COPiL CAMBRIDGE OCCASIONAL PAPERS *in* LINGUISTICS
Volume **14(1)**, Article 4: pp. 124–144, 2022 | ISSN: 2050-5949

Accelerating Human Translation of Public Health Information into Low-Resource Languages with Machine Translation

Dimitra Stasinou
Computer Laboratory,
University of Cambridge

Theresa Biberauer
Theoretical & Applied Linguistics,
University of Cambridge

Ebele Mọgọ
MRC Epidemiology Unit, School of Clinical
Medicine, University of Cambridge

Andrew Caines
Computer Laboratory,
University of Cambridge

Abstract We discuss the potential role of machine translation technology in the cross-linguistic dissemination of reliable information during health emergencies. It is clear that translation into low-resource languages is one way to enable greater access to high quality health information around the world. We consider two scenarios: the first being the set of languages already served by publicly available translation services and the degree of human correction needed on outputs from these; the second being those languages for which the only extant option is the development of research models with available training data. In both cases the human-in-the-loop aspect is essential, in order to ensure the accuracy of public health advice is not lost in translation. We report on experiments in translating COVID-19 information from English into Swahili. This language pair allows us to examine both scenarios. We show on the one hand that translations available from Google Translate are good enough for humans to work on and correct rather than re-writing from scratch; and on the other hand that data collection focused on the public health domain is needed for model training from scratch.

1 Introduction

In this paper we envisage 'human-in-the-loop' machine translation scenarios to aid in the dissemination of reliable public health information during future health emergencies. The proposal applies to all languages, but is especially relevant to those languages which are low-resource in natural language processing terms (i.e. in terms of available datasets and pre-trained models).

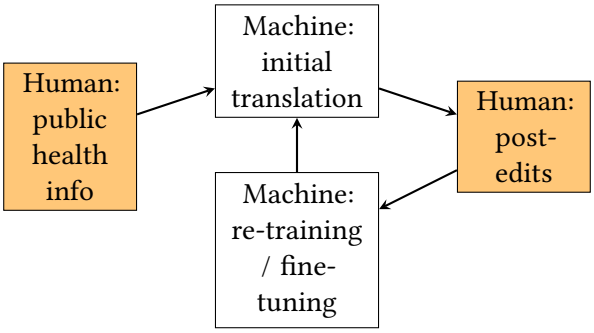

Figure 1 Proposed human-machine workflow for rapid translation of public health information.

In essence, the idea is that, given a reliable source of public health information, we can rapidly translate that information into many languages by making a first pass at translation with machine models, followed by post-editing of the outputs by bilingual speakers of source and target language. This is a conceptually simple workflow, illustrated in Figure 1, with a strong emphasis on the *human-in-the-loop*. Note that, more accurately, we deem it *machine-in-the-loop* with humans at either end, and the ability to update the model based on human correction.[1] The updated model can then play a role in future translations for the language pair, potentially producing better outputs and making the human post-editors' task easier and less time-consuming.

In this introductory section we consider Figure 1 in overview, firstly examining the human elements. The source of expertly-crafted public health information could be any of many suitable organisations. We consider the example of the World Health Organisation in section 2. The human post-editors of machine translations is a topic we look at more closely in section 3. In essence we consider this a potential task for crowdsourcing, harnessing the urge witnessed among many in the COVID-19 pandemic to aid others and contribute in whatever way possible during a health emergency.

The 'machine' in Figure 1 represents a machine translation model for a given language pair. In section 4 we discuss two of the possibilities for accessing such a model. The first is the from-scratch option familiar to machine translation researchers: making use of any relevant existing resources to train models. In low-resource contexts, training data are by definition scarce and unlikely to be in-domain for public health documents. We show how well we can do with available training data for translating from English into Swahili, make use of published techniques for low-resource contexts (namely, triangular methods and data augmentation), and conclude that further collection of in-domain data is urgently needed if machine translation is to play a supporting role in future health emergencies.

[1] Whether model updating should be in the form of re-training or fine-tuning is an open question; we hope that researchers will try fine-tuning in the first instance, for environmental reasons (Strubell, Ganesh & McCallum 2019).

