More than Tweets: A Critical Reflection on Developing and Testing Crisis Machine Translation Technology

Patrick Cadwell, Sharon O'Brien & Eric DeLuca

Dublin City University & Translators without Borders

The application of machine translation (MT) in crisis settings is of increasing interest to humanitarian practitioners. We collaborated with industry and non-profit partners: (1) to develop and test the utility of an MT system trained specifically on crisis-related content in an under-resourced language combination (French-to-Swahili); and (2) to evaluate the extent to which speakers of both French and Swahili without post-editing experience could be mobilized to post-edit the output of this system effectively. Our small study carried out in Kenya found that our system performed well, provided useful output, and was positively evaluated by inexperienced post-editors. We use the study to discuss the feasibility of MT use in crisis settings for low-resource language combinations and make recommendations on data selection and domain consideration for future crisis-related MT development.

Keywords:

crisis translation; crisis; machine translation (MT); post-editing; evaluation; training; citizen translators; data sets

Acknowledgments:
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734211.

Introduction

The potential benefits and harms of translation and technology in crisis settings have begun to be explored by researchers and practitioners inspired by the pivotal work, *Disaster Relief 2.0*, published by Harvard Humanitarian Initiative (2011), which called on scholars to examine potential contributions that technology can make when crisis strikes. The contribution of machine translation (MT) is an area of particular interest. However, research into this area is challenging and requires the resolution of several problems related to core concepts, availability and viability of language technologies for deployment in crisis settings, the

human input required to make such technologies useful, and provision of appropriate data sets needed for the development of language technologies. In order to address some of these problems, we carried out a research project in Kenya to develop and test the utility of an MT system trained specifically on crisis-related content (referred to by us as Crisis MT) in an under-resourced language combination (the French-to-Swahili language pair).[1] The feasibility study involved participants translating crisis-related content, post-editing the output of a specially-designed machine translation (MT) system, comparatively evaluating the output of this against a market-leading MT system, and evaluating the quality of the human translations, post-edits, and raw MT output produced in the study. The goal of the study was to test the reasonableness of mobilizing 'citizen translators' (Federici and Cadwell 2018) to work with MT output in a crisis setting in an under-resourced language combination.

In this paper, we begin by reviewing recent research on crisis translation, crisis MT, and the data sets required to develop language technologies for crisis settings. We then describe the design and conduct of a small feasibility study of Crisis MT carried out by us in Kenya in January 2019. We present findings from the study and discuss issues arising from these findings before drawing conclusions from the study and presenting considerations for future Crisis MT development and testing.

Crisis translation: research foci and blind spots

A review of literature written on the communication that takes place during a crisis such as a disease outbreak, a mass migration, or the aftermath of an earthquake quickly establishes that the potential benefits and risks of translation have gone largely unrecognized in policy and research (O'Brien et al. 2018; Federici et al. 2019). There is an extensive body of literature on crisis or disaster management (e.g., Fischer 2008; Haddow et al. 2011; Sphere Project 2013; Thomas et al. 2013; Waugh and Tierney 2007), but only a small number of commentators have acknowledged negative consequences – for instance, misunderstood risks or poor response decisions – when crisis communication is in a second or third language for those affected, or in a language they do not understand at all (e.g., Santos-Hernández and Morrow 2013). Researchers and practitioners have begun to fill this gap by exploring the potential of translation to contribute to positive outcomes in crisis settings (see, e.g., Federici 2016; Federici and Cadwell 2018; O'Brien and Cadwell 2017; Shackleton 2018). One of the

[1] We adopt the following convention when referring to language pairs, MT systems, and parallel corpora: A-to-B (hyphen 'to' hyphen) for unidirectional language pairs and MT systems and A–B (en-dash) for bidirectional language combinations and parallel corpora.

main trajectories of this research has been to explore the role of translation technologies and MT in crisis (Cadwell 2016; Cadwell and O'Brien 2016; Federici et al. 2019; O'Brien forthcoming). However, researchers confront an array of problems when carrying out research on MT for crisis contexts.

The research context itself lacks some conceptual clarity. Definitions of crisis are debated (see, e.g., Sellnow and Seeger 2013). Also, practitioners and researchers treat crisis, disaster, and emergency as synonyms or near synonyms (see, e.g., Al-Dahash, Thayaparan, and Kulatunga 2016; Seki 2008, 11). We adopt a broad perspective on crisis in our research. We acknowledge Sellnow and Seeger's (2013, 4–20) review of defining characteristics of a crisis and use it to propose that a crisis is a non-routine event that violates expectations, poses a threat to a social group, and requires a response to mitigate the harm. In our treatment of crisis, we emphasize that responding to one may require external assistance – rendering communication and coordination more significant – and may involve varied timelines (short, medium, and long term) and varied levels of action (individual, local, national, supranational, etc.). As such, a crisis setting in our research could include a road transport accident involving a few casualties or equally a cross-border disease outbreak involving many fatalities. When language and culture are significant elements of the communicative scene in such crisis settings, translation has a role to play.

Availability and viability of MT systems are also problematic issues for crisis settings. MT systems have been deployed successfully in industry to allow human translators to deal with increasingly large volumes of texts speedily (e.g., Castilho et al. 2014; Doherty and O'Brien 2014; Gaspari et al. 2014; Guerberof 2009; Koponen 2012; Moorkens et al. 2015; Plitt and Masselot 2010; Teixeira 2014). At times of crisis, too, large volumes of texts need to be translated quickly. However, the texts most needing quick translation in crises are often not well served by MT: the language combinations required are usually not economically viable enough to sustain a pool of professional translators and associated infrastructure, and data sets of sufficient size and quality are often not available (e.g., in the French-to-Swahili language pair). Recently, there has been an increasing focus on MT for low-resource languages (see, e.g., Liu 2018). Low-resource languages lack the linguistic software and large parallel corpora of quality data required to train MT systems (Karakanta, Dehdari, and van Genabith 2018), though there is not always agreement on what languages can, therefore, be considered as having low resources (Liu 2018). Various approaches to mitigate the low-resource problem are being examined in MT development. Approaches being adopted include 'unsupervised MT', where an MT system is trained on multiple

languages at the same time without using parallel data, or 'pivoting', where the MT system translates first into an intermediate language and then into the desired target language.[2] The issue of domain-specific adaptation of MT systems is also important to consider. Work has been done to show the benefits for target text quality of domain adaptation and control in statistical MT engines (Kobus, Crego and Senellart, 2017) and research is underway to explore the benefits of domain adaptation for engines constructed within the more recent neural MT paradigm (Chu and Wang 2018).

The only two examples that we know of concerning a rapid, practical deployment of MT for crisis response involve support in Haitian Creole following the Haiti earthquake and Kurdish during the European refugee crisis. Lewis (2010) and Lewis, Munro and Vogel (2011) developed an MT system over a period of days for the Haitian Creole-to-English language pair to assist with communication between responders and the local community in the aftermath of the earthquake that struck Haiti in 2010. Translators without Borders (2016) worked with Prompsit to develop a basic rules-based MT system for the Kurmanji and Sorani dialects of Kurdish. This project was designed to be available offline where it could be more easily integrated into communication efforts on rescue ships in the Mediterranean or remote islands in Greece where connectivity was limited. The lack of ability of international responders to communicate with the locals, the lack of professional translators and interpreters, the lack of a commercial MT system for these languages, and limited digital linguistic resources were the main features that drove these innovations, and these are not likely to be unique to these two crises. A digital divide persists in crisis settings, and many of the world's poorest, least educated, most vulnerable populations are being disadvantaged (Ansari and Petras 2018).

Even when MT exists and can be successfully deployed in a crisis, people may have little experience of or expertise in post-editing the output of these systems to make it more useful for end users because of the infrequent nature of the language combination (Ansari and Petras 2018). To achieve adequate levels of quality from machine translated texts typically requires human intervention in the form of post-editing. The level of intervention required may depend on the use to which the machine translated text will be put. If the text is intended to be published and disseminated widely, significant post-editing may be required. If the text

[2] The unsupervised approach can be used when there is no directly parallel training data, a situation that can arise for low-resource languages. For a deeper explanation and related work, see: https://iconictranslation.com/2018/09/issue-11-unsupervised-neural-mt/; https://iconictranslation.com/2019/03/issue-28-hybrid-unsupervised-machine-translation/ (Accessed 19 June 2019).

is intended to be gisted or assimilated – for instance, to triage important texts for further dissemination – light or even no post-editing may be required. Successful post-editing by humans requires special expertise and skills, for which training is required (Flanagan and Pulsen Christensen 2014; Koponen 2015), and this is even more the case if the post-editors do not already have experience of translation. Humans must evaluate the output of MT systems to decide on its usefulness and this, too, is a complex and demanding undertaking that benefits from training and experience (Castilho et al. 2018).

