Improving humanitarian needs assessments through natural language processing

T. Kreutzer
P. Vinck
P. N. Pham
A. An
L. Appel
E. DeLuca
G. Tang
M. Alzghool
K. Hachhethu
B. Morris
S. L. Walton-Ellery
J. Crowley
J. Orbinski

An effective response to humanitarian crises relies on detailed information about the needs of the affected population. Current assessment approaches often require interviewers to convert complex, open-ended responses into simplified quantitative data. More nuanced insights require the use of qualitative methods, but proper transcription and manual coding are hard to conduct rapidly and at scale during a crisis. Natural language processing (NLP), a type of artificial intelligence, may provide potentially important new opportunities to capture qualitative data from voice responses and analyze it for relevant content to better inform more effective and rapid humanitarian assistance operational decisions. This article provides an overview of how NLP can be used to transcribe, translate, and analyze large sets of qualitative responses with a view to improving the quality and effectiveness of humanitarian assistance. We describe the practical and ethical challenges of building on the diffusion of digital data collection platforms and introducing this new technology to the humanitarian context. Finally, we provide an overview of the principles that should be used to anticipate and mitigate risks.

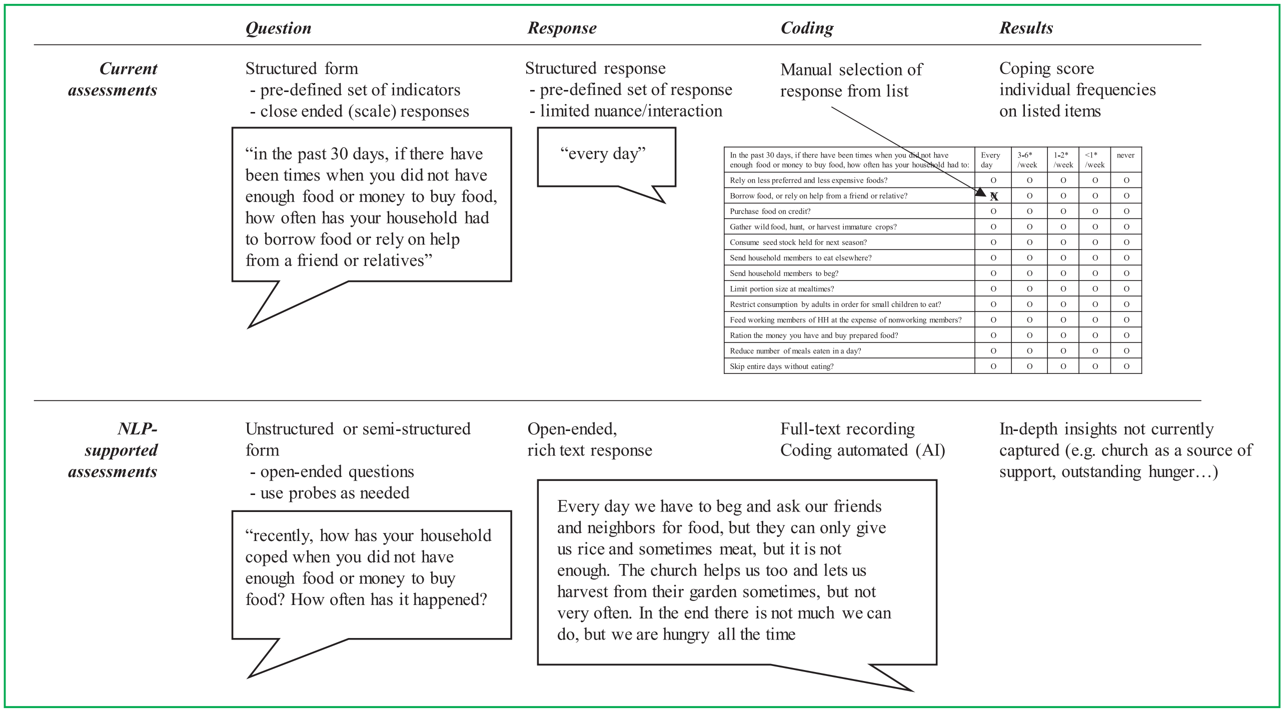
Introduction

The urgent humanitarian needs of people affected by conflict, natural disasters, and climate change have increased significantly in recent years. Today, 69 million people have been forced from their homes, and more than 200 million people need some form of humanitarian assistance [1, p. 81], [2, p. 2]. Affected individuals and communities are an indispensable source of information about needs, preferences, and existing resources for effective disaster response. Humanitarian organizations predominantly conduct successive quantitative interviews with affected people, both to understand initial needs as well as to monitor, improve, and evaluate the response throughout the program cycle. Today, this is accomplished largely through face-to-face surveys using mobile data collection applications, while a small but growing number of organizations also use computer-assisted telephone interviews (CATI). Like paper-based methods, these tools require interviewers or respondents themselves to convert complex responses into categorical or numeric data that can then be objectively analyzed with relative ease and speed.

More nuanced insights into the lives of affected people continue to require the use of qualitative survey methods as well as manual coding and analysis that are time-consuming and hard to conduct at scale during any fixed period in an unfolding humanitarian crisis.[1] At the same time, responses to qualitative questions often contain important contextual information not captured in a quantitative survey. For example, food insecurity is common among people impacted by disasters and complex emergencies. Understanding how households cope in such situations is a key objective of many humanitarian needs assessments (HNA). However, an open-ended question such as “How do you cope with the lack of food?” often cannot be asked because of the challenges in objectively capturing the response in a manner that can be analyzed to inform operational decision-making. Rather, multiple questions with scale-like response keys are used, which can include a

[1] Humanitarian crisis is understood here as “an event or series of events representing a critical threat to the health, safety, security, or wellbeing of a community, usually over a wide area” [3].

Digital Object Identifier: 10.1147/JRD.2019.2947014


Question
Response
Coding
Results
Current assessments
Structured form
- pre-defined set of indicators
- close ended (scale) responses
Structured response
- pre-defined set of response
- limited nuance/interaction
Manual selection of response from list
Coping score individual frequencies on listed items
"in the past 30 days, if there have been times when you did not have enough food or money to buy food, how often has your household had to borrow food or rely on help from a friend or relatives"
"every day"
NLP-supported assessments
Unstructured or semi-structured form
- open-ended questions
- use probes as needed
Open-ended, rich text response
Full-text recording Coding automated (AI)
In-depth insights not currently captured (e.g. church as a source of support, outstanding hunger…)
"recently, how has your household coped when you did not have enough food or money to buy food? How often has it happened?
Every day we have to beg and ask our friends and neighbors for food, but they can only give us rice and sometimes meat, but it is not enough. The church helps us too and lets us harvest from their garden sometimes, but not very often. In the end there is not much we can do, but we are hungry all the time

In the past 30 days, if there have been times when you did not have enough food or money to buy food, how often has your household had to:	Every day	3-6* /week	1-2* /week	<1* /week	never
Rely on less preferred and less expensive foods?	O	O	O	O	O
Borrow food, or rely on help from a friend or relative?	X	O	O	O	O
Purchase food on credit?	O	O	O	O	O
Gather wild food, hunt, or harvest immature crops?	O	O	O	O	O
Consume seed stock held for next season?	O	O	O	O	O
Send household members to eat elsewhere?	O	O	O	O	O
Send household members to beg?	O	O	O	O	O
Limit portion size at mealtimes?	O	O	O	O	O
Restrict consumption by adults in order for small children to eat?	O	O	O	O	O
Feed working members of HH at the expense of nonworking members?	O	O	O	O	O
Ration the money you have and buy prepared food?	O	O	O	O	O
Reduce number of meals eaten in a day?	O	O	O	O	O
Skip entire days without eating?	O	O	O	O	O

Figure 1

Schematic comparison between current interviewing methodology and NLP-supported assessments.

common coping index measure. These, however, may miss important local coping mechanisms and limit respondents' input, as illustrated in **Figure 1**. Here, qualitative information was difficult to process, requiring labor-intensive and time-consuming manual transcription, translation, and coding. Natural language processing (NLP), a type of artificial intelligence (AI), can provide potentially far-reaching new opportunities to rapidly analyze voice responses for relevant content to inform humanitarian assistance decisions. NLP capabilities are becoming widely available through commercial applications released by companies such as Amazon, Microsoft, Google, and IBM, as well as open-source alternatives. However, these AI tools remain largely unavailable in many humanitarian emergency settings where local languages have not (yet) been incorporated by NLP technologies.