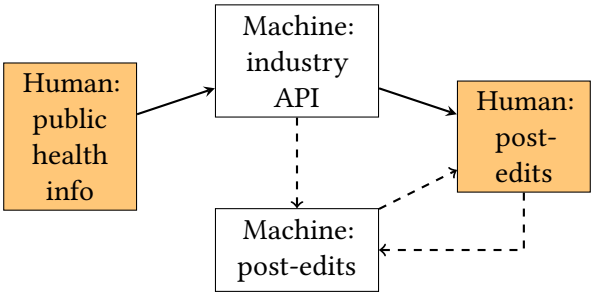

Figure 2 Proposed human-machine workflow for rapid translation of public health information – alternative scenario in which industry translation APIs play a role.

The second scenario is one in which – rather than attempting to outdo what the technology industry can do with machine translation – we start by obtaining automatic translations from an industry API ('application programming interface', a software intermediary – as opposed to a 'user interface'), and ask bilingual speakers to correct them for fluency and fidelity to the original. This approach, making use of a translation API in a low or zero-cost fashion, is a set up that is ready for immediate implementation – in contrast to a long-running and potentially expensive data collection project which might not result in models which are better than industry ones. This workflow is illustrated in Figure 2, in which there is a definite flow from source language to target language via an API, and there is the possibility to train post-editing technology off the human post-edits – which could then be inserted into the workflow after calls to the API (the path of the dashed lines).

However, this second scenario is only applicable to languages served by industry APIs; of course there are many languages not in this category. For these languages the challenge remains of building research models which are good enough to give human editors a hypothesised translation to work with and save them time, as opposed to humans translating from scratch. Discovering whereabouts the threshold for 'good enough' is, in this context, is a matter for future work. Note that for these languages, the data collection proposals we make in this paper are particularly relevant.

2 The COVID-19 Pandemic and Sources of Public Health Information

The COVID-19 pandemic of 2020 onwards[2] highlighted the need for global cooperation and the benefit of information and resource sharing across national boundaries. The President of the European Commission, Ursula von der Leyen, summarised the point with clarity: 'No one is safe until everyone is safe.'[3]

[2] The virus was first identified in December 2019, but it was not until January 2020 that The World Health Organization deemed it a Public Health Emergency of International Concern, and subsequently declared a pandemic on 11 March 2020.

[3] Twitter, https://twitter.com/vonderleyen/status/1301591418327183363, 3 September 2020. The words also featured in her speech to the State of the Union conference of the European University Institute, https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_21_2284, 6 May 2021.

It became apparent that, in a scenario which was unprecedented for so many, and with an unknown virus about which new information accumulated rapidly, the delivery of clear, high quality public health recommendations from reliable sources would be extremely beneficial for the general population in combatting COVID-19. It seems uncontroversial to assert that people will best understand these public health recommendations if they are communicated in (one of) their first or dominant language(s).

However, for example: the primary world authority on public health, the World Health Organisation (WHO), publishes advice relating to COVID-19 on their website in six of the most widely used languages[4] – Arabic, English, French, Mandarin Chinese, Russian and Spanish. This is presumably a resource-intensive effort in itself but it means that the majority of the world's several thousand languages,[5] and the majority of people, have not been directly served by the WHO during the COVID-19 pandemic. Besides, for many, their contact with public health information about COVID-19 was through government services in their own countries. This is itself not a straightforward process since it is fraught with political complications and competing socio-economic priorities: there was the case of Tanzania where for a period the government denied the seriousness of the situation, refused the vaccine, and stopped publishing COVID-19 data (Makoni 2021, Buguzi 2021).[6] Many also relied on social media for health information, and an 'infodemic' of public health *mis*-information arose (Cinelli, Quattrociocchi, Galeazzi, Valensise, Brugnoli, Schmidt, Zola, Zollo & Scala 2020, Dash, Parray, De Freitas, Mithu, Rahman, Ramasamy & Pandya 2021).[7]

It is true that the six languages served by the WHO are learned as a second language by billions around the world, and that many can understand a second language well enough to follow public health direction. Nevertheless it seems uncontroversial to assert that there is a greater risk of misunderstanding if attempting to understand health advice in a second language. This in turn can mean that information is not passed on reliably within communities, and that community-specific planning is hard. Without the appropriate lexicon or phrasing to talk about recommended health measures, rumour, misinformation and fear can spread. Moreover, a confluence of social factors means that a region with fewer resources for government funding of public health agencies, where citizens are often multilingual but indigenous languages are not always the languages of government or education (Bunyi & Schroeder 2017), and where there are a great number of quite different native languages is sub-Saharan Africa. Therefore in this work we are motivated to

[4] Correct at time of writing, February 2022.

