, cultural and economic barriers to phone ownership and internet access.

More research with communities themselves is needed to understand how voice-enabled tools could be designed or deployed to reach those affected by such barriers. Applications that work offline or using 2G, low-data or automated callback systems rather than WhatsApp or dedicated apps, for instance, could be game-changing for a range of excluded Rohingya speakers. So too could deployment methods that women and girls can safely access from their own homes.

### 3.5 Is the output good enough for human communication?

Voice technology is not yet advanced enough to provide an illustration for Rohingya on this point. But this is surely the critical metric for NLP research in crisis preparedness and response, and only contextual human understanding can measure against it.

The 2024 Pariksha study (Watts et al., 2024) highlighted the limitations of AI benchmarking for 10 Indic languages. Human evaluations of 30 LLMs used culturally grounded benchmarks developed by members of the relevant language communities. They found that AI benchmarks were of moderate to low value for evaluating culturally nuanced tasks, even in relatively well resourced languages like Hindi.

"Good enough" may of course vary depending on the user and the use case. TTS that works for madrasa-educated men may very well not make sense to women in the camps of Bangladesh. An STT application may perform poorly as part of automating real-time responses to refugees' questions but be useful for speeding up transcription with a human in the loop. Again, research with the intended users themselves is needed to gauge this.

## 4 Research can drive impactful integration of human understanding

When the needs are this vast and the budgets typically small, the cost, ethics and practicalities of working with language community members on NLP research can be challenging. This is where NLP research can gain by exploring where and how the input of linguists and non-linguists from each community can be optimized.

### 4.1 Transcription guidance as a potential tool of optimization

To take an example from corpus creation, a foundational challenge for low-resource languages: human transcription can easily result in partially unusable data. A non-standardized language like Rohingya presents obvious risks of variation in spelling and diacritics, in addition to dialectal variations and individual habits of punctuation and capitalization. As part of building speech datasets for 10 low-resource languages in collaboration with George Mason University, CLEAR Global worked with linguists from the 10 communities to develop contextually informed language-specific transcription guidance. The guidance for each language specifies the script to use, standard treatment of dialectal variations and foreign words, and how to deal with non-speech sounds, repetition, hesitation and filler words.

A resource like this, replicated at scale, has significant potential to reduce wastage and improve speed and efficiency in low-resource voice technology development. By measuring the actual impact, NLP research could make a valuable contribution by determining whether this is an initiative worth scaling.

### 4.2 Potential research questions on how to optimize human input

Two categories of question flow from this. Some things only members of a language community can decide; here the research question is how those decisions are made and by whom. In other cases the questions are about allocation: community time is scarce, and we lack evidence on how best to spend it.

**Where is a community's time best spent?** Recording, transcribing, validating, developing terminology and generating synthetic data all draw on the same limited pool of hours. We have some information on costs: human-corrected transcription for Bambara has been measured at around 30 hours of labor per hour of audio (Diarra et al., 2025), but the comparison with the downstream costs of not correcting the transcript is missing. Synthetic speech can expand the data available (DeRenzi et al., 2025), but costs time and compute of its own and depends on a generative model being able to produce the language at all – excluding many low-resource languages. The cost-benefit analysis of each additional layer of validation for system quality remains to be determined.

**Who decides which variety the technology serves?** Collecting balanced data across varieties is not the solution: dialect-balanced training data for Irish did not equalize accuracy across dialects (Lonergan et al., 2023), and on a large multilingual corpus, increasing the number of distinct speakers did more for fairness than demographic balancing (Maison and Estève, 2023). The variation produced by displacement is almost absent from this literature: no speech technology research we can find treats displacement and new language contacts as something to be modeled.

**Where does the standard come from, when there isn't one?** For e.g. Kanuri or Rohingya there is no authoritative digital reference for how a word is written, so every project readjudicates the same words and the inconsistency surfaces later as apparent model error. What is the cheapest process that produces a usable reference, and how much of a language's measured error rate is attributable to reference instability rather than to the technology? This is necessarily language-specific, and reduces the cost of everything downstream.

**What counts as a correct transcription when spellings legitimately differ?** How accuracy is measured determines which system scores better. Metrics that tolerate variation need a set of acceptable alternatives, and only speakers can supply it. Who produces it, and at what cost, is unexamined – and existing approaches assume one script, where a language like Rohingya has three. Transcription guidance of the kind outlined above addresses the same problem upstream and is equally untested: we could find no study that varies annotation guidance for a low-resource language and measures the effect on the resulting system.

**Who decides which errors matter most?** Common metrics give the same weight to a mistranscribed greeting and a mistranscribed dosage. In

other fields the impact of this has been assessed: translation quality frameworks give errors carrying health, safety or legal consequences far more weight than minor ones (Specia et al., 2021); risk weights elicited from air traffic controllers exposed systems that scored well conventionally (Chang et al., 2026); machine-translated emergency discharge instructions were accurate at sentence level in over 90 percent of cases yet a proportion carried potential for clinical harm (Khoong et al., 2019). No comparable analysis exists for crisis communication, nor a protocol for eliciting such weights from the communities who bear the risks.

These questions need access to speakers, transcribers and crisis expertise that few research groups can assemble alone. Bringing together a community of linguists across the globe working towards a language-inclusive digital world, CLEAR Global can. We would welcome collaboration on any of the topics outlined.

## References


Chang et al. 2026. Safety-oriented evaluation of language understanding systems for air traffic control. *Preprint*, arXiv:2605.11769.

B. DeRenzi, A. Dixon, M. A. Farhi, and C. Resch. 2025. Synthetic voice data for automatic speech recognition in African languages. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 152–186.

M. Diarra et al. 2025. Cost analysis of human-corrected transcription for predominately oral languages. *Preprint*, arXiv:2510.12781.

Elaine C. Khoong, Eric Steinbrook, Cheyenne Brown, and Alicia Fernandez. 2019. Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Internal Medicine*, 179(4).

L. Lonergan et al. 2023. Towards dialect-inclusive recognition in a low-resource language: Are balanced corpora the answer? In *Interspeech 2023*.

L. Maison and Y. Estève. 2023. Some voices are too common: Building fair speech recognition systems using the Common Voice dataset. In *Interspeech 2023*.

L. Specia et al. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*.

TWB. 2018. The language lesson: What we've learned about communicating with Rohingya refugees. Technical report, Translators without Borders.

TWB. 2018a. *Hefazot* transforms to *nirapotta*; *janela* becomes *kirkiri*. Technical report, Translators without Borders.

TWB. 2019. Misunderstanding + misinformation = mistrust: How language barriers reduce access to humanitarian services, reduce the quality of those services and aggravate social exclusion for Rohingya communities. Technical report, Translators without Borders.

TWB. 2021. A healthy discussion: Rohingya language guidance for healthcare workers. Technical report, Translators without Borders.

TWB. 2021a. Language characteristics and perspectives of long-term Rohingya refugees. Technical report, Translators without Borders.

TWB. 2021b. Discussing gender in the Rohingya language: Building a better dialogue around gender issues. Technical report, Translators without Borders.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.