Despite rapid technical progress and growing interest, NLP capabilities have not yet been used in humanitarian settings to understand population needs [4]. In this article, we describe a feasibility strategy for use of NLP in humanitarian crises. Our vision is to develop a new way of engaging with affected populations while at the same time providing humanitarian responders with augmented information so that they can respond more effectively and efficiently. As organizations become able to systematically transcribe, translate, and analyze their dialogue with individuals in affected communities—as opposed to merely extracting responses—we suggest that humanitarian assistance may become more effective and efficient, while potentially increasing trust from affected communities. This will require identifying the best NLP methods to analyze different kinds of responses, and—for many emergency settings—creating new transcription models for languages not yet amenable to NLP tools.

The number of people in need of humanitarian assistance will likely grow significantly as the effects of climate change intensify and infectious disease outbreaks become harder to control in the growing number of precarious urban settlements [5]. The innovation we describe here may be highly scalable, save a considerable amount of resources, and produce actionable data in close to real time. Already, funding for humanitarian assistance only covers 56% of estimated requirements [6, p. 8], highlighting the need for innovative strategies and tools to provide rapid, more accurate evidence on how to best use these limited resources.

Employing this new approach is not without ethical challenges. Without upfront and ongoing identification of the socio-political complexity that often leads to or accompanies humanitarian emergencies, and without recognizing the limits and potential biases of NLP techniques, humanitarians may exacerbate context biases that make a particular group vulnerable in the first instance, replicate NLP biases, or expose populations to new risks (especially in the domain of security), all with potentially severe consequences for individuals and population groups. Engineers and humanitarian innovators planning to use NLP tools in humanitarian assistance should understand this potentially complex ethical playing field and anticipate and evaluate the potentially harmful consequences that new technologies might bring with them.

In this article, we will first describe the limitations of current approaches and technologies being used for primary data collection in humanitarian assistance. Second, we outline our methods and how expected results from using NLP could improve qualitative information in various types of humanitarian data collection. Finally, we examine the need for and implications of the growing array of humanitarian, ethical principles, and standards and their implications for the development and use of NLP.

Understanding people affected by crisis quantitatively

Humanitarian assistance[2] refers to coordinated actions that save lives, alleviate suffering, and maintain human dignity during and after human-made crises and disasters caused by natural hazards, and is guided by the fundamental principles of humanity, impartiality, neutrality, and independence [8]. Decisions on necessary assistance and the best methods for delivering it are increasingly driven by a wide array of primary and secondary data to address the needs of the affected population. HNA and other humanitarian surveys are typically conducted quantitatively during the early stages of a crisis to inform organizations and donors of specific gaps and long-term needs, and during the crisis to monitor and assess progress. This often includes several instances of primary data collection over the first several weeks, either to inform specific sectors (such as shelter needs or access to water and sanitation services) or through a single multisectoral initial rapid assessment conducted by multiple organizations [9]. These assessments are conducted using structured questionnaires, often with a selection of key informants at the community level. Other types of quantitative data collection, usually occurring after the first two weeks of the crisis onset, may include household-level surveys to assess needs in more detail, to monitor progress of a particular intervention, or to evaluate it after completion. A major advantage of such quantitative surveys is that information provided by the respondent is immediately coded by the interviewer, typically by selecting preprogrammed choices in a questionnaire. Data can, therefore, be analyzed immediately at the end of data collection—or even meaningfully collated while survey work is still underway.

The growing availability of the Internet, and especially the widespread use of mobile phones, has led to numerous innovations that have changed (and continue to shape) the practice of data collection in humanitarian crises. In recent years, most humanitarian face-to-face surveys have moved from paper forms to handheld computer-assisted personal interviewing (CAPI) technologies, resulting in high-quality data and faster results [10]. KoBoToolbox, a free and open-source platform based on the OpenDataKit [11], was embraced in 2014 as the preferred humanitarian survey tool by the United Nations (UN) Office for the Coordination of Humanitarian Affairs (OCHA), leading to widespread adoption by a broad range of international and national humanitarian agencies [12]. As of 2019, public KoBoToolbox servers received more than 73 million survey submissions and had more than 200,000 users (figures from the authors). For example, 63% of humanitarian organizations working in Syria in 2017 reported using KoBoToolbox for primary data collection [13, p. 19]. Similar to traditional paper-based surveys, these tools are generally employed in face-to-face interactions, requiring staff to travel to sampled locations. Under ideal circumstances, a single interviewer can conduct, for example, ten 30-minute household-level interviews per day in a dense urban environment. However, humanitarian crises often pose extreme limits on interviewers' ability to travel, either for logistical or security reasons. This can significantly slow down face-to-face data collection in practice. A sharp increase in violence against aid workers [14] in recent years has further hampered physical access to affected areas. For some contexts, such as in the Pacific region, access to certain communities requires long and expensive travel even in nondisaster circumstances. Time, security, and budgetary limitations sometimes force humanitarian organizations to exclude some groups and communities from face-to-face surveys, leading to large information gaps and unrepresentative data.

In an effort to overcome these challenges, a number of surveys in low- and middle-income countries have moved from face-to-face interviews to phone-based interviews [15]. Cell phone networks can now be accessed by 96% of the world's population [16, p. 8], and smartphones are seen as an essential lifeline among displaced groups [17, 18]. Remote data collection most commonly involves CATI, which are conducted by a trained interviewer (often a call center operator) who asks questions in the local language using structured questionnaires. Similar to face-to-face interactions, responses are classified immediately in close-ended categories to facilitate rapid quantitative analysis. The World Food Programme (WFP) has conducted CATI surveys in 31 humanitarian crises to date to collect critical food security and nutrition information [19, 20]. Other forms of remote phone-based interviews are conducted by text messages and interactive voice response (automated either through voice recognition or

[2] Humanitarian assistance here is considered to include "protection," which "encompasses all activities aimed at obtaining full respect for the rights of the individual in accordance with the letter and the spirit of the relevant bodies of law" [7].

Table 1 Schematic comparison between current interviewing methodology and NLP-supported assessments.

	Face-to-face interviews	Remote interviews
Technology	CAPI (e.g., KoBoToolbox, OpenDataKit, CommCare) for quantitative; digital audio recording or personal notes for qualitative methods	CATI call centers; Interactive Voice Response, text messages
Example use cases	Humanitarian needs assessments, in-depth, or cross-sectional surveys	Rapid surveys with fewer questions (needs assessments, situation monitoring, program monitoring)
Advantages	Allows longer interviews; representativeness; no bias due to technology access	Faster data collection/shorter turnout time; allows more frequent data collection and larger samples; cheaper; allows the collection of data from hard to access and insecure areas
Disadvantages	Expensive, slow, restricted by physical access	Exclusion of people without access to mobile phones or mobile networks; exclusion of people with low literacy for text message surveys; not suited for long and complex surveys
Qualitative data	Can be captured, but rarely done with proper transcription, translation	Can be captured, but rarely done with proper transcription, translation

responding by pressing a number on the phone keypad)—all of which are used primarily for collecting quantitative data [21, pp. 2–3].

The key advantages of CATI surveys are speed and precise categorical data to inform immediate humanitarian operational decision-making. Compared to face-to-face interviews, CATI methods are also more practical for collecting information in physically inaccessible areas and less costly for conducting real-time monitoring [22]. For example, WFP estimates $3–$9 per complete CATI survey versus $20–$40 for a complete face-to-face survey. The most obvious limitation of CATI methods is their inability to reach people who may not have access to mobile technologies (either directly or through a family member). This particularly includes rural populations as well as women, elderly, and disabled people as the only mobile phone is frequently controlled by the male household head. In low-income countries, cell phones are often shared across a larger network of friends or family members [23], while ownership is skewed toward males and urban households with higher income [24, p. 14]. This challenge can be addressed by using complex statistical weighting procedures to reduce bias in the data, which requires reliable demographic baseline data and a minimum level of access to cell phones among the population [24]. Information provided by respondents may also depend on the type of data collection method used, whereby face-to-face interviews can be both conducive and a hindrance to more honest responses, depending on the topic [21], [25, p. 23], [26]. A comparison between face-to-face and remote data collection methods is shown in **Table 1**.

Understanding people affected by crisis qualitatively

The key limitation of both face-to-face and CATI surveys in humanitarian emergencies is the lack of resources to properly handle qualitative information. There have been many initiatives over the past three decades to engage more meaningfully with the people affected by conflicts and disasters by complementing quantitative surveys with qualitative data collection methods. These initiatives include focus group discussions with affected populations to evaluate the effectiveness of a specific program or using hotlines to collect feedback from program participants. However, the use of these methods to better inform humanitarian assistance or give affected people a greater sense of ownership still varies widely across different emergencies [27, pp. 12 and 40]. In these instances, the information provided by respondents or participants is typically captured through handwritten notes, and—in the case of focus group discussions—through audio recordings for later analysis. Nuanced human expression and thinking cannot be easily captured by quantitative methods. An important reason for the limited use of qualitative methods is the time and cost associated with transcription, translation, and content analysis. For example, ten unstructured interviews of 30 minutes each may take 50–100 hours to transcribe and translate, while analysis—even with the help of software—can require an additional 20–40 hours.