[5] Estimates vary, and it depends how you define a 'language', but for instance the Glottolog identifies 7613 'spoken L1 languages (i.e. spoken languages traditionally used by a community of speakers as their first language)' (Hammarström, Forkel, Haspelmath & Bank 2021). The number is declining as languages go extinct at an alarming rate (Caines, Bentz, Alikaniotis, Katushemererwe & Buttery 2016).

[6] Note also another side to the story (Mfinanga, Mnyambwa, Minja, Ntinginya, Ngadaya, Makani & Makubi 2021).

[7] WHO: Managing the COVID-19 infodemic, https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation, 23 September 2020.

investigate the quality of current machine learning technology for the translation of public health information into African languages, but of course, the proposals we make in this paper could apply to any language pair.

We think that the COVID-19 pandemic has demonstrated the need for rapid translation of public health information to aid in the dissemination of high quality, evidence-based advice from a central authority such as the WHO. The WHO would in our view be a good choice as an information source for translation into multiple target languages, but it is apparent that any reliable information source could be used.

3 Crowdsourcing Corrections of Machine Translation

Crowdsourcing is widely used for quantitative and qualitative tasks involving natural language data. Post-editing machine translations is one such task which can be construed as follows: please consider source sentence x and the hypothesised translation y, and rewrite or edit y such that a new version y' is appropriately fluent and faithful to x (Parton, Habash, McKeown, Iglesias & de Gispert 2012, Chatterjee, Weller, Negri & Turchi 2015). The human post-editors needed for the task of correcting machine translated public health information should have a good working knowledge of both source and target languages. Exactly what linguistic proficiency level they should have in both languages could be a matter for future investigation, but we do not assume that they need to be native speakers of either source or target language, though of course that would help. Rather, perhaps if they have at least an upper intermediate knowledge of both languages (level B2 to put it in terms of the CEFR[8]) that could be enough.

The other important desideratum is that the post-editors feel adequately motivated to perform the task accurately, since health information can be of vital importance. This motivation could be intrinsic, i.e. the post-editors volunteer to correct the translations out of a wish to contribute to societal defences against a current health emergency. This lies behind the internet meme where 'everyone' was suddenly a virologist at the start of the COVID-19 pandemic: this may have been counter-productive in many ways, but it demonstrates that in crisis situations people want to try and help. This urge could be harnessed in positive ways by making use of people's multilingualism, but trying our best to keep the task time-efficient by starting the translation process through machine models. Indeed, the spark for our project was noticing ongoing community mobilisation efforts to translate health information from English or French into other languages – in this case African languages – and we aimed to explore if there would be value in using machine translation to cut down the time spent on such tasks in future mobilisation efforts. Hence we asked to use those translations as our test set, in order to investigate how close we could get to texts of similar quality.

However, we recognise that intrinsic motivation may wane as a health crisis goes on, with evidence of population fatigue also observed during the COVID-19

[8] Council of Europe, https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions, accessed 14 November 2021.

pandemic (Bartusevičius, Bor, Jørgensen & Petersen 2021, Lindholt, Jørgensen, Bor & Petersen 2021, Petersen, Rasmussen, Lindholt & Jorgensen 2021). Thus it would be important to explore opportunities to build ongoing capacity for improving access to health information through translation efforts. This would also involve mechanisms for sustaining, financing and learning from community mobilisation efforts as well as building partnerships between such efforts and institutionalised efforts. This is especially important given that future global health disruptions such as pandemics are anticipated. One issue here is that the response can be more rapid with voluntary efforts – and rapidity is a *very* important consideration in this context – rather than one delayed by the time taken to recruit appropriately-skilled employees and negotiate legal and contractual matters.

These are the issues we wish to put forward on the matter of human post-editors. Machine translation researchers can collect and evaluate post-editing data, but wider questions about workforce motivation and financial models would best be tackled with experts from other disciplines in the social sciences.