Sourcing appropriate data to train MT engines for crisis settings is also challenging in various ways. Many data sets of crisis-related information focus on one element in the timeline of a crisis – the 'response' phase that occurs immediately after onset. The disaster response and management field emphasizes a number of phases of crises, typically in a cyclical pattern along short-, medium-, and long-term timelines. A variety of cycles have been used since the 1980s to describe and explain the management of crisis, emergency, and disaster events. Alexander (2002, 5) proposed a now authoritative version comprising four phases: *mitigation* (to reduce future impacts, e.g., through building codes); *preparedness* (to reduce imminent impacts, e.g., through training); *response* (to deal with the immediate aftermath, e.g., through search and rescue); and *recovery* (to restore affected populations to normality, e.g., through temporary housing). Typically, when crisis is mentioned, people think of the response phase and of its immediate information requirements; for instance, entering the keywords 'crisis communication' and 'response' in an academic database will return more than five times as many results as entering 'crisis communication' and any other stage of a crisis.[3] By focusing on 'response' only, crisis technology developers risk overlooking considerable variations in content and text types. Let us point to three examples to illustrate.

TREC-IS is an information retrieval challenge supported by the US National Institute of Standards and Technology (NIST), PSCR (or Public Safety Communications Research Division), and the University of Glasgow.[4] TREC-IS creates information retrieval task challenges for academia and industry to automatically process social media streams during emergencies in order to categorize them and prioritize aid requests. The 'event type profiles' include bombing, earthquake, flood, typhoon/hurricane, wildfire, or shooting profiles. This initiative is very valuable, but it is worth pointing out that the event type profiles are limited

[3] At the time of writing, entering 'crisis communication' + 'response' into the Web of Science database returned 408 hits, while entering 'crisis communication' + 'preparedness', 'recovery', or 'mitigation' respectively returned 75, 38, and 14 hits.

[4] See: http://dcs.gla.ac.uk/~richardm/TREC_IS/ (Accessed 5 March 2019).

(there is no medical emergency, for example), the text type is limited (it analyses social media streams only), and, most importantly for our discussion, (machine) translation is not considered; all tweets and other social media data searched and retrieved are in English. This is valid if the task is only about information retrieval for responders who only speak English. Nonetheless, important tweets from non-English speaking affected locals (or other types of information) are not included.

A similar initiative is the Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP). The first workshop was held in 2017.[5] In this workshop, the data challenges focus on extraction of information and summarization of social media (tweets/microblogs). Though there appears to be some focus on preparedness in one paper, the data involved consists of "[m]icroblogs posted *during* the earthquake in Italy in August 2016" (Mehra and Chandra 2017, 2, our emphasis). One of the papers in this workshop mentions multilingualism in its title (Patel et al. 2017), but the paper has no mention of any specific languages.

A final example worth noting is the US DARPA's LORELEI project. LORELEI stands for Low Resource Languages for Emergent Incidents.[6] According to the description on their website, the goal of this project is "to dramatically advance the state of computational linguistics and human language technology to enable rapid, low-cost development of capabilities of low-resource languages" (Onyshkevych 2014, no pagination). Although MT is factored into this project, the description explicitly states that the research will "not be focused solely on machine translation" and the emphasis is, again, on "situational awareness" in an emerging incident, i.e., the response phase (ibid.). The reason given for not focusing solely on MT, but rather on language technologies to aid situational awareness, is that there would be too much (translated) data to be analysed for *rapid* response to emergent incidences.

This sample of initiatives demonstrates an increasing focus on the importance of information in crisis response, but also a focus from the technological community on the response phase and on information retrieval of English social media content, with little attention given to other languages, to two-way communication, or to other content types and other phases of crises. It could be suggested that this focus is entirely valid because it represents the majority of communication needs in crises. This assumes that 'crisis'

[5] See: https://www.computing.dcu.ie/~dganguly/smerp2017/ (Accessed 5 March 2019).
[6] See: https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents (Accessed 5 March 2019).

represents a neatly bounded domain in which knowledge, terminologies, functions, text types, and so on are shared. However, as we can show in the following section based on data from a related unpublished study carried out by us, the type of information content involved in crises and in crisis translation can be much more varied, encompassing a range of domains rather than a (non-existent) 'crisis' domain.

Social media messaging during the immediate response phase of a crisis is a specific use case which is most relevant during sudden-onset disasters. Yet, this use case is only a small component of overall translation needs during global crisis response work. Non-disaster related crises – arising especially from political insecurity and conflict – are the primary drivers of long-term displacement (see, e.g., IDMC 2018) and often require external support from the United Nations and international non-profit organizations. These institutions primarily operate in English and French, which results in significant needs for translation.

Rapid social media messaging becomes less relevant in conflict-induced displacement, support for refugees and resettlement, and protracted crises. To illustrate the diversity of crisis translation requests in such a setting, Translators without Borders (TWB) analyzed their translation corpora from the Rohingya refugee response in Bangladesh. This particular crisis was selected because it represents a sudden-onset event (around 600,000 Rohingya refugees fled Myanmar for Bangladesh in a matter of weeks) which has turned into a protracted crisis (the vast majority of the displaced population remains in Bangladesh and there is a large presence of national and international organizations supporting the response). Between October 2017 and February 2019, TWB's work in Bangladesh supported 37 different organizations. In total, nearly 700 documents and over 1 million words were translated. For the purposes of analysis, this corpus was refined to 557 unique documents where English was the source language, and the attributes 'deadlines', 'word count', 'type of organization', and 'sector' were analyzed. A 20 percent random sample was selected to focus analysis on two key attributes – target audience and format of the content – and these are reported here.

It should be pointed out however, that there are several areas of potential selection bias in this dataset. TWB did not begin translating in Bangladesh until October 2017; over one month after the initial displacement began. As such, there are emergency messages and rapid response materials that likely occurred in the early days of the response which are not reflected in this data set. However, like TWB, many organizations were still setting up operations during that initial month so there was limited communication for the most part. This data set is also not representative of all types of organizations equally. Among this

corpus, approximately 75 percent of the requests were from international NGOs and 25 percent were from UN agencies. TWB's eligibility criteria for organizations to whom translation services are offered exclude governments, religious institutions without a secular separation, and for profit companies, even if they are active in humanitarian response efforts (TWB 2019). Additionally, TWB does not accept all categories of translations. Legal documents, fundraising or lobbying materials, and human resource documents are all excluded (ibid.). Still, TWB is by far the largest humanitarian translation organization in the world, and the fact that they work with a variety of organizations means the data set is less biased by the specific activities or areas of focus of one organization. So while this corpus is not necessarily representative of all crisis translations that occurred during the Rohingya refugee crisis, it does offer one of the most diverse and informative cross sections of crisis translation work to date.

Generally speaking, translations were neither urgent nor comprised of short strings in the Rohingya refugee crisis. Based on the overall corpus of work, the average translation deadline was 11.2 days (269 hours) and the median deadline was 4.2 days (100 hours). The average word count per request was 1,662 words with a median word count of 794 words. The shortest translation request was for a sign above a building and consisted of 2 words, while the longest request involved 30,972 words of an informal education curriculum targeted at teachers. In four quartile bins, 25 percent of the translation requests had less than 286 words, 25 percent had between 287 and 794 words, 25 percent had between 795 and 1684 words, and the final 25 percent had more than 1,685 words. While many of the translation requests in the fourth quartile could be considered statistical outliers by standard definitions, we have decided not to omit them in the majority of this analysis. These large translation requests represent a significant amount of work and time commitment and are important to consider in the portfolio of translation requests.

Translations were requested in 16 different humanitarian sectors, but the top 3 sectors by word count accounted for 74 percent of all translation work. Table 1 offers three examples of content from each sector selected at random.

INSERT TAB 1 ABOUT HERE

Table 1. Examples of sectoral content selected at random

There is also often an assumption about crisis translation work that it is focused solely on communicating directly with affected communities. Based on the 20% random sample of the corpus (111 documents), only a quarter of all translation requests were public-facing messages, typically delivered over loudspeakers or the radio in the case of the refugee camps in Bangladesh. Since these messages are usually short, they do not comprise a significant amount of translation work. When controlling for word count, public-facing messaging only accounted for 8 percent of TWB's overall translation workload in Bangladesh. Ordered by word count, the type of content translated included: training materials (60 percent), surveys (21 percent), public messaging (8 percent), newsletters (6 percent), and other content (5 percent). Of this content, only 10 percent of the word count was actually targeted at the affected community, while 52 percent was targeted at humanitarian workers, 37 percent was targeted at teachers, and 1 percent was targeted at doctors or nurses.

In summary, our review of the relevant literature and our own research has shown that translation has gone largely unrecognized in crisis communication until recently, but that the role of translation technologies and MT in crisis settings – viewed broadly by us as anything from a small-scale traffic accident to a large-scale disease outbreak – is of increasing interest. The literature has also shown that MT in the languages required in many crisis settings is either unavailable or unviable and few examples of successful deployment exist so far. When MT does exist, those who use it may have little experience of post-editing and the data on which the MT system has been trained may be overly focused on a limited element of an overall crisis timeline and may fail to consider the wide variety of domains, text types, and forms of communication that can be characteristic of a crisis setting. With this context in mind, we collaborated with our non-profit partner, TWB, and our industry partners, Microsoft Research and Unbabel, to train a French-to-Swahili MT system on crisis-related content and to have it used and evaluated by a group of French- and Swahili-speaking citizen translators in a Kenyan setting to gauge whether Crisis MT could be a feasible solution to communication needs in a future crisis.