Both CATI and CAPI can be used for qualitative methods, such as for recording complex open-ended responses in a largely quantitative survey or even entire unstructured or semistructured interviews. However, most surveys that inform humanitarian operations, monitor progress, or track

Table 2 Comparison between existing quantitative and qualitative data collection methods.

	Quantitative	Qualitative
Example of humanitarian data collection	Household survey on nutrition needs	Semi-structured impact evaluation interviews; feedback collection hotlines
Technologies for data collection	Paper, CAPI, CATI	Paper based note taking, digital audio recording
Speed for data to be available	Fast: Can be coded immediately by the interviewer	Very slow and expensive to scale: Requires recording, transcription, translation, manual coding, and post-hoc analysis
Information depth	Complex questions are hard to code or summarize during the interview (and attempts by interviewers can be very unreliable)	Allows for deeper analysis of knowledge, sentiments, perceptions

opinions about aid actors are conducted using quantitative research instruments. Even in instances where qualitative information is collected, it is rarely systematically analyzed. The main reasons for this limitation are the time and cost associated with using qualitative data well: Transcribing, translating, and coding open-ended responses should ideally be done by trained professionals who process collected data in the language of the affected population. Already, there is too few staff with such skills. Fees for trained translators, transcribers, as well as staff with experience in qualitative analysis can quickly exhaust small budgets. Likewise, the time needed for these activities is often a multiple of the original interview response, creating a lag of days—and often weeks—between data collection and final analysis. Instead, organizations are forced to use a single staff member with limited training to cover all these tasks—or collect fewer qualitative data to begin with.

These challenges, summarized in **Table 2**, prevent organizations from fully using qualitative methods at scale, particularly at the early stages of emergencies when representative survey samples require a large number of interviews and when rapid analysis is essential. Complex questions requiring careful qualitative analysis are, therefore, largely lacking in humanitarian surveys. Instead, a common approach is to replace them with simpler, less nuanced alternatives using closed-ended categorical or ordinal response options. Interviewers are required to interpret and categorize the information immediately, turning each response into a single variable. However, this approach also bears significant risks: Interviewers are often hired rapidly in a crisis and given only minimal training. Similar responses may be coded differently and incorrectly, depending on the interviewer's biases and comprehension of key terms used.

When considering survey methods, significant linguistic challenges can emerge in many of the most urgent humanitarian crises. Human languages can vary in dialect, morphology, grammar, syntax, and semantic structure—all of which define and affect meaning, which itself can be culturally specific, and which can change over space and time [28]. Linguistic forms and meanings can evolve with use in a given culture, and as cultures and their languages interact. Ensuring that the original meaning is fully captured in the target, language can be extremely challenging, particularly for qualitative information [29, 30]. These challenges often make translation from a source to a target language an ongoing interpretive process—even for dedicated professionals.

There are over 3,000 languages spoken in the 42 countries currently experiencing humanitarian crises or situations of concern [31]. For example, there are over 40 languages spoken in the six conflict-affected states of northeast Nigeria alone [32]. National citizens hired as humanitarian interviewers in Nigeria are often asked to read surveys in English, sight translate them into Hausa or a variety of other local languages, and then instantly try to match the respondent's answers with one of the listed categories in English. One study with humanitarian interviewers found that these challenges are pervasive even among highly experienced staff [33, p. 6]. Furthermore, a recent review found that two quantitative assessments in the Rohingya crisis yielded wildly different results on the same indicators [34]. Such findings lead to the question of how the quest for speed may have rendered many humanitarian assessments less reliable.

Even in contexts where there is no significant language diversity, there can be real challenges. Nearly all of the refugees in Cox's Bazar, Bangladesh, speak Rohingya, a language very closely related to Chittagonian, which is spoken by nationally hired humanitarian field staff. Yet, a study found that the differences between the two languages are significant enough that nearly a third of Rohingya refugees were unable to understand a basic sentence in Chittagonian [35]. And since both Rohingya and Chittagonian lack formalized written scripts, survey instruments are often written in two or more languages, sometimes even offering transliterations of Rohingya using Bengali script to help local responders pronounce the questions appropriately. These workarounds are not simple and require significant training and support to maintain an often elusive consistency.

There has been some "bottom-up" innovative data collection in humanitarian assistance. These include crowdsourcing based on text messages [36], feedback collection on aid deliveries [37], voluntary reporting on conflict events [38], and social media used to communicate personal needs [39]. The growing number of humanitarian call centers, which allow affected people to find reliable information or communicate specific complaints with aid delivery, is also included in this category (such as during the 2014–2016 Ebola outbreak in Sierra Leone or for displaced people in Iraq [25, p. 21]). However, for survey purposes, these methods generally do not allow for randomness in the sample, are only practical for a small number of questions, and require large numbers of staff to transcribe and categorize unstructured audio or text data. In short, none of the innovations to date enables a reliable, rapid, and effective use of qualitative data, overcoming the challenges identified above. NLP may offer an opportunity to address this gap.

Using NLP to analyze qualitative data in humanitarian assistance

AI can be defined as scientific and technical attempts to build machines that act rationally, with the capacity to mimic human cognitive functions to perceive, understand, predict, or manipulate [40, pp. 1–30]. When such techniques are deployed to real-life contexts, they are often referred to as AI systems (AIS). AIS are understood as "software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal" [41, p. 6].[3]

In recent years, NLP has received increasing research and commercial attention. NLP is understood as "the use of computational methods to analyze and process spoken or written statements in a language commonly used by humans" [43, p. 2]. Breakthroughs in transcribing, translating, and understanding human speech have been propelled by innovations in the field of AI, including data mining, machine learning, deep learning, and reinforcement learning [44]. The complex tasks of machine translation and natural language understanding have, in turn, been based on advances in information extraction and related subfields [45]. The rapid advances in these fields have led to a boom in commercial applications for digital assistants such as Apple's Siri, Google Assistant, and Amazon Alexa. At the same time, open-source software and training models are being created to replicate or surpass these systems' performance.

The effective use of NLP may enable humanitarian assistance organizations to vastly increase the use of qualitative methods in their interactions with affected people to better enable affected people to communicate their needs. Doing this requires linguistic, technological, and methodological approaches for three separate fields: 1) transcription of voice data into written language; 2) translation from a source language into a target language; and 3) various types of NLP analysis of target language transcripts of what was said, as displayed in **Figure 2**.

Transcription

Proper transcription from audio recordings can be challenging even for skilled annotators. Today, transcription software is already widely available through commercial services such as from Google, IBM Watson, Microsoft Azure, Amazon Web Services, and open-source alternatives such as Transformer [46], DeepSpeech [47], and OpenNMT [48]. For example, Google's speech-to-text service supports 64 languages [49].[4]

The most technical challenge for automating transcription (also referred to as speech recognition systems) is due to the informal nature of survey responses. Spontaneous speech is an "unplanned, non-rehearsed, naturally occurring, and non-experimental type of speech that forms the means of communicating information between individuals" [50, p. 1]. State-of-the-art speech recognition technologies have achieved high recognition accuracy for read texts or constrained-spoken interactions (such as broadcast news). However, accuracy is still rather poor for spontaneous speech, which is often not well structured and contains many disfluencies, leading to a higher error rate for automatic speech recognition systems and to redundant information. The four most popular disfluencies are filler words, repetitions, repairs, and restarts [51]. Repetitions are redundant pieces of information that occur when the speaker pauses for a while, considering what to say next, and then repeats the previous information. Repairs occur when the speaker says something wrong and corrects themselves immediately. Restarts occur when a whole part of a sentence is abandoned, and the speaker starts another one. Personal speaker characteristics also affect transcribed text, such as heavy accents, age-related co-articulations, speaker and language switching, and emotional speech.

[3] Because of the vagueness of many definitions of AI, the Institute of Electrical and Electronics Engineers (IEEE) proposed using the more narrow term "autonomous and intelligent systems" (A/IS) [42].

[4] The platform lists 120 "languages and variants," but 54 refer to dialects of the same language (e.g., Australian English versus American English).

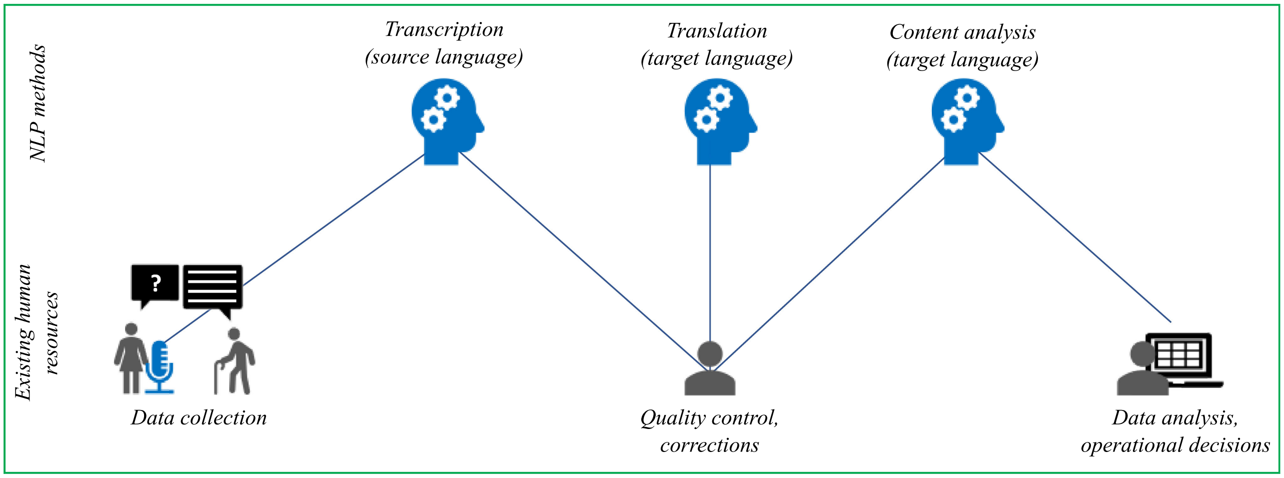

NLP methods
Existing human resources
Transcription (source language)
Translation (target language)
Content analysis (target language)
?
Data collection
Quality control, corrections
Data analysis, operational decisions

Figure 2

Schematic overview of NLP-supported humanitarian assessments.

For this reason, human intervention to correct falsely transcribed speech remains necessary to ensure that the transcript is accurate—even under ideal circumstances (such as formal speech, high audio recording quality, or highly advanced NLP language models).

Translation

As described above, correct translation of qualitative information through human interpreters faces numerous challenges. Advances in deep learning over the last decade have led to a rapid increase in the quality and quantity of available machine translation tools. For example, the quality of machine translation systems for commercially viable languages such as Chinese is now nearing parity with professional human translators in particular domains [52]. Typically, machine translation first requires creating a written transcript in the source language, which is then translated into a target language. However, recent methods also suggest that it may soon become feasible to directly translate source audio into the target language text [53, 54].

Similar to the limitations of automatic speech recognition systems, trained translators are still required to correct NLP-generated translations, ranging from obvious terminology mistakes to maintaining semantic nuances existing in the source language. For historical and commercial reasons, however, there are very few languages in countries affected by humanitarian crises for which machine transcription and translation is available, such as Hausa, Rohingya, Swahili, Fulfulde, or Amharic [55, 56]. Despite a recent willingness from companies such as Google, Microsoft, and Facebook to also support these languages, there remains insufficient training data for machine learning in many of these languages to satisfy the more data-intensive approaches of neural machine translation [57].

These languages also often have little to no existing text or audio datasets that can be used for NLP training purposes. Open-source NLP software, coupled with crowdsourcing of labeled recordings, are becoming a promising option to bridge this new digital divide by creating open-source models for source languages common in contemporary humanitarian emergencies. For example, the Common Voice platform created by Mozilla makes it possible to conduct speech recognition in many of the world's languages, with volunteers already having contributed 2,198 hours of training data across 29 languages as of June 2019 [58]. Efforts to collect more training data for three priority humanitarian languages are also underway among several humanitarian partners, led by the nonprofit organization Translators Without Borders [59]. However, the lack of training data in these source languages is not a quick problem to solve. At a minimum, it involves a coordinated effort to engage with communities of linguists to actively build and improve these datasets. This can be a challenging task for languages such as Rohingya that lack a formalized translation industry or experienced translators.

Analysis

For our purposes, qualitative text analysis refers to the process of establishing the content and meaning of interview responses and other unstructured data. This has traditionally been done by humans at a linear scale (more text requires correspondingly more time), and only with limited use of software, such as NVivo. However, recent advances in multiple NLP disciplines have led to an explosion of the tools available to analyze large amounts of data, including commercial platforms such as Google's Cloud Machine Language Engine or IBM's Watson Natural Language Classifier. A growing number of sophisticated open-source NLP tools can also be used to process qualitative survey data, depending on the type of questions asked and the data sought from responses. Examples include the Natural Language Toolkit for classification (such as the severity of needs) [60], TextRank for summarization (such as identifying most urgent needs) [61], GATE for information extraction (such as the respondent's age) [62], and MALLET for topic modeling (such as understanding a household's top priorities for disaster recovery) [63]. Some of these methods rely on supervised machine learning, which requires creating a human-labeled dataset of real-life responses (e.g., education level) that result in a model for predicting how future responses can be classified. Unsupervised learning, such as for clustering responses based on similar content, does not require data specifically labeled by humans, but instead solely relies on the data collected. Each approach needs to be carefully calibrated to overcome a long list of challenges, such as the use of double meaning or corrections to what was said previously.

There remain several challenges and limitations to employing this approach for surveys. First, the accuracy of NLP analysis varies significantly depending on the amount of data available for training new models and the amount of variation in survey responses. Human verification of any NLP-generated data is, thus, essential to correcting mistakes and improving the model's algorithms. Second, many existing NLP analysis tools require English text as input. As a result, responses collected in other source languages usually need to be translated into English as a target language, adding an additional potential for errors. Alternatively, it is possible to train some NLP toolkits to work in other source languages, but this approach requires substantial resources to generate high-quality models. In such cases, analysis can be done in the source language first, thereby only translating coded results into English (or other languages), if necessary.

Process

This section outlines a proposed functional design of NLP-supported HNAs (as well as similar types of primary data collection) and describes some of the changes needed to adapt existing processes. First, during initial data collection, audio recordings for each question–response pair would need be saved on a mobile device (for face-to-face interaction in CAPI) or on a call center computer (for CATI), along with proper timestamps to mark the beginning and end of each response. The recordings would then be transmitted to a server for further processing.

Second, audio recordings in the source language would be converted into text using an NLP transcription model. Third, in many cases, the transcribed text in the source language would then be translated into a different target language (such as English) for analysis. Quality-control by native speakers and professional translators should be ensured to correct semantic and terminological mistakes stemming from the automatic transcription and translation steps.

Fourth, content analysis based on transcribed, translated, and human-corrected text would require a combination of different NLP analysis techniques (as mentioned in the previous section). Ideally, unsupervised NLP methods should be used wherever possible to reduce the need for both training data and creating custom analysis models. Where training data are required for supervised machine learning methods, enumerators would need to categorize (or label) a sufficient set of initial responses, either at the time of the interview or by reviewing the audio recordings at a later stage. As a fifth and final step, trained specialists should conduct quality control of all NLP-generated categorizations to correct mistakes and improve machine learning models in the process.

The result of this proposed functional design would be a systematically coded dataset that extracts, classifies, and clusters information from spontaneous speech in response structured and unstructured data collection methods. More research is required to 1) test the performance of various NLP toolkits for the same qualitative analysis task; 2) establish standard approaches for analyzing common question types (including by combining different NLP tasks); and 3) establish the feasibility of creating human-labeled training datasets during ongoing humanitarian assistance operations.

Anticipated Benefits

We expect that creating a system to utilize NLP in humanitarian emergency settings would significantly improve the quality of information collected for humanitarian operational purposes. First, by integrating more open-ended questions into surveys or using more qualitative methods overall, humanitarian organizations would be able to gain a more nuanced and accurate understanding of the topic under investigation. Second, as shown in Figure 1, questionnaires can be designed to include fewer questions, making the interview less rigid and flow more naturally. Instead of asking questions that each result in a single variable needed for analysis, questions can be asked in more open-ended ways and followed up with probing questions as needed. This would invite respondents to elaborate on a topic, giving them more of an opportunity to describe what is important to them. Later analysis of these responses is likely to contain much more information (which can be coded into subsequent survey variables) than what is possible with the current question–answer style of quantitative interviews. Third, standard surveys—even those that are primarily quantitative—could eventually be conducted more quickly, as interviewers would not need to spend time entering responses into predesigned multiple-choice options or text boxes. Instead, they can simply focus on conducting the interview and move on to the next question as soon as a satisfactory response has been received, as all audio is recorded for later processing.