Crisis machine translation: a feasibility study

The Crisis MT engine leveraged corpora of crisis-related translations provided by TWB and allowed us to build, with the help of Microsoft Research and Unbabel, a system that would translate French-to-Swahili.

The decision to focus on French-to-Swahili was a pragmatic one, grounded in the growing need of TWB to provide better language support in the Ebola crisis in the eastern part of the Democratic Republic of the Congo (DRC). The DRC Ministry of Health and many humanitarian organizations operate primarily in French, yet French is a language mostly understood by the educated and urban populations of eastern DRC. Swahili is a much more effective *lingua franca* for reaching vulnerable communities directly. Despite these needs, it has been difficult to find sufficient translators in these language pairs. The majority of Swahili translators are based in Kenya, Tanzania, or Uganda – three countries where English is the primary international language and French is not widely spoken. As a result, this use case represents a situation where MT could help improve the efficiencies or speed of existing translation services.

Full explanations of the process of development of our Crisis MT engine creation can be found in Cruz Silva et al. (2018) and Liu, Cruz Silva, and Way (2018). In summary here, the French-to-Swahili MT output was created by pivoting on English, that is, by translating from French-to-English and then from English-to-Swahili, due to a lack of parallel data in French–Swahili. The French-to-English engine was trained using one million pairs of randomly selected sentences from the UN Parallel Corpus v1.0 (Ziemski at al. 2016). This model was trained using Marian NMT, a neural MT framework. For the English–Swahili pair, Moses (Koehn et al. 2007) was employed to build the phrase-based Statistical MT (SMT) models. SMT was used for this due to the small size of the parallel corpus. It is generally understood from experiments that SMT outperforms Neural MT when sentence pairs are fewer than one million, depending on the language pairs, however. The 5-gram language models were trained using the SRI Language Toolkit (Stolcke 2002). To obtain word alignment, GIZA++ (Och and Ney 2003) was run on the training data. Minimum error rate training was used to optimize the feature weights. The maximum length of sentences was set as 80. The English–Swahili corpus was prepared from three corpora: 1) 10,406 pairs of parallel data from the general health domain obtained from Microsoft Research; 2) 928 pairs from TWB–Microsoft Research disaster and disease messages; and 3) 6,740 pairs extracted from the TWB Translation Memory. (1) is parallel data collected by Microsoft Research that covers topics such as maternity and health matters in crisis response such as vaccination, food and shelter provision. (2) and (3) were created in collaboration between TWB and Microsoft Research as part of TWB's crisis response translation efforts. These data were shared with us

since DCU, TWB, and Microsoft Research are all partners in INTERACT, the International Network on Crisis Translation.[7]

Our objective in this research was to design a study that would evaluate the extent to which the Crisis MT engine would produce useful text outputs for translators and text users and to evaluate ways in which citizen translators on the ground in a Kenyan setting could be mobilized and trained to post-edit the output of these systems effectively. We follow Federici and Cadwell (2018) in viewing citizen translators as trained or untrained linguists working for a common good, often in a voluntary and ad-hoc capacity. Our participants all spoke French, Swahili, and English, most had little or no experience of translation, either paid or unpaid, and none had prior experience of post-editing.[8] The main steps in the study are summarized in Table 2. Table 2 is intended as a simple overview to orient readers: all steps are explained and discussed in detail below.

INSERT TAB 2 ABOUT HERE

Table 2. Main steps taken in the study

The first element of our study involved sourcing crisis-related messages for translation and post-editing by citizen translators as well as for processing through our Crisis MT. Our goal was to find short, authentic, computer-mediated communication written in French in the domains of health, shelter, and safety and security sent out to affected populations during a crisis. We intended to select texts across all major phases of a crisis as described above and to include a variety of domains and text types. However, our efforts to select a wide variety of crisis-related information across a long timescale were hindered by the fact that many authentic data sets of this kind already available in French had been used by us to train the Crisis MT engine and could not, therefore, be used in any translating, post-editing, or evaluation tasks. Instead, we ran a series of Internet searches online using a selection of keywords in French related to our communicative focus (*alerte, vigilance, risque, santé, abri, securité*) and identified potential sources of relevant communication.

[7] https://sites.google.com/view/crisistranslation/home (Accessed: 3 July 2019).

[8] Approximately one third of our participants had no formal experience of working as a translator, either paid or unpaid. Approximately one third again had carried out fewer than ten paid and ten unpaid translation assignments ever. The remainder identified themselves as experienced translators. No participant had post-editing experience.

This element of the study presented several challenges. It was difficult to claim that many of the sources returned in our searches – particularly those of large, international organizations, such as the UN or Red Cross – produced original source text (ST) messages in French and not target text (TT) messages created in French following an original, usually in English. For this reason, we focused on communication produced by ministries and governments of French-speaking states: France's Interior Ministry, the Ministry of Health of DRC, and the Government of Côte d'Ivoire. It was also difficult to claim that all the communication found had been produced in the context of a real crisis. For instance, a large number of hits were template messages created by crisis communication experts to train local government employees in how to write effectively in crisis settings. For this reason, we focused on communication sent out via Twitter and Facebook – often based on another text type such as a larger report or guideline – for which a timestamp of sending could be verified. We manually trawled the social media accounts of the three sources for STs that would satisfy our search criteria. This time-consuming process returned approximately 100 appropriate messages, mostly sent in the immediate aftermath of a crisis. We verified with TWB that these ST messages had not been included in data used to train the Crisis MT engine.

From this body of 100 crisis-related messages in French, we needed to create two smaller subsets that would be broadly comparable in terms of length, domain, crisis phase, and sender. This would allow us to mitigate a learning effect when a post-editing task was carried out following a translation task; i.e., the same participant would not post-edit the same sentences they had just translated but would post-edit something broadly comparable instead. We also aimed for messages that were generic enough not to require localizing to a Kenyan context; i.e., we attempted to select messages that did not contain proper nouns, acronyms, or other information that would identify the sender as being outside Kenya. We selected 28 messages in French in total: 14 for human translation and 14 for post-editing. The number of messages selected was a pragmatic choice based on an estimate of the largest number of messages that a participant would be able to process within the total participation time for which we had ethical approval and to which participants would consent. Estimates proved accurate and most participants completed their sessions just within the timeframe agreed. Two STs were slightly localized to remove information specific to a French context, and a further five STs were cleaned of new line markers or square brackets before being processed by our Crisis MT system; otherwise, the STs were translated as they had been sent. A small number of typographic errors and non-standard forms of French can be found in the STs and

may have influenced the quality of MT output. The STs were authentic tests disseminated during real-world crises. As errors were not corrected when the messages were originally sent, we did not want to correct the errors during our study as our goal was to test the feasibility of an MT system in as close an approximation to a real-world crisis context as possible. The selected STs and their comparability can be examined in Tables 3 and 4.[9]

INSERT TAB 3 ABOUT HERE

Table 3. Crisis-related messages for human translation

INSERT TAB 4 ABOUT HERE

Table 4. Crisis-related messages for post-editing

We recognize that the selection of STs is a limiting factor in our feasibility study. Many appropriate, authentic texts that would have been ideal for our study had already been used to train the Crisis MT engine and had to be excluded from our selection. Sourcing alternative texts proved to be time-consuming, generated only a small amount of data from a limited crisis time frame, and our sample proved to be only partially representative of authentic translation in crisis settings, as was explained above. Despite having contacts in the academic field of crisis and disaster response, we found it challenging to locate data and to verify that texts were originally written in French. Nevertheless, we succeeded in sourcing a sufficiently large number of appropriate, authentic texts in French from several sources, encompassing more than one text function – ranging from general safety instructions, to factual reports about shelter, to descriptions of basic medical symptoms – across three domains of

[9] Abbreviations have been used in Tables 3 and 4. The key to the abbreviations is as follows:
Sa = Safety and Security;
He = Health;
Sh = Shelter;
Re = Response;
Pr = Preparedness;
Fr = French Interior Ministry;
Co = Cote D'Ivoire Government;
Dr = DRC Ministry of Health;
Fs = French Solidarity and Health Ministry.

knowledge, and from two of the four major phases of a crisis: we were satisfied that the selected data met the needs of our feasibility study, and we moved on to recruit participants.

Our study was to be a comparison of human translations, post-editing, and machine translations into Swahili of the selected French crisis messaging. Through TWB, we recruited two sets of participants: ten speakers of Swahili, French, and English with no prior experience of post-editing to carry out the human translation and post-editing tasks (our 'Citizen Translator' group); and two experienced translators of Swahili, French, and English to evaluate the quality of Swahili output produced by translation, post-editing, and machine translation (our 'Translation Expert' group). English, an official language of Kenya along with Swahili, was chosen as a vehicular language to facilitate the conduct of the study.