Shorter interviews are primarily in the interest of crisis-affected respondents, who can return to their important post-disaster activities more quickly. However, they also benefit humanitarian organizations who can increase the number of sampled respondents (thereby increasing the representativeness for smaller subgroups or regions). Time savings are much greater, of course, for qualitative methods where manual transcription and content analysis would otherwise require hundreds of hours of staff time. Better information and fewer resources used to generate it should result in a more effective, efficient, and nuanced response to needs and a reduction in suffering of people impacted by crises.

The proposed system would substantially augment and integrate well with existing techniques and technologies for field data collection used in humanitarian assistance. As described earlier, the vast majority of humanitarian data collection is conducted between a human interviewer and an individual respondent, either face-to-face using CAPI or remotely in CATI surveys. CAPI technologies, such as KoBoToolbox or the OpenDataKit, so far only allow recording audio responses to individual questions, though work is underway to automatically record background audio without the need for interviewer interaction. Similarly, many organizations using CATI already routinely record calls for quality assurance. Call centers lend themselves in particular to this planned system, as limitations in the field on uploading potentially large numbers of audio files through slow Internet connections are avoided. However, as the following section will show, we also need to focus on the potential risks of using NLP in the context of humanitarian crises.

Anticipating and mitigating ethical challenges

NLP and AIS in general can pose significant new risks to people affected by humanitarian crises. This risk stems both from the increased use of information and communication technologies (ICT) to collect and process ever more detailed personal data, but also from entirely new technologies that can automate analysis and decision-making. Increasing availability of mobile technology, access to the Internet, and the use of social media have led to considerable enthusiasm to deploy new ICT as an answer to many complex societal challenges, including for humanitarian assistance [36, 64, 65]. Growing interest by technology companies in the field of disaster response has only heightened this trend, including forays by Google and Facebook that aim to transform humanitarian operations. Over the last several years, there has been a growing awareness around the ethical risks of increased use of ICT in humanitarian crises, as well as calls to mitigate them [66–68]. Many specific risks have been identified—both from successful and failed technology deployments [69]. Among many issues, this includes remote survey respondents who may be targeted by armed groups [13, p. 19], wrong decisions based on an erroneous sense of accuracy [70], misleading information sourced from social media and crowdsourcing [71], increased vulnerability by excluding those who do not own phones from participation [17] and by using exploitive data mining practices [72], and worsening power imbalances between responders and affected people [73]. One particular area of concern has been the premature introduction of novel technologies during humanitarian crises, using crisis settings as a testing ground for experimentation [74]. Humanitarian practitioners have created a number of specific technical guidelines in recent years to address such concerns. A notable example among these is the ICRC *Handbook on Data Protection in Humanitarian Action*, which is the most detailed and stringent resource to date [75]. Nonetheless, careful balancing of risks and benefits of particular ICT options (including the decision to not use any technology) remains rare [76].

The use of remote data collection technologies faces additional risks, especially when used to enable remote project management in conflict environments. Here, ICT for remote surveys and other forms of information transmission (including apps such as WhatsApp or Signal) are increasingly used in violent conflicts such as in Syria where many aid organizations were not allowed to operate openly, enabling humanitarian actors to engage with affected people from afar [77]. While justified by the need to keep (international) staff safe in conflict environments, remote management can also be motivated by cost reduction or convenience reasons, which clashes with the valued principle that proximity to populations in need is essential for humanitarian action.

Data collection technologies supported by AI face yet another layer of risks. Applying such systems to understand people and their behavior has been shown to reinforce or even exacerbate human biases: If the AI training data reflects existing social biases or is constructed based on easily accessible but unrepresentative data, the results can not only repeat but amplify social inequities. For example, commercially available face recognition software performed very poorly for dark-skinned women while excelling with pictures of white men [78], while voice recognition was found to be less accurate for women and speakers of minority dialects [79]. Many of these often-proprietary tools do not publish their source code and underlying training data. As a result, many well-intentioned systems have resulted in negative or controversial results due to different forms of biases that were only discovered after they were deployed [80, 81]. However, aiming for greater transparency can also pose risks: The authors of a recently created system that can write complete news articles given only a headline decided not to publish their source code, fearing it might be used to spread "fake news" maliciously [82].

NLP used to transcribe and analyze human speech is associated with very low immediate risk, but if such information is used to make recommendations for human operators—as in the case of German immigration officials using NLP to detect identity fraud among refugee claimants [83]—the level of potential risk increases significantly. In general, for example, a model could perform better for young, educated, urban males (whose voice or written data may be more easily accessible), thereby leaving women and many other members of society misrepresented in surveys. In order to counter the risk of entrenching societal biases in NLP methods, it is essential to include a large demographically representative training sample used in the process of creating new models (especially women, the elderly, and minorities)—and to use data weighting that increases the use of outliers in order to optimize the performance for all likely users [84]. Such a process can only reduce biases, not eliminate them. It is, therefore, important to rigorously document the methods for establishing training data, such as publishing the demographic composition of speakers used for compiling voice training data and to continually test model outputs to identify and combat bias. Imbalances in carelessly collected AI training data can lead to unintended biases that can have severe real-life consequences, making it imperative to anticipate, measure, and mitigate these early on.

A recent review has found that existing ethical principles may have little impact in reality as software developers were found to ignore ethics codes in a behavioral experiment [85]. Recent scandals involving Facebook and other companies underline the challenge of embedding and enforcing ethical frameworks and practices in everyday business decisions, especially where unintended consequences may not be apparent for months or years.

Scandals involving AIS and controversially acquired personal data received significant attention in 2018 [86]. Such concerns about the ethical implications of AI tools have only recently been recognized as research priorities. Public concerns around the implications of AI (both current but especially in the future) [87] have resulted in many public policy initiatives to address a wide array of concerns. National governments have largely focused on making AI development a national priority for economic competitiveness, framing it as a national security issue and providing increased research funding. For most countries, there are as yet no laws or policies on the ethical development and use of AI [88]. The European Commission has set up an independent panel to establish specific policy recommendations to this end. In April 2019, the panel released a set of ethical guidelines and operational recommendations for creating "trustworthy AI" with the goal of creating concrete policy recommendations [89]. In May 2019, 42 countries (including all 36 members of the Organisation for Economic Co-operation and Development) adopted an intergovernmental policy guideline to ensure AI systems are designed "in pursuit of beneficial outcomes for people and the planet" [90, Para. 1.1].

Other public policy initiatives include the 23 Asilomar AI Principles as well as the *Montreal Declaration for Responsible Development of AI,* which hopes to "be translated into political language and interpreted in legal fashion" [91, p. 10]. There are also several examples of private sector initiatives to guide managers and engineers to remain ethical. Some of these are brief, nonspecific, and at a high level, such as Google's *AI Principles* and the industry-civil society collaboration Partnership on AI's *8 Tenets*. Others provide more detailed guidance, particularly the Association for Computing Machinery, IEEE, and Microsoft. Separately, the humanitarian sector has developed its own array of principles and operational guidance around data and technology in the field. However, practical implementation remains challenging [92, 93] while none have addressed AI to date.

It is essential for the humanitarian sector to interact with the growing body of ethical AI principles as well as building on the work done to improve data protection with "traditional" ICT. Private and public sector initiatives to provide guidance for ethical development and deployment of autonomous and intelligent systems should, where possible, be used by humanitarian innovators and data scientists instead of creating new guidelines. The use of ethical review boards is essential for shaping, ensuring, and enforcing responsible behavior internally. The increasingly standardized and professionalized governance structures can serve as platforms to promote and review the responsible use of AIS. These include, for example, the Inter-Agency Standing Committee, NGO associations such as InterAction or the International Council of Voluntary Agencies, as well as the cluster coordination system under the UN OCHA.

Conclusion

Current state-of-the-art humanitarian assessments do not capture the complex and rich context in which humanitarian crises unfold, particularly the experience, needs, and resources of affected communities. Furthermore, they rely on primarily extractive quantitatively oriented methods that can miss nuanced qualitative information. Routinely analyzing rich interview and dialogue data could help generate more tailored assessments, improving disaster response, and would arguably improve the relationship between communities and the agencies that seek to assist them. NLP methods offer unique capabilities to systematically transcribe, translate, and analyze interview

Table 3 Components of a proposed pilot phase in a humanitarian crisis for which no transcription and translation models exist.

Workstream	Activities
Preparation	• Together with relevant partners, select the assessment questionnaires most appropriate to use • Modify assessments to include more open-ended questions to cover more information needs
Generate transcription and translation model	• Manually transcribe and translate sample responses (related to humanitarian assistance and general domain language) to serve as training data • Collect speech recordings from volunteers as audio training data (from volunteers recruited on the ground and online). Voice collection will specifically seek out women, older age groups, speakers of minority dialects, among others to avoid biases favoring urban young men. • Verify accuracy and correct as needed, using crowdsourcing and professional translators • Use weighting methods to train the transcription model as accurately as possible for all population groups
Create analysis models	• Transcribe samples of interview responses (first manually, later using transcription model), using categorization from two separate reviewers as labels • Measure the accuracy of interviewer classification through secondary qualitative analysis to establish the human accuracy rate. • Use and compare different NLP approaches to achieve (or surpass) human accuracy rate
Toolkit	• Create a methodological toolkit that can be applied to all humanitarian data collection contexts and integrated with other methods (CATI, mobile data collection) • Provide detailed recommendations for replication and scaling up of approach in other emergencies • Release all documentation, language models, classification algorithms, and software code through open source licenses
Supporting research and documentation	• Identify practical, data governance, and ethical issues that need to be addressed by future humanitarian AIS using NLP • Conduct a scoping review of prevailing NLP algorithms and training data • Test available transcription models to document existing biases • Evaluate and publish all results

responses. Advancing the use of NLP, however, will first depend on research to establish the viability of the five-step proposed functional design of an NLP tool, described earlier. It will also require the support of many partners, including UN agencies, international and local NGOs, donors, as well as private sector organizations. It will also require ensuring the ethical design and use of the NLP tool in a manner consistent with the specific complexities and needs of humanitarian crises. **Table 3** briefly describes the components of a proposed pilot phase in a humanitarian crisis for which no transcription and translation NLP model (commercial or open source) exists so far.

NLP should complement, not replace, face-to-face interviews: Surveys conducted in person allow much more control over the interviewing environment, are able to extend over a longer time (e.g., for cross-sectional studies), and can go into more depth than phone interviews. Furthermore, in-person interviews offer a more personable way to interact with people who have suffered trauma and are struggling to recover. Increasing the usage of existing NLP methods in humanitarian assistance operations has enormous potential benefits by enabling better two-way communication through which affected people can better communicate their needs. We hope that the proposed approach allows humanitarian responders (as well as other survey-intensive domains, such as public health) to rethink the current methodological paradigm that holds that qualitative methods are not compatible with large population samples.

There is an urgent need to bridge the growing gulf between the people affected by humanitarian emergencies and response professionals through improving the quality and quantity of information provided by the affected population. Research and software development are needed to make NLP technology relevant, accessible (and free) for all humanitarian assistance organizations in all crises globally. However, new transcription, translation, and analysis models may not yet prove sufficiently accurate for large-scale deployment in the complex environments that often characterize humanitarian emergencies. Regardless of success, lessons learned on bias reduction and automated classification, as well as the practical results such as trained transcription models, will be immensely useful for humanitarian assistance and academic research, as well as the overall use of NLP in applied domains.

References

1. P. K. Clarke, “The state of the humanitarian system 2018,” ALNAP/ODI, London, U.K., 2018.
2. UNHCR, “Global trends 2017,” Geneva, Switzerland, 2018.
3. World Health Organization, “Risk reduction and emergency preparedness,” Geneva, Switzerland, 2007.
4. Emergency Data Science Workshop, “Natural language processing for humanitarian survey work,” 2018. [Online]. Available: https://emergencydatascience.org/challenge_nlp_surveys/
5. D. A. Ghazali, M. Guericolas, F. Thys, et al., “Climate change impacts on disaster and emergency medicine focusing on mitigation disruptive effects: An international perspective,” *Int. J. Environ. Res. Public Health*, vol. 15, no. 7, pp. 1–13, 2018, doi: 10.3390/ijerph15071379.
6. OCHA, “Global Humanitarian Overview 2019,” Geneva, Switzerland, 2018.
7. IASC, “Protection of internally displaced persons,” Geneva, Switzerland, 1999.
8. J. Pictet, “The Fundamental Principles of the Red Cross,” Geneva, Switzerland, 1979.
9. IASC, “Multi-Sector Initial Rapid Assessment Guidance,” Geneva, Switzerland, 2015.
10. N. Mock, N. Morrow, and A. Papendieck, “From complexity to food security decision-support: Novel methods of assessment and their role in enhancing the timeliness and relevance of food and nutrition security information,” *Global Food Secur.*, vol. 2, no. 1, pp. 41–49, 2013.
11. Y. Anokwa, C. Hartung, and W. Brunette, “Open source data collection in the developing world,” *Computer*, Long Beach, CA, USA, vol. 42, no. 10, pp. 97–99, 2009.
12. OCHA, “World humanitarian data and trends 2015,” New York, NY, USA, 2015.
13. Building Markets and Orange Door Research, “What is the point if nothing changes: Current practices and future opportunities to improve remote monitoring and evaluation in Syria,” New York, NY, USA, 2018.
14. Humanitarian Outcomes, “Aid worker security database,” 2019. [Online]. Available: https://aidworkersecurity.org. Accessed: Feb. 8, 2019.
15. D. G. Gibson, A. Pereira, B. A. Farrenkopf, et al., “Mobile phone surveys for collecting population-level estimates in low- and middle-income countries: A literature review,” *J. Med. Internet Res.*, vol. 19, no. 5, p. e139, May 2017.
16. ITU, “Measuring the information society report 2018,” ITU, Geneva, Switzerland, Rep., 2018.
17. D. Poole, M. Latonero, and J. Berens, “Refugee connectivity: A survey of mobile phones, mental health, and privacy at a Syrian refugee camp in Greece,” Cambridge, MA, USA: Harvard Humanitarian Initiative and New York, NY: Data & Society Research Institute, 2017.
18. UNHCR, “Connecting refugees: How internet and mobile connectivity can improve refugee well-being and transform humanitarian action,” Geneva, Switzerland, 2016.
19. N. Morrow, N. Mock, J. M. Bauer, et al., “Knowing just in time: Use cases for mobile surveys in the humanitarian world,” *Procedia Eng.*, vol. 159, no. June, pp. 210–216, 2016.
20. A. Robinson and A. Obrecht, “Using mobile voice technology to improve the collection of food security data: WFP’s mobile vulnerability analysis and mapping,” ODI/ALNAP, London, U.K., 2016.
21. C. Lamanna, K. Hachhethu, S. Chesterman, et al., “Strengths and limitations of computer assisted telephone interviews (CATI) for nutrition data collection in rural Kenya,” *PLoS One*, vol. 14, no. 1, 2019, Art. no. e0210050.
22. J.-M. Bauer, K. Akakpo, M. Enlund, et al., “Tracking vulnerability in real time: Mobile text for food security surveys in eastern democratic republic of Congo,” *Africa Policy J.*, vol. 9, p. 36, 2013.
23. T. Kreutzer, “Internet and online media usage on mobile phones among low-income urban youth in Cape Town,” in *Proc. Int. Commun. Assoc. Conf.*, 2009.
24. B. Leo, R. Morello, J. Mellon, et al., “Do mobile phone surveys work in poor countries?” Center for Global Develop, Working Paper no. 398, 2015. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2623097
25. R. Dette, J. Steets, and E. Sagmeister, “Technologies for monitoring in insecure environments,” Global Public Policy Institute, Berlin, Germany, 2016.
26. L. F. Langhaug, Y. B. Cheung, S. Pascoe, et al., “How you ask really matters: Randomised comparison of four sexual behaviour questionnaire delivery modes in Zimbabwean youth,” *Sexual Transmiss. Infection*, vol. 87, no. 2, pp. 165–173, 2011.
27. HHI and ICRC, “Engaging with people affected by armed conflicts and other situations of violence – Taking stock. mapping trends. looking ahead. Recommendations for humanitarian organizations and donors in the digital era,” Cambridge, MA, USA, 2018.
28. G. Deutscher, *Through the Language Glass: Why the World Looks Different in Other Languages*. New York, NY, USA: Henry Hold, 2010.
29. A. Bowden and J. A. Fox-Rushby, “A systematic and critical review of the process of translation and adaptation of generic health-related quality of life measures in Africa, Asia, Eastern Europe, the Middle East, South America,” *Social Sci. Med.*, vol. 57, pp. 1289–1306, 2003.
30. R. Al-Amer, L. Ramjan, P. Glew, et al., “Language translation challenges with Arabic speakers participating in qualitative research studies,” *Int. J. Nursing Stud.*, vol. 54, pp. 150–157, 2016.
31. Translators without Borders, “Language mapping: Putting communication needs on the map,” 2019. [Online]. Available: https://translatorswithoutborders.org/language-mapping/. Accessed: Jun. 24, 2019.
32. Translators without Borders, “Communications dashboard: Internally displaced people in northeast Nigeria,” Apr. 19, 2018. [Online]. Available: https://translatorswithoutborders.org/communications-dashboard-internally-displaced-people-in-north-east-nigeria/
33. Translators without Borders, “The words between us: How well do enumerators understand the terminology used in humanitarian surveys? A study from northeast Nigeria,” Danbury, CT, USA, 2018.
34. ACAPS, “Lessons learned: Needs assessments in Cox’s Bazar,” Geneva, Switzerland, 2019.
35. M. M. Hasan, “The language lesson: What we’ve learned about communicating with Rohingya refugees,” Danbury, CT, USA, 2018.
36. P. Meier, *Digital Humanitarians: How Big Data is Changing the Face of Humanitarian Response*. Boca Raton, FL, USA: CRC Press, 2014.
37. F. Bonino, I. Jean, and P. Knox Clarke, “Humanitarian feedback mechanisms: Research, evidence and guidance,” ALNAP/ODI, London, U.K., 2014.
38. P. Van der Windt and M. Humphreys, “Crowdseeding in eastern Congo,” *J. Conflict Resolution*, vol. 60, no. 4, pp. 748–781, Jun. 2016.
39. Facebook, “A new center for crisis response on Facebook,” Menlo Park, CA, USA: Facebook, Sep. 14, 2017. [Online]. Available: https://newsroom.fb.com/news/2017/09/a-new-center-for-crisis-response-on-facebook/. Accessed: Jul. 30, 2018.
40. S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.
41. High-Level Expert Group on Artificial Intelligence, “A definition of artificial intelligence: Main capabilities and scientific disciplines,” Brussels, Belgium, 2019.
42. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being With Autonomous and Intelligent Systems*, 1st ed. Piscataway, NJ, USA: IEEE Press, 2019.
43. H. Assal, J. Seng, F. Kurfess, et al., “Semantically-enhanced information extraction,” in *Proc. Aerosp. Conf.*, 2011, pp. 1–14.