Over a period of three weeks in January 2019, participants in the two groups – the 'Citizen Translator' group and the 'Translation Expert' group – were invited to the local office of TWB in Nairobi to carry out their tasks. Each participant in the 'Citizen Translator' group was first asked to read a translation and post-editing brief. This brief was designed to provide participants with a hypothetical context that would help guide their decisions. It tasked them with preparing Swahili messages from French originals that would be sent by mobile phone to a wide range of Swahili-speakers in Kenya during a crisis, and it instructed them to prepare the messages in the shortest timeframe needed to create a message that was adequate, accurate, fluent, and clear. Specifically, participants were asked to consider that their job in the translation and post-editing tasks was to produce messages in Swahili: that expressed all the meaning expressed in the French messages accurately (adequate and accurate); that used well-formed grammar and spelling and terms in common use (fluent); and that would be understood and accepted by a wide range of Swahili speakers in Kenya (clear). The full brief given to participants in the 'Citizen Translator' group can be seen in Appendix A.

Translation and post-editing tasks were carried out online using the free, open-source computer-aided translation tool MateCat.[10] Individual MateCat sessions were prepared by us for each participant in the 'Citizen Translator' group so that they could: (1) translate 14 crisis-related messages in one session from scratch with no MT or translation memory match assistance, nor access to any dictionaries or other language resources; and (2) post-edit 14 broadly similar crisis-related messages in another session with only Crisis MT output prepopulated in the target-text entry box and no other language assistance available.

[10] https://www.matecat.com/about/ (Accessed: 4 March 2019).

Conducting a translation task without supporting resources (dictionaries, prepared glossaries, pre-existing translation memories, online research, etc.) would not be realistic in a contemporary human translation setting. However, translation and technology in crisis settings are partly characterized by their lack of available resources (Cadwell and O'Brien 2016), and this would likely be exacerbated in a low-resource language pair. We felt justified that a clear comparison of human translation (with *no* MT or other resources involved) and post-editing (with *only* MT and no other resources involved) would closely approximate a low-resource crisis translation context. As participants were selected because they had no experience of post-editing, we also asked them, prior to carrying out the post-editing step in (2) above, to take and evaluate a specially-developed, short, online course on the fundamentals of post-editing prepared by us in advance of the study.

The final task asked of our citizen translators was to compare the output of our Crisis MT engine with that of Google Translate and rate a relative preference for adequacy and fluency. Participants were presented once again with all 28 ST messages in French from earlier tasks. This time, the STs were accompanied by two Swahili translations – one produced by our Crisis MT engine and one by Google Translate – with the order randomized. The goal of this relative quality evaluation task (in which A is better than B, B is better than A, or A and B are equal) was not to establish any claim of absolute quality about either engine, but merely to determine the performance of our engine compared to the market leader from a user perspective. In all, each individual session with a participant in the 'Citizen Translator' group took approximately four hours to complete, and participants were recompensed for their time and effort accordingly.

We now had three main bodies of data: human translations in Swahili of 14 crisis-related messages; post-edits of 14 broadly similar crisis-related messages; and machine translation output of all 28 messages. This allowed translation experts to indicate their subjective judgments of the quality of these three types of Swahili translation. Again, we prepared a task brief to guide the participants in their work. In the brief, we explained that participants would be asked to rate each Swahili translation for adequacy and fluency on separate four-point scales. We defined adequacy and fluency for participants, we provided participants with an error typology should they wish to further comment on a translation, and we illustrated the task with a sample evaluation created by a French- and Swahili-speaking member of the research team. Definitions of adequacy and fluency were adapted from TAUS (2013). Participants were asked to think of adequacy as the extent to which the source meaning was expressed in the translation and of fluency as the extent to which the translation

was well-formed in terms of grammar, spelling, term use, etc. The error typology for adequacy consisted of addition, omission, mistranslation, and non-translation of meaningful content. The error typology for fluency consisted of grammar, spelling, and formatting errors, and unintelligible sections. The full brief given to participants in the 'Translation Expert' group – including the definitions and error typologies used and an example evaluation of a French-to-Swahili translation – can be seen in Appendix B.

We were restricted at this step by time and resources and were able to recruit only two participants to our 'Translation Expert' group. The experts were asked to evaluate the work of two citizen translator participants each (meaning that the work of four participants chosen at random could be evaluated in total). The experts were presented with one evaluation form per participant containing 21 French STs accompanied in seven instances by human translations in Swahili, in seven other instances by post-editing in Swahili, and in seven other instances again by the raw output of our Crisis MT engine in Swahili.

Findings from the study

The amount of data generated in our study was small. Nevertheless, we can make some tentative claims from the work of our ten crisis translators and from the evaluations by two translation experts of the work of four of these participants.

Participants in our study were faster at post-editing than at the broadly comparable translation task in most instances. For the nine out of ten participants who were faster post-editing than translating, speed savings for the total tasks ranged from two to 43 minutes, with an average saving of approximately 18 minutes between the two tasks. The participant who saved approximately two minutes between tasks completed the translation task in about 24 minutes (at a rate of roughly 4.69 seconds per word) and completed the post-editing task in about 22 minutes (at a rate of roughly 4.29 seconds per word). The participant who saved approximately 43 minutes between tasks completed the translation task in about 63 minutes (at a rate of roughly 12.31 seconds per word) and completed the post-editing task in about 20 minutes (at a rate of roughly 3.91 seconds per word). For the one participant who was slower post-editing than translating, the post-editing task took approximately 42 minutes longer than the translation task.

No participant in the 'Citizen Translator' group had formal experience of post-editing prior to the study, though some had used MT (e.g., for gisting, as a dictionary, etc.). Participants' use of MT prior to the study varied significantly, and it was a 2:2:1 split for the

responses ‘frequently’, ‘sometimes’, and ‘never’ when participants were asked about their levels of MT experience. Nevertheless, all participants reported in the debriefing following their tasks that they were now positively disposed to MT and post-editing should a real crisis translation task arise; they all thought that MT would be useful, but that post-editing would be required. In addition, all participants evaluated the content and format of the short, online training course on the fundamentals of post-editing positively and found it to be useful in guiding the decisions that they made during the hypothetical crisis translation and post-editing tasks.

Table 5 summarizes the results of the quality evaluation carried out by the two translation experts of the three forms of Swahili messages produced – human translations, post-edits, and raw MT. In Table 5, responses given by evaluators were weighted according to the response’s place on the relevant scale. For instance, if none of the ST meaning was expressed in the TT, such evaluations were awarded only 1 point as ‘None’ was the lowest rating possible on a four-point scale. Similarly, if a TT was flawlessly formed in terms of grammar, spelling, term use, etc., such evaluations were awarded 4 points, as ‘Flawless’ was the highest rating possible on a four-point scale. Calculating weighted scores in this way then produced a total weighted count that could be used to indicate the relative favourability of the evaluations given by the translation experts for each of the three forms.[11]

INSERT TAB 5 ABOUT HERE

Table 5. Summary of expert evaluations of three forms of Swahili messages

It can be seen from Table 5 that the most favourable ratings of adequacy and fluency were for the content that was post-edited, though only by a small margin. The results of human translation were rated the next most favourably, with equal adequacy to post-editing and with fluency only marginally less highly-evaluated. The results of MT were rated the least favourably of the three forms, but not significantly so; the evaluations of raw MT output were only approximately 10–15% less favourable than either post-editing or human translation. The difference between evaluations of human versus machine became much clearer when instances of the top evaluations were compared (i.e., an evaluation where everything in the

[11] In Table 5: ‘Weighted Count’ is the ‘Total Count’ multiplied by the relevant points; ‘Relative Favourability’ is the ‘Total Count’ divided by the ‘Weighted Count’.

ST was expressed in the TT with flawless fluency). MT output achieved ‘Everything’ for Adequacy or ‘Flawless’ for Fluency in only one instance each, while the results of both post-editing and human translation achieved this six- to 11-times more frequently. Nonetheless, the comparison was much more favourable at the second-highest evaluation (i.e, an evaluation where most meaning in the ST was expressed in the TT with good fluency). Here MT performed about half as well as human participants.

To illustrate the type of content that was evaluated by the translation experts, examples of one human translation, one post-edit, and one piece of raw MT output rated by one of the experts have been included in the bulleted lists below. The expert provided a ‘Most’ rating for adequacy and a ‘Good’ rating for fluency for each of these Swahili sentences, meaning that they thought that most of the source meaning was expressed in each Swahili sentence and that the grammar, spelling, term use, etc. of each was achieved at a good level.

Human Translation

- French ST: Si l'un de vos proche [sic] présente des symptômes suspects, appelez gratuitement les équipes d'urgence ou allez au centre de santé public agréé le plus proche pour une prise en charge adéquate.
- Swahili TT: Iwapo mmoja wa wendani wako ana dalili za kushuku,wapigie bila malipo timu ya haraka ama tembelea kituo cha afya kilicho karibu nawe kwa tiba ya haraka.
- Evaluator Comment: The translation contains a grammar error "kushuku,wapigie". Meaningful content "agréé" has been omitted.