44. S. Sajad Mousavi, M. Schukat, and E. Howley, "Deep reinforcement learning: An overview," in *Proc. SAI Intell. Syst. Conf.*, 2018.
45. S. Singh, "Natural language processing for information extraction," 2018. [Online]. Available: http://arxiv.org/abs/1807.02383
46. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017.
47. A. Hannun, C. Case, J. Casper, et al., "Deep speech: Scaling up end-to-end speech recognition," 2014. [Online]. Available: http://arxiv.org/abs/1412.5567
48. G. Klein, Y. Kim, Y. Deng, et al., "OpenNMT: Open-source toolkit for neural machine translation," Jan. 2017. [Online]. Available: http://arxiv.org/abs/1701.02810
49. Google Cloud, "Language support," [Online]. Available: https://cloud.google.com/speech-to-text/docs/languages. Accessed: Apr. 20, 2019.
50. M. Alzghool, "Investigating different models for cross-language information retrieval from automatic speech transcripts," Univ. Ottawa, Ottawa, ON, Canada, 2009.
51. S. Furui, M. Nakamura, T. Ichiba, et al., "Why is the recognition of spontaneous speech so hard?"
52. Microsoft, "Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English—Microsoft News Center Hong Kong," 2018. [Online]. Available: https://news.microsoft.com/en-hk/2018/03/15/microsoft-reaches-a-historic-milestone-using-ai-to-match-human-performance-in-translating-news-from-chinese-to-english/. Accessed: May 24, 2019.
53. S. Bansal, H. Kamper, K. Livescu, et al., "Low-resource speech-to-text translation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, vol. 2018-Sep., pp. 1298–1302.
54. R. J. Weiss, J. Chorowski, N. Jaitly, et al., "Sequence-to-sequence models can directly translate foreign speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, vol. 2017-Aug., pp. 2625–2629.
55. J. Z. Abbott and L. Martinus, "Towards neural machine translation for African languages," 2018. [Online]. Available: https://arxiv.org/abs/1811.05467v1
56. J. Gu, Y. Wang, Y. Chen, et al., "Meta-learning for low-resource neural machine translation," Aug. 2018. [Online]. Available: http://arxiv.org/abs/1808.08437
57. P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proc. First Workshop Neural Mach. Translation*, 2017, pp. 28–39, doi: 10.18653/v1/w17-3204.
58. CommonVoice, "Languages," 2019. [Online]. Available: https://voice.mozilla.org/en/languages. Accessed: Jun. 20, 2019.
59. A. Ansari and R. Petras, "Gamayun: The language equality initiative," Translators without Borders, 2018.
60. E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proc. ACL Workshop Effective Tools Methodologies Teaching Natural Lang. Process. Comput. Linguistics*, 2002.
61. R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in *Proc. ACL 2004 Interactive Poster Demonstration Sessions*, 2004, doi: 10.3115/1219044.1219064.
62. H. Cunningham, V. Tablan, A. Roberts, et al., "Getting more out of biomedical documents with GATE's full lifecycle open source text analytics," *PLoS Comput. Biol.*, vol. 9, no. 2, Feb. 2013, Art. no. e1002854.
63. A. K. McCallum, "MALLET: A machine learning for language toolkit," 2002. Accessed: Jun. 1, 2019. [Online]. Available: http://mallet.cs.umass.edu/
64. P. Meier, "New information technologies and their impact on the humanitarian sector," *Int. Rev. Red Cross*, vol. 93, no. 884, 2011.
65. Harvard Humanitarian Initiative, "Disaster relief 2.0: The future of information sharing in humanitarian emergencies," Washington, D.C., USA, 2011.
66. K. B. Sandvik, M. Gabrielsen Jumbert, J. Karlsrud, et al., "Humanitarian technology: A critical research agenda," *Int. Rev. Red Cross*, vol. 96, no. 893, pp. 219–242, 2014.
67. P. N. Pham and P. Vinck, "Technology, conflict early warning systems, public health, and human rights," *Health Humans Rights*, vol. 14, no. 2, pp. E106–E117, 2012.
68. T. Scott-Smith, "Humanitarian neophilia: The 'innovation turn' and its implications," *Third World Quart.*, vol. 37, no. 12, pp. 2229–2251, 2016.
69. K. L. Jacobsen, *The Politics of Humanitarian Technology: Good Intentions, Unintended Consequences and Insecurity*. New York, NY, USA: Routledge, 2015.
70. M. Hunt, J. Pringle, M. Christen, et al., "Ethics of emergent information and communication technology applications in humanitarian medical assistance," *Int. Health*, vol. 8, no. 4, pp. 239–245, 2016.
71. K. Crawford and M. Finn, "The limits of crisis data: Analytical and ethical challenges of using social and mobile data to understand disasters," *Geo J.*, vol. 80, no. 4, pp. 491–502, 2015.
72. P. G. Greenough, J. L. Chan, P. Meier, et al., "Applied technologies in humanitarian assistance: Report of the 2009 applied technology working group," *Prehospital. Disaster Med.*, vol. 24, no. SUPPL.2, pp. 2–5, 2009.
73. K. B. Sandvik and N. A. Raymond, "Beyond the protective effect: Towards a theory of harm for information communication technologies in mass atrocity response," *Genocide Stud. Prevention, An Int. J.*, vol. 11, no. 1, pp. 9–24, 2017.
74. K. B. Sandvik, K. L. Jacobsen, and S. M. McDonald, "Do no harm: A taxonomy of the challenges of humanitarian experimentation," *Int. Rev. Red Cross*, vol. 99, no. 904, pp. 319–344, Apr. 2017.
75. C. Kuner and M. Marelli, *Handbook on Data Protection in Humanitarian Action*. Geneva, Switzerland: International Committee of the Red Cross, 2017.
76. R. Dette, "Do no digital harm: Mitigating technology risks in humanitarian contexts," in *Technologies for Development*, S. Hostettler, S. Najih Besson, and J.-C. Bolay, Eds. New York, NY, USA: Springer, 2018, pp. 13–29.
77. J. Steets, E. Sagmeister, and L. Ruppert, "Eyes and ears on the ground: monitoring aid in insecure environments," Global Public Policy Institute, Berlin, Germany, 2016.
78. J. Buolamwini, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Mach. Learning Res.*, vol. 81, pp. 1–15, 2018.
79. R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proc. 1st Workshop Ethics Natural Lang. Process.*, 2017, pp. 53–59.
80. P. Molnar and L. Gill, "Bots at the gate: A human rights analysis of automated decision-making in Canada's immigration and refugee system," Univ. Toronto, Toronto, ON, Canada, 2018.
81. A. Paul, C. Jolley, and A. Anthony, "Reflecting the past, shaping the future: Making AI work for international development," USAID, Washington, DC, USA, 2018.
82. J. Vincent, "AI researchers debate the ethics of sharing potentially harmful programs," *Verge*, 2019. [Online]. Available: https://www.theverge.com/2019/2/21/18234500/ai-ethics-debate-researchers-harmful-programs-openai. Accessed: Mar. 18, 2019.
83. G. Wood, "The refugee detectives," *Atlantic*, Apr. 2018.
84. A. Amini, A. Soleimany, W. Schwarting, et al., "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proc. AAAI/ACM Conf. Artif. Intell., Ethics, Society*, 2019.
85. A. McNamara, J. Smith, and E. Murphy-Hill, "Does ACM's code of ethics change ethical decision making in software development?," in *Proc. 6th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, 2018, pp. 729–733.
86. AI Now Institute, "AI in 2018: A year in review," *Medium*, 2018. [Online]. Available: https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e. Accessed: Mar. 18, 2019.
87. M. Brundage, S. Avin, J. Clark, et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," Feb. 2018. [Online]. Available: http://arxiv.org/abs/1802.07228
88. Future of Life Institute, "AI policy resources," [Online]. Available: https://futureoflife.org/ai-policy-resources/. Accessed: Mar. 5, 2019.
89. High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," Brussels, Belgium, 2019.