Post-Edit

- French ST: Attendez impérativement les consignes diffusées sur les radios avant de changer de posture.
- Swahili TT: Lazima usubiri maelekezo kutoka kwa redio kabla ya kubadilisha pose.
- Evaluator Comment: The translation contains something unintelligible in Swahili i.e. "pose"

Machine Translation

- French ST: Retrouvez toutes les infos sur les opérations de mise à l'abri, l'accueil et l'accompagnement dans les centres d'accueil et d'orientation.
- Swahili TT: Pata habari zote kuhusu kituo cha kukimbia, mapokezi na ushirikiano kwenye kituo cha mapokezi na maelekezo.
- Evaluator Comment: Meaningful content "les opérations" has been omitted. "Les centres" is in plural in the source text while it is rendered in singular form "kituo" in the target text.

While the output of our Crisis MT engine performed satisfactorily in absolute terms in this small study, it also slightly outperformed the market leader Google Translate. Output from our engine was deemed by participants to be better than Google's free online Google Translate engine in 40% of instances, Google output was deemed to be better than our engine's output in 30% of instances, and the outputs of the two engines were deemed to be equal in a further 30% of instances. It must be remembered that this was a purely relative evaluation of quality, and we can make no claims about absolute quality from this exercise: i.e., participants could have thought that the output of both engines was unsatisfactory, but that the output of our engine was less so.

In short, we can argue that, in our small study, output from our Crisis MT engine: performed at least as well as the market leader; improved the speed of inexperienced citizen translators working under stressful conditions while maintaining a favourable level of quality compared to pure human translation; and was perceived to provide useful output, once a human was still available to post-edit the results. We can also assert that these inexperienced citizen translators were positively disposed to receiving a short, online course in fundamental principles of post-editing prior to a crisis translation task and evaluated the content and format of the training favourably.

Discussion

The study above is admittedly limited in scope, but it did provide some initial encouraging results for our Crisis MT engine and for the capacity of citizen translators to post-edit output from such systems successfully. At the same time, this experiment opened up new and important questions for us, which have to do with 'crisis data sets', their definition and their selection for the development and evaluation of translation technologies that serve crisis settings, which we feel are important to raise here.

A conceptualization of crisis translation as being necessarily rapid and always directed at affected communities assumes certain requirements – namely that speed and quality are primary priorities of translation systems and processes. While this is often the case, there are also many instances where the information in need of translating is not extremely time sensitive and requires a 'fit-for-purpose' approach to quality. For instance, the messages in our feasibility study sourced from the preparedness phase are likely to have been less time sensitive than those sourced from the response phase. In a real-world crisis setting, such time sensitivity should be factored into the approach to quality taken and to the method of translation that would achieve the desired level of quality. Thinking beyond the narrow use cases of social media messaging towards a more nuanced understanding of crisis translation is important in designing more appropriate technologies and human support initiatives in the future. This approach would allow for more targeted training data to reduce the dependencies on large parallel datasets. There are many fit-for-purpose language technologies that could offer genuine utility in humanitarian crises. These include, for instance: speech-to-text tools, which could be used by humanitarian responders to process large amounts of audio-recorded feedback from affected populations; text-to-speech tools in low-resource languages, to help support communication for low-literacy populations; speech-to-speech interpreting tools, to help facilitate basic, two-way communication between affected populations and humanitarian responders; and MT already optimized for education and health domains. These tools may not deliver perfect results across all applications, and they would need to be available either offline or in contexts with limited bandwidth in low-resource languages that are relevant in crises. This approach could also allow for citizen translators to fill a valuable gap in the translation process, especially in languages with limited professional translation capacity, as agents of human input to make these language technologies more effective.

Our study was also limited by the fact that the data sets we selected had to be focused on just two phases of a standard crisis timeline (preparedness and response) and on a limited and hard-to-source body of STs. We know from the corpus of authentic crisis translations generated during the Rohingya humanitarian crisis in Bangladesh, and presented above, that texts relevant to crisis settings relate to a much wider variety of content and to more diverse use cases than simply safety and security, shelter, or health messages at the preparedness and response phases of a crisis. Many translations in the Rohingya crisis corpus had longer timelines and a more 'fit-for-purpose' quality requirement than might have been assumed: for instance, a crisis translation assignment to translate a massive amount of information for coding or triage purposes might require speed but not full, high-quality, human translation. A

'fit-for-purpose' approach to crisis translation using MT and its post-editing, especially in languages with limited professional translation capacity, would enable human translators in crisis settings to focus their efforts on timely two-way communication with crisis-affected populations – for instance, through social media messaging – where an understanding of nuance, local context, slang, or text conventions like hashtags is required. Obviously, the type of messaging will vary depending on the type of crisis response that is required. For instance, message types differ depending on whether there has been an earthquake, a tsunami, a disease outbreak, a terrorist attack, or a combination of such or other events. For building Crisis MT engines, where domain has been shown to be an important factor for target language quality, the types of crisis and messaging are important topics that also need to be considered.

Further research to make MT and post-editing in crisis settings more useful would be worthwhile. Domain-tuning is known to improve MT output, as we indicated in our literature review, and training MT engines on out-of-domain parallel corpora along with much smaller volumes of specific, in-domain monolingual texts may be an approach to mitigate some of the problems low-resource languages face. In addition, tags can be added in front of sentences to 'direct' the MT system in order to improve aspects such as politeness, gender, or even the target language required. We argue that a domain/content-specific MT engine would be beneficial for crises, rather than a generic engine for an overly broad 'crisis' domain, or indeed an engine that focuses on social media texts only.

Conclusions

In this research we tested the utility of a French-to-Swahili Crisis MT pivot engine developed in collaboration with industry and non-profit partners. We found in a small study of 'citizen translator' participants in Kenya that the output of the engine was considered useful by them in a hypothetical crisis translation setting, especially once they had completed a short training course on the fundamentals of MT and post-editing. Despite these small-scale, positive findings, the design and conduct of our Crisis MT evaluation exercise revealed that the technological community focuses disproportionately on the response phase of a crisis and on English when gathering data to build and evaluate informational supports for crisis settings and ignores much of the variety of content and domains being translated in real-world crises.

There are also limitations to the work that we carried out. The number of participants in the translation and post-editing tasks was small, and the amount of their translations that could be evaluated by experts was restricted by time and resources. Also, sourcing authentic,

appropriate, and varied STs for evaluation in our feasibility study proved to be time-consuming and generated only a small amount of limited data. Furthermore, evaluations of the three forms of Swahili messages produced – human translations, post-edits, and raw machine translation – were based on subjective judgments. Having established in this pilot study that participants were positively – if subjectively – disposed to Crisis MT output, we intend to mitigate the above limitations by expanding our research to other contexts. Our team has developed Crisis MT in another low-resource language combination (Arabic–Greek) and we intend to evaluate this in the context of the migration crisis in Europe using more standard, less subjective measures of usefulness or text quality including, for example, post-editing effort or intelligibility, and selecting texts for evaluation from as wide a variety of sources, text types, and domains as possible.

Finally, we conclude from our study that further research to make MT and post-editing in crisis settings more useful is worthwhile. We will share our Crisis MT engines with the academic and practitioner community once our research project has finished (further details will be made available on the INTERACT website) and they will be available for use and further testing.[12] We will also make available online the short training course developed to instruct citizen translators in the fundamental principles of MT and post-editing, and we continue to have it evaluated and will improve it further before its final release.

References

Al-Dahash, Hajer, Menaha Thayaparan, and Udayangani Kulatunga, 2016. "Understanding the Terminologies: Disaster, Crisis and Emergency." In *Proceedings of the 32nd Association of Researchers in Construction Management (ARCOM)*, 1191–1200. Manchester, UK: Association of Researchers in Construction Management.

Alexander, David. 2002. *Principles of Emergency Planning and Management*. Oxford: Oxford University Press.

Ansari, Aimee, and Rebecca Petras. 2018. *Gamayun: The Language Equality Initiative*. Accessed March 3, 2019. https://translatorswithoutborders.org/wp-content/uploads/2018/03/Gamayun-Language-Equality-Initiative-March-2018.pdf

Cadwell, Patrick. 2016. "A Place for Translation Technologies in Disaster Settings: The Case of the 2011 Great East Japan Earthquake." In *Conflict and Communication: A Changing Asia in a Globalising World*, edited by Minako O'Hagan, and Qi Zhang, 169–194.New York: Nova Science Publishers, Inc.

[12] See https://sites.google.com/view/crisistranslation/home following completion of the INTERACT project in March 2020 for further information (Accessed: 3 July 2019).

Cadwell, Patrick, and Sharon O'Brien. 2016. “Language, Culture, and Translation in Disaster ICT: An Ecosystemic Model of Understanding.” *Perspectives* 24 (4): 557–575. DOI 10.1080/0907676X.2016.1142588

Castilho, Sheila, Sharon O'Brien, Fabio Alves, and Morgan O'Brien. 2014. “Does Post-Editing Increase Usability? A Study with Brazilian Portuguese as Target Language.” In *Proceedings of the 17th Conference of the European Association for Machine Translation*, edited by Marko Tadić, Philipp Koehn, Johann Roturier, and Andy Way, 183–190. Dubrovnik: EAMT.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. “Approaches to Human and Machine Translation Quality Assessment.” In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 9–38. Cham: Switzerland.

Chu, Chenhui, and Rui Wang. 2018. “A Survey of Domain Adaptation for Neural Machine Translation.” In *Proceedings of the 27th International Conference on Computational Linguistics*, edited by Emily M. Bender, Leon Derczynski, and Pierre Isabelle, 1304–1319. Santa Fe, New Mexico: Association for Computational Linguistics. http://aclweb.org/anthology/C18-1

Cruz Silva, Catarina, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. “Extracting In-Domain Training Data for Neural Machine Translation Using Data Selection Methods.” In *Proceedings of the Third Conference on Machine Translation*, 224–231. Brussels, Belgium: Association for Computational Linguistics. http://www.statmt.org/wmt18/WMT-2018.pdf

Doherty, Stephen, and Sharon O'Brien. 2014. “Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking.” *International Journal of Human-Computer Interaction* 30 (1): 40–51. DOI 10.1080/10447318.2013.802199

Federici, Federico. M. (ed). 2016. *Mediating Emergencies and Conflicts*. Houndmills: Palgrave Macmillan.

Federici, Federico. M., and Patrick Cadwell. 2018. “Training Citizen Translators: Design and Delivery of Bespoke Training on the Fundamentals of Translation for New Zealand Red Cross.” *Translation Spaces* 7 (1): 23–43. DOI 10.1075/ts.00002.fed

Federici, Federico. M, Brian. J. Gerber, Sharon O'Brien, and Patrick Cadwell. 2019. The *International Humanitarian Sector and Language Translation in Crisis Situations. Assessment of Current Practices and Future Needs*. London; Dublin; Phoenix, AZ: INTERACT The International Network on Crisis Translation.

Fischer, Henry W. 2008. *Response to Disaster: Fact versus Fiction and Its Perpetuation: The Sociology of Disaster*. Lanham, MD: University Press of America.

Flanagan, Marian, and Tina Pulsen Christensen. 2014. “Testing Post-Editing Guidelines: How Translation Trainees Interpret Them and How to Tailor Them for Translator Training Purposes.” *The Interpreter and Translator Trainer* 8 (2): 257–275. DOI 10.1080/1750399X.2014.936111

Gaspari, Federico, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. “Perception vs Reality: Measuring Machine Translation Post-Editing Productivity.” In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Workshop on Post-Editing Technology and Practice (WPTP3)*, edited by Sharon O’Brien, Michel Simard, and Lucia Specia, 60–72. Vancouver: AMTA.

Guerberof Arenas, Ana. 2009. “Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation.” *Localisation Focus* 7 (1): 11–21.

Haddow, George D., Jane A. Bullock, and Damon P. Coppola. 2011. *Introduction to Emergency Management*. Burlington, MA: Butterworth Heinemann.

Harvard Humanitarian Initiative 2011. *Disaster Relief 2.0: The Future of Information Sharing in Humanitarian Emergencies*. Washington, D.C. and Berkshire, UK: UN Foundation & Vodafone Foundation Technology Partnership.

IDMC (Internal Displacement Monitoring Centre). 2018. *Global Report on Internal Displacement 2018.* Accessed March 3, 2019. http://www.internal-displacement.org/global-report/grid2018/

Karakanta Alina, Jon Dehdari, Josef van Genabith J. 2018. “Neural Machine Translation for Low-Resource Languages without Parallel Corpora.” *Machine Translation*. 32 (1–2), 167–189. DOI 10.1007/s10590-017-9203-5

Kobus Catherine, Josep Crego, and Jean Senellart. 2017. “Domain Control for Neural Machine Translation.” In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*, 372–378. Varna, Bulgaria: Association for Computational Linguistics. DOI 10.26615/978-954-452-049-6_049

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. “Moses: Open Source Toolkit for Statistical Machine Translation.” In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 177–180. Prague, Czech Republic; Association for Computational Linguistics.

Koponen, Maarit. 2012. “Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations.” In *Proceedings of the 7th Workshop on Statistical Machine Translation*, 181–190. New York: Association for Computational Linguistics.

Koponen, Maarit. 2015. “How to Teach Machine Translation Post-Editing? Experiences from a Post-Editing Course.” In *Proceedings of the 4th Workshop on Post-Editing Technology and Practice (WPTP4)*, 2–15. Miami: Association for Computational Linguistics.

Lewis, William, D. 2010. “Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 Days, 17 Hours, & 30 Minutes.” In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010)* (no pagination). Saint-Raphaël, France: EAMT. http://www.mt-archive.info/EAMT-2010-Lewis.pdf

Lewis, William D., Robert Munro, and Stephan Vogel. 2011. “Crisis MT: Developing a Cookbook for MT in Crisis Situations.” In *Proceedings of the 6th Workshop on Statistical Machine Translation*, 501–511. Edinburgh, Scotland: UK Association for Computational Linguistics.

Liu, Chao-Hong. 2018. *Workshop Proceedings of Technologies for MT of Low Resource Languages (LoResMT 2018)*. Accessed March 3, 2019. http://aclweb.org/anthology/W18-2200

Liu, Chao-Hong, Catarina Cruz Silva, Longyue Wang, and Andy Way. 2018. “Pivot Machine Translation Using Chinese as Pivot Language.” In *Proceedings of the 14th China Workshop on Machine Translation*, 1–12. Wuyishan, China: Springer Nature Singapore Pte Ltd. DOI 10.1007/978-981-13-3083-4 _ 7.

Mehra Kanav, and Vibhash Chandra. 2017. “Summarizing Microblogs for Emergency Relief and Preparedness.” In *Proceedings of the First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2017)*, 104–108. Aberdeen, UK: CEUR. http://ceur-ws.org/Vol-1832/

Moorkens, Joss, Sharon O'Brien, Igor A. L. da Silva, Norma B. de Lima Fonseca, and Fabio Alves. 2015. “Correlations of Perceived Post-Editing Effort with Measurements of Actual Effort.” *Machine Translation* 29 (3-4): 267–284. DOI: 10.1007/s10590-015-9175-2

O'Brien, Sharon. Forthcoming. “Translation Technology and Disaster Management.” In *The Routledge Handbook of Translation and Technology*, edited by Minako O’Hagan. Abingdon, Oxon: Routledge.

O'Brien, Sharon, and Cadwell, Patrick. 2017. “Translation Facilitates Comprehension of Health-Related Crisis Information: Kenya as an Example.” *JoSTrans: The Journal of Specialised Translation* 28: 23–51. https://www.jostrans.org/issue28/art_obrien.pdf

O’Brien, Sharon, Federico M. Federici, Patrick Cadwell, Jay Marlowe, and Brain Gerber. 2018. “Language Translation During Disaster: A Comparative Analysis of Five National Approaches.” *International Journal for Disaster Risk Reduction*, 31: 627–636.

Och, Franz Josef and Hermann Ney. 2003. “A Systematic Comparison of Various Statistical Alignment Models.” *Computational Linguistics* 29 (1): 19–52. DOI: 10.1162/089120103321337421

Onyshkevych, Boyan. 2014. *Low Resource Languages for Emergent Incidents (LORELEI)*. Accessed March 3, 2019. https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

Patel Sindur, Nirav Bhatt, Chandni Shah, and Rutvika Nanecha. 2017. “Multilingual Microblog Summarization.” In *Proceedings of the First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2017)*, 116–121. Aberdeen, UK: CEUR. http://ceur-ws.org/Vol-1832/

Plitt, Mirko, and François Masselot. 2010. “A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context.” *The Prague Bulletin of Mathematical Linguistics* 93: 7–16. DOI: 10.2478/v10108-010-0010-x

Quarantelli, Enrico L. (ed). 1998. *What Is a Disaster? A Dozen Perspectives on the Question*. New York: Routledge.

Santos-Hernández, Jenniffer, and Betty Hearn Morrow. 2013. “Language and Literacy.” In *Social Vulnerability to Disasters*, 2nd ed., edited by Deborah S.K. Thomas, Brenda D. Phillips, William E. Lovekamp, and Alice Fothergill, 265–280. Boca Raton, FL: CRC Press.

Sellnow, Timothy L., and Matthew W. Seeger. 2013. *Theorizing Crisis Communication*. Malden, Mass: Wiley-Blackwell.

Shackleton, Jamie. 2018. “Preparedness in Diverse Communities: Citizen Translation for Community Engagement.” In *Proceedings of the Information Systems for Crisis Response and Management Asia Pacific 2018 Conference*, 400–406. Wellington, New Zealand: Massey University. http://ndhadeliver.natlib.govt.nz/delivery/DeliveryManagerServlet?dps_pid=IE37914290

Sphere Project. 2013. *Humanitarian Charter and Minimum Standards in Humanitarian Response*, 3rd ed. Geneva: The Sphere Project.

Stolcke, Andreas. 2002. “SRILM - An Extensible Language Modeling Toolkit.” In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002-INTERSPEECH 2002)*, edited by John H. L. Hansen and Bryan L. Pellom, 901–904. Denver, Colorado: International Speech Communication Association.

TAUS. 2013. *Adequacy/Fluency Guidelines.* Accessed March 3, 2019. https://www.taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines

Teixeira, Carlos S. C. 2014. “Perceived vs. Measured Performance in the Post-Editing of Suggestions from Machine Translation and Translation Memories.” In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Workshop on Post-Editing Technology and Practice (WPTP3)*, edited by Sharon O’Brien, Michel Simard, and Lucia Specia, 450–59. Vancouver: AMTA.