90. OECD, *Recommendation of the Council on Artificial Intelligence*. Paris, France: OECD, 2019.
91. Montreal Declaration, *Montreal declaration for a responsible development of artificial intelligence 2018*, Montreal, QC, Canada, 2018.
92. N. A. Raymond and B. L. Card, “Applying humanitarian principles to current uses of information communication technologies: gaps in doctrine and challenges to practice,” Harvard Humanitarian Initiative, Cambridge, MA, USA, 2015.
93. D. Hilhorst and N. Schmiemann, “Development in practice humanitarian principles and organisational culture: Everyday practice in Médecins sans Frontières-Holland,” *Dev. Pract.*, vol. 12, no. 3–4, pp. 490–500, 2002.

Received December 14, 2018; accepted for publication October 1, 2019

Tino Kreutzer *Dahdaleh Institute for Global Health Research, York University, Toronto, ON M3J 1P3, Canada (kreutzer@yorku.ca).* Mr. Kreutzer received an M.A. degree in communication from the University of Cape Town, Cape Town, South Africa, and is currently working toward a Ph.D. degree in health with the York University, Toronto, ON, Canada. He previously was with the United Nations Development Programme and other organizations in nine countries in conflict or natural disaster settings. He is an Executive Director of KoBoToolbox and an Advisor to NetHope.

Patrick Vinck *Harvard Medical School and Harvard T. H. Chan School of Public Health, Cambridge, MA 02138 USA (pvinck@hsph.harvard.edu).* Dr. Vinck received a Ph.D. degree from Tulane University, New Orleans, LA, USA, in 2006. He is the Director of research with the Harvard Humanitarian Initiative, Cambridge, MA, USA. He is an Assistant Professor with the Harvard Medical School and Harvard T. H. Chan School of Public Health, Cambridge, and Lead Investigator with the Brigham and Women’s Hospital, Boston, MA, USA. He is a Co-Founder of KoBoToolbox and the Data-Pop Alliance on Big Data.

Phuong N. Pham *Harvard Medical School and Harvard T. H. Chan School of Public Health, Cambridge, MA 02138 USA (ppham@hsph.harvard.edu).* Dr. Pham received a Ph.D. degree from Tulane University, New Orleans, LA, USA, in 2001. She is an Assistant Professor with the Harvard Medical School and Harvard T. H. Chan School of Public Health, Cambridge, MA, USA, and a Lead Investigator with the Brigham and Women’s Hospital, Boston, MA, USA. She is a Co-Founder of KoBoToolbox.

Aijun An *Department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada (aan@cse.yorku.ca).* Dr. An received a Ph.D. degree in computer science from the University of Regina, Regina, SK, Canada. She is a Professor with the Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada. She has authored and coauthored extensively in various well-respected journals and conferences on data mining, databases, machine learning, and NLP.

Lora Appel *Faculty of Health, York University, Toronto, ON M3J 1P3, Canada (lora.appel@yorku.ca).* Dr. Appel received a Ph.D. degree from Rutgers University, New Brunswick, NJ, USA, in 2016. She is an Assistant Professor of health informatics with York University, Toronto, ON, Canada, and a Collaborating Scientist at OpenLab, University Health Network, where she leads “Prescribing Virtual Reality (VRx)” and is Principal Investigator on three CABHI-funded clinical trials.

Eric DeLuca *Translators without Borders, Danbury, CT 06810 USA (eric@translatorswithoutborders.org).* Mr. DeLuca received an M.A. degree in geography from the University of Minnesota, Minneapolis, MN, USA. He was previously a Graduate Teaching Assistant with the University of Minnesota and a Response Team Leader with ShelterBox. He is currently the Monitoring, Evaluation, and Learning Manager with Translators without Borders, Danbury, CT, USA.

Grace Tang *Translators without Borders, Danbury, CT 06810 USA (grace@translatorswithoutborders.org).* Ms. Tang received an M.A. degree in social responsibility from Steinbeis University, Berlin, Germany. She was with Médecins Sans Frontières as a Senior Manager in strategy development, organizational design, and emergency response in complex contexts. She is currently leading the Gamayun program with Translators without Borders, Danbury, CT, USA.

Muath Alzghool *York University, Toronto, ON M3J 1P3, Canada (alzghool@cse.york.ca).* Dr. Alzghool received a Ph.D. degree in computer science from the University of Ottawa, Ottawa, ON, Canada, in 2009. He is currently a Postdoctoral Fellow with York University, Toronto, ON, Canada, in the field of information retrieval (IR) and NLP.

Kusum Hachhethu *United Nations World Food Programme (WFP), Rome 00148, Italy (kusum.Hachhethu@wfp.org).* Ms. Hachhethu received an M.S. degree in food policy and applied nutrition from Tufts University, Medford, MA, USA, in 2014. She was with UNICEF, Tufts University, and various USAID-funded Nutrition Projects. She is currently a Food Security and Nutrition Analyst with WFP’s Vulnerability Analysis and Mapping team, United Nations World Food Programme (WFP), Rome, Italy.

Bobi Morris *International Rescue Committee, New York, NY 10168 USA (bobi.morris@rescue.org).* Ms. Morris is currently working toward a Dr.PH. degree in epidemiology with the Columbia University Mailman School of Public Health, New York, NY, USA, and an M.H.S. degree from the Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, in 2009. She is the Associate Director of Research and Accountability with the International Rescue Committee, New York, NY, USA.

Sandie L. Walton-Ellery *ACAPS, Geneva 1202, Switzerland (swe@acaps.org).* Mrs. Walton-Ellery received a Post Grad Dip. Ed degree from the University of Western Australia, Crawley WA, Australia, in 1990. She has worked in a range of emergency contexts, including Pakistan, Kosovo, Bhutan, Bangladesh, Ukraine, Nepal, Fiji, and Papua-New Guinea. She has contributed to several academic papers and many of the assessment resources developed by ACAPS.

John Crowley *NetHope, Fairfax, VA 22030 USA (john.crowley@nethope.org).* Mr. Crowley received an M.P.A. degree from the Harvard Kennedy School of Government, Cambridge, MA, USA, and an M.A. degree in history of ideas from Boston University, Boston, MA, USA. He was with the IFRC, the UN Secretariat, the World Bank Group/GFDRR, and the Harvard Humanitarian Initiative. He is the Director of information management and crisis informatics with NetHope, Fairfax, VA, USA.

James Orbinski *Dahdaleh Institute for Global Health Research, York University, Toronto, ON M3J 1P3, Canada (orbinski@yorku.ca).* Dr. Orbinski received an M.D. degree from McMaster University, Hamilton, ON, Canada, in 1990, and the M.A. degree in international relations from the University of Toronto, Toronto, ON, Canada, in 1998. He is a Professor with the Faculty of Health and the inaugural Director of York University’s Dahdaleh Institute for Global Health Research, York University, Toronto, ON, Canada.