TWB (Translators without Borders). 2019. *Becoming a TWB Partner*. Accessed March 3, 2019. https://translatorswithoutborders.org/partners/Eligibility/

Seki, Kaoruko. (ed). 2008. *Civil-Military Guidelines & Reference for Complex Emergencies*. New York, Geneva: United Nations.

Translators without Borders. 2016. *Translators without Borders Develops the World’s First Crisis-Specific Machine Translation System for Kurdish Languages.* Accessed June 25, 2019. https://translatorswithoutborders.org/translators-without-borders-develops-worlds-first-crisis-specific-machine-translation-system-kurdish-languages/

Waugh, William L., and Kathleen J. Tierney. 2007. Emergency Management: Principles and Practice for Local Government. Washington, D.C.: ICMA Press.

Ziemski, Michał, Junczys-Dowmunt, Marcin, Pouliquen, Bruno. 2016. “The United Nations Parallel Corpus v1.0.” In *Proceedings of the 2016 International Conference on Language Resources and Evaluation (LREC),* edited by Nicoletta Calzolari et al., 1–5. Portoroz, Slovenia: European Language Resources Association (ELRA).

Appendix A

Translation and Post-Editing Task Brief given to participants in the ‘Citizen Translator’ group.

Translation and Post-Editing Brief

Imagine a crisis has occurred in Kenya. You have been asked by Translators without Borders to help communicate some crisis-related messages to Swahili speakers around Kenya. The messages are currently written in French, but they will be sent in Swahili to the mobile phones of people all around Kenya.

These messages will be sent to a wide range of Swahili-speakers: men and women, young and old, people who live in big cities as well as people who live in small villages, people with a high level of education and people with more basic education.

Your job is to produce messages in Swahili:

1. That express all the meaning expressed in the French messages accurately (adequate and accurate);
2. That use well-formed grammar and spelling and terms in common use (fluent);
3. That will be understood and accepted by a wide range of Swahili speakers in Kenya (clear).

You should complete the Swahili messages in the shortest amount of time needed to achieve the goals above (timeliness). Do not spend extra time checking or editing if you think your Swahili message is already adequate, accurate, fluent, and clear.

You will not have access to any dictionaries or other language resources when you are working - this is a crisis situation and you must do the best you can.

For some messages, you will be given a pre-prepared message in Swahili. This message was produced by a machine translation system. You will be asked to complete a short training course on machine translation post-editing. This training will help you to decide how to work with these machine translation messages.

You have been assigned a participant number for today's tasks. You participant number is _____ . (Please ask your facilitator for your participant number if you have not been given one already.)

Appendix B

Translation and Evaluation Task Brief given to participants in the 'Translation Expert' group.

Translation Evaluation Task

In this task, you will be presented with two sets of 21 short messages in French. Under each French message, there is a Swahili translation. In 7 cases, it is a human translation, in 7 other cases it is raw machine translation output, and in 7 other cases again it is machine translation output post-edited by a human. You will not know which is which, and the order in which the different translations are presented has been randomised.

Please evaluate the Swahili translations for each French message for adequacy and fluency. Consider the following definitions of adequacy and fluency in your evaluations.

> Adequacy
>
> "How much of the meaning expressed in the source is also expressed in the target translation" (Linguistic Data Consortium).
>
> Fluency
>
> To what extent the translation is "one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker" (Linguistic Data Consortium).

There will be space for you to comment on your evaluation choices, and we would like you to use this space to comment on any errors in the Swahili translations. When considering errors, please consider the follow error types.

Adequacy	Fluency

Meaningful content has been **added**	The translation contains a **grammar** error
Meaningful content has been **omitted**	The translation contains a **spelling** error
Meaningful content has been **mistranslated**	The translation contains a **formatting** error
Meaningful content has been left **untranslated**	The translation contains something **unintelligible**

Let us give you an example:

FR: N'utilisez pas de feu à flamme nue.
SW: Usitumie moto wa uchi.

	Everything	Most	Little	None
Adequacy (how much of the source meaning is expressed in the translation)		√		
	Flawless	Good	Dis-fluent	Incomprehensible
Fluency (how well-formed is the grammar, spelling, term use, etc.)		√		
Please comment on the translation, if desired.	The Swahili sentence starts very well, using the correct subject marker. It would appear that the sentence is structurally well formed. The only mistake is that something has been omitted before "uchi" to adequately express "à flamme nue".			

Table 1

Name of sector	**% of translation requests (by word count)**	**Examples of content**
Education	32%	● Teacher training material
		● Informal education in emergencies curriculum
		● Disaster risk reduction activity for children
Communicating with communities	21%	● Public messaging around monsoon preparedness in the camps
		● Feedback from the community about humanitarian operations
		● Training material for humanitarian interpreters and field workers
Multi-sectoral	20%	● Survey to be conducted with affected communities covering a variety of topics
		● Self-care training material for humanitarian staff
		● Focus-group discussion guide for a qualitative research project
Protection	7%	● Resource for child-friendly spaces
		● Training materials for teaching humanitarian staff to protect unaccompanied children
		● Guidelines for preventing sexual exploitation and abuse of crisis affected populations by humanitarian workers
Health	7%	● Written and audio public messages during a diptheria outbreak
		● Patient intake form for medical clinics
		● Training handout for humanitarian staff handling psychosocial support activities
Water, Sanitation, and Hygiene (WASH)	6%	● Video content about hand washing
		● Survey about hygiene practices
		● Poster instructing the use of pit latrines
Other	7%	● Survey following up after an emergency shelter distributions
		● Poster about cash distributions
		● Humanitarian newsletter about different fuel sources in the camp

Table 2

	Step 1	→	Step 2	→	Step 3	
	Researchers		Citizen translators (n=10)		Translation experts (n=2)	

	Sourced crisis-related content		Read a task brief		Read a task brief	
	Prepared content in a CAT tool for translation		Translated content in a CAT tool		Evaluated raw MT output from Step 1	
	Prepared content in a CAT tool for post-editing		Took a post-editing course online		Evaluated human translations from Step 2	
	Prepared raw MT output for evaluation		Post-edited content in a CAT tool		Evaluated post-edits from Step 2	
			Evaluated content online			

Table 3

	HUMAN TRANSLATION					
	Cleaned and localized ST followed by our Crisis MT output for reference (in italics)	**ST word count**	**Primary domain**	**Crisis phase**	**ST sender**	
1	N'utilisez pas de feu à flamme nue. *Usitumie moto ulio wazi.*	7	Sa	Re	Fr	
2	Restez à l'abri et NE SORTEZ SOUS AUCUN PRETEXTE. *Endelea salama na usiende kwa PRETEXT yoyote.*	9	Sa	Re	Fr	
3	Réfugiez-vous dans la pièce la plus sûre de votre habitation. *Jikimbilia kwenye chumba kilicho salama kabisa nyumbani kwako.*	10	Sa	Re	Fr	
4	Eloignez-vous des ouvertures pour éviter les projections de verre en cas de bris. *Ondoa mbali na ufunguzi ili uepuke kuacha kioo wakati umeharibiwa.*	13	Sa	Re	Fr	

5	METTEZ-VOUS EN SECURITE. Installez-vous en hauteur à l'abri dans un bâtiment. Coupez l'électricité et le gaz. Mettez les produits d'entretien et produits chimiques en hauteur. *Salama. Kusimama juu ya urefu wa jengo. Zima umeme na gesi. Weka bidhaa za kusafisha na kemikali kwenye mahali pa juu.*	25	Sa	Pr	Co	
6	Prévoyez des moyens d'éclairages de secours et faites une réserve d'eau potable. N'intervenez en aucun cas sur les toitures et ne touches pas aux fils électriques tombes au sol. Prenez contact avec vos voisins et organisez-vous. *Kutoa taa za dharura na kuhifadhi maji ya kunywa. Usichukue hatua yoyote juu ya paa au kugusa waya kwenye sakafu. Endelea kuwasiliana na majirani zako na kujiandaa mwenyewe.*	36	Sa	Pr	Fr	
7	Si l'un de vos proche présente des symptômes suspects, appelez gratuitement les équipes d'urgence ou allez au centre de santé public agréé le plus proche pour une prise en charge adéquate. *Ikiwa mpendwa wako ana dalili zenye tuhuma, tafadhali piga simu ya timu ya dharura kwa bure au kwenda kwenye kituo cha afya cha umma kilichoidhinishwa karibu na huduma inayofaa.*	31	He	Re	Dr	
8	Pour la semaine 47 allant du 19 au 25 novembre 2018, nous avons enregistré : 309 cas suspects investigués et testés au laboratoire, 46 nouveaux cas confirmés, 24 décès de cas confirmés, 12 nouvelles personnes guéries. *Katika juma la 47 la Novemba 19-25, 2018, tuliandika: kesi 309 za watuhumiwa zilifuatiliwa na kupimwa katika maabara, kesi 46 zilizotambuliwa, matukio 24 yaliyothibitishwa, na kesi 12 zilizoponywa.*	36	He	Re	Dr	
9	Partage ce message et sauve des vies. *Shiriki ujumbe huu na uhifadhi maisha.*	7	He	Pr	Dr	
10	#SavoirCestPouvoirAgir. Les premiers symptômes du Poliovirus sont : A. Fièvre et maux de têtes, B. Fatigue, C. Vomissements, D. Raideurs au cou et douleurs dans les membres, E. Paralysie souvent permanente. *#KuwezeshaCestPouvoirAgir. Dalili za awali za virusi vya polio ni: A. Homa na kichwa, B. Uchovu, C. Kuvunja, D. Ugumu wa shingo na maumivu kwenye viungo, E. mara nyingi ni ya kudumu.*	31	He	Pr	Dr	
11	#LesSymptômesduPalu. Le paludisme se manifeste par la fièvre, les maux de têtes, les vomissements ainsi qu'autres signes de type grippal. Ces symptômes apparaissent entre 10 et 15 jours après la piqure de moustique. *#LesSymptômesduPalu. Malaria inajulikana na homa, maumivu ya kichwa, kutapika na dalili nyingine za mafua.*	33	He	Pr	Dr	

	Dalili hizi hutokea siku 10 hadi 15 baada ya kuumwa kwa mbu.					
12	Le théâtre est un art important pour sensibiliser la population sur Ebola. A travers des sketchs comiques ou tragiques, les acteurs font vivre d'intenses émotions au public qui laissent une trace indélébile dans leur mémoire. *Theatre ni sanaa muhimu ya kuongeza ufahamu wa virusi vya Ebola. Kupitia picha za katuni au maafa, wasanii huleta hisia kali kwa umma, na kuacha athari zisizoweza kukubalika katika kumbukumbu zao.*	35	He	Pr	Dr	
13	La coopération se poursuit pour qu'une réponse adaptée à chaque cas soit trouvée. *Ushirikiano unaendelea kuhakikisha kwamba majibu yanapatikana yanafaa kwa kila kesi.*	13	Sh	Re	Fr	
14	Au total depuis le début du démantèlement, ce sont donc 4014 personnes qui ont d'ores et déjà été mises à l'abri. *Tangu mwanzo wa uharibifu, jumla ya watu 4,014 wamehifadhiwa.*	21	Sh	Re	Fr	

Table 4

	POST-EDITING					
	Cleaned and localized ST followed by our Crisis MT output for reference (in italics)	**ST word count**	**Primary domain**	**Crisis phase**	**ST sender**	
1	Coupez le courant électrique du réseau. *Zima nguvu kwenye mtandao.*	6	Sa	Re	Fr	
2	Essayez de rester calme. La situation va évoluer rapidement. *Jaribu kukaa. Hali itaendeleza haraka.*	9	Sa	Re	Fr	
3	Préparez-vous à subir des coupures d'électricité et d'eau potable. *Kuwa tayari kupata uzoefu wa nguvu na maji safi.*	9	Sa	Re	Fr	
4	Attendez impérativement les consignes diffusées sur les radios avant de changer de posture. *Lazima unasubiri maelekezo kutoka kwa redio kabla ya*	13	Sa	Re	Fr	

	kubadilisha pose.					
5	PRENEZ DES NOUVELLES DE VOS PROCHES. Souciez-vous des personnes proches (familles, voisins, personnes vulnérables). Signalez votre départ, destination et arrivé à vos proches. *Pata habari kutoka kwa umaarufu wako. Jihadharini na watu karibu na wewe (familia, majirani, vikundi vya hatari). Ripoti sehemu yako ya mwanzo, marudio na wapendwa wako.*	23	Sa	Pr	Co	
6	Mettez-vous à l'abri. N'utilisez pas votre véhicule. Fermez portes, fenêtres et volets. Placez les groupes électrogènes à l'extérieur de la maison. Prenez vos précautions si vous utilisez un dispositif d'assistance médicale (contactez l'organisme qui en assure la gestion). *Dodge. Usitumie gari lako. Funga milango, madirisha na vipofu. Weka jenereta nje ya nyumba. Ikiwa unatumia vifaa vya matibabu, tumia tahadhari (wasiliana na shirika linaloweza kuitunza).*	38	Sa	Pr	Fr	
7	J'ai été en contact avec une personne malade de la rougeole, même brièvement…Dans ma famille…Dans un lieu d'accueil collectif : crèche, chez l'assistante maternelle, école…Sur mon lieu de travail : bureau, cantine… *Nimekuwa nikiwasiliana na majani, hata kwa muda mfupi ... katika familia yangu ... katika eneo la mapokezi ya pamoja: kitalu, msaidizi wa mama, shule ... mahali yangu ya kazi: ofisi, Canteen ...*	32	He	Re	Fs	
8	Rougeole. Si vous êtes malades : restez à la maison, portez un masque chirurgical, prévenez rapidement les personnes avec qui vous avez été en contact, surtout s'il s'agit de nourrissons, femmes enceintes, personnes immunodéprimées ou sous traitement immunosuppresseur. *Vipimo. Ikiwa wewe ni mgonjwa: kukaa nyumbani, kuvaa mask na haraka kukuonya watu ambao umekuwa wazi, hasa kama wao ni watoto wachanga, wanawake wajawazito, kutokuwa na maambukizi ya kinga au immunosuppression.*	38	He	Re	Fs	
9	Prévention contre la maladie à virus Ebola. *Kuzuia ugonjwa wa virusi vya Ebola.*	7	He	Pr	Dr	
10	Vaccination. #StopEbola. La vaccination contre Ebola est entièrement VOLONTAIRE. Voici toutes les étapes du processus si vous acceptez de vous faire vacciner en tant que contact ou prestataire de première ligne. *Kinga. #StopEbola. Chanjo dhidi ya Ebola ni kabisa kwa hiari. Ikiwa unakubaliana na chanjo kama hatua yako kuu ya mawasiliano au mtoa huduma, zifuatazo ni hatua zote katika mchakato.*	31	He	Pr	Dr	

11	#Prévention. L'utilisation des moustiquaires imprégnées à longue durée d'action est actuellement le seul moyen de protection individuelle contre le Paludisme. A cela il faut associer l'assainissement du cadre de vie des communautés. *#Prévention. Matumizi ya nyavu za kutibiwa na wadudu kwa muda mrefu ndiyo njia pekee ya kupambana na malaria. Hii lazima ihusishwe na afya ya mazingira ya viumbe hai.*	32	He	Pr	Dr	
12	Vitesse de propagation d'Ebola. Une personne malade non-isolée contamine en moyenne 70 personnes en milieu urbain. L'isolement rapide de la personne malade et la vaccination de ses contacts permettent d'arrêter la propagation du virus. *Kasi ya maambukizi ya virusi vya Ebola. Mgonjwa asiyetambuliwa ana wastani wa watu 70 waliosababishwa katika maeneo ya mijini. Kutengwa haraka kwa wagonjwa na chanjo ya mawasiliano husaidia kuzuia kuenea kwa virusi.*	34	He	Pr	Dr	
13	3242 majeurs ont donc été accueillis en CAO et 772 mineurs ont rejoint le CAP. *Kwa hiyo, CAO ilipokea wataalamu 3,242 na watoto 772 walijiunga na CAP.*	15	Sh	Re	Fr	
14	Retrouvez toutes les infos sur les opérations de mise à l'abri, l'accueil et l'accompagnement dans les centres d'accueil et d'orientation. *Pata habari zote kuhusu kituo cha kukimbia, mapokezi na ushirikiano kwenye kituo cha mapokezi na maelekezo.*	20	Sh	Re	Fr	

Table 5

	Human Translation				**Post-Editing**				**Machine Translation**				
	Adequacy	**Total Count**		**Weighted Count**	***Adequacy***	**Total Count**		**Weighted Count**	***Adequacy***	**Total Count**		**Weighted Count**	
	Everything (4 points)	10		40	Everything (4 points)	11		44	Everything (4 points)	1		4	
	Most (3 points)	16		48	Most (3 points)	14		42	Most (3 points)	8		24	
	Little (2 points)	1		2	Little (2 points)	2		4	Little (2 points)	5		10	
	None (1 point)	1		1	None (1 point)	1		1	None (1 point)	0		0	
		28		**91**		**28**		**91**		**14**		**38**	

	Relative Favourability	**3.25**			**Relative Favourability**	**3.25**			**Relative Favourability**	**2.71**			
	Fluency	**Total Count**		**Weighted Count**	***Fluency***	**Total Count**		**Weighted Count**	***Fluency***	**Total Count**		**Weighted Count**	
	Flawless (4 points)	6		24	Flawless (4 points)	7		28	Flawless (4 points)	1		4	
	Good (3 points)	15		45	Good (3 points)	17		51	Good (3 points)	8		24	
	Dis-fluent (2 points)	6		12	Dis-fluent (2 points)	3		6	Dis-fluent (2 points)	5		10	
	Incomprehsible (1 point)	1		1	Incomprehsible (1 point)	1		1	Incomprehsible (1 point)	0		0	
		28		**82**		**28**		**86**		**14**		**38**	
	Relative Favourability	**2.93**			**Relative Favourability**	**3.07**			**Relative Favourability**	**2.71**			