4 Machine Translation of Public Health Information

We identify two options for achieving the initial machine translations introduced in section 1: making use of industrial translation APIs, if available for the languages in question (Figure 2), or developing new models for languages outside that set (Figure 1). The question in both cases is whether the machine translations obtained are good enough to save humans time as a post-editing task as opposed to translating from scratch. For the first case (translation APIs) we propose that the translations are good enough – for English into Swahili at least – through a human evaluation study. For the latter case (new models), we outline the kind of data collection efforts which need to be carried out in order to make good-enough machine translation feasible.

We consider both scenarios through English-Swahili translation, choosing English because it is the best-resourced in NLP terms, and because it is among the WHO's website languages for COVID-19 information. Moreover it is likely to be well served in health emergencies of the forseeable future. Since we wish to translate into African languages, we have a low-resource situation in natural language processing terms – that is, corpora and pre-trained models are scarce compared to well-resourced languages, especially English. We choose Swahili as the target language because (a) it is widely spoken across sub-Saharan Africa, being an official language in Kenya, Rwanda and Tanzania, among other countries, and a lingua franca elsewhere; (b) some NLP training resources do exist, so this is a low-resource machine translation situation, rather than 'zero-shot' (Firat, Sankaran, Al-onaizan, Yarman Vural & Cho 2016); (c) it commonly features in industry translation APIs, such as Google Translate, and so we can compare the two machine translation scenarios we have set up.

In overview our experiments may be summarised as follows:

i. Use of a large but *out-of-domain* training corpus of religious texts for English-Swahili machine translation;

ii. Use of triangular machine translation methods: augmentation of training data with target languages linguistically related to Swahili, namely Igbo and Yoruba[9] which are also in the corpus of religious texts, based on the notion of transfer learning through a shared vocabulary (Lakew, Erofeeva, Negri, Federico & Turchi 2018);

iii. Use of an *in-domain* corpus of biomedical texts which is small: with augmentation of the training data by automatically generating new texts with language models;

iv. Data augmentation by obtaining machine translations of English biomedical texts into Swahili from Google Translate, and incorporating this 'silver-standard' data into the training set.

In all cases we are attempting to mitigate the shortage of in-domain biomedical training data, either by training on larger out-of-domain corpora (experiment 1) or data augmentation of various kinds (experiments 2 to 4).

The experiments are reported below in the order given above. In all cases we use the JoeyNMT toolkit for neural machine translation (Kreutzer, Bastings & Riezler 2019), making use of an example code notebook for training a Transformer model published by the Masakhane research community on their GitHub page.[10] Aside from adjusting the commands to point to our datasets, we largely use the notebook unchanged.[11]

4.1 Evaluation

We report results with CHRF (out of 1.0) – the character n-gram F-score which has been shown to be independent of tokenisation choices, correlates well with human evaluation, and can give credit for correct lexical choices even if morphological form is incorrect (Popović 2015). This makes the metric more language-independent than commonly used word n-gram metrics such as BLEU (Papineni, Roukos, Ward & Zhu 2002), since Swahili for instance is morphologically more complex than languages such as English.

We use the default CHRF metric available in JoeyNMT, which in turn comes from the SACREBLEU software repository (Post 2018). We use the default SACREBLEU settings for evaluation (notably a character n-gram order of 6 and case sensitivity).

[9] These are all Niger-Congo languages belonging to the Volta-Congo branch, but even if this indicates typological relatedness nevertheless the three languages are quite distant from each other in reality: Yoruba and Igbo for instance developed from the same language but are no longer mutually comprehensible. However, it is acknowledged that there is a substantial amount of shared vocabulary among the three languages and therefore it is not unreasonable to expect that there can be gains in translation performance through resource combination.

[10] https://github.com/masakhane-io/masakhane-mt/blob/master/starter_notebook_from_English_training.ipynb

[11] For the sake of reproducibility: configuration details include use of the Adam optimizer, a learning rate of 0.0003, dropout of 0.3, and a beam search of 5 for decoding. We report results on all experiments after training for 30 epochs. Training with more epochs and tuning the given hyperparameters could be attempted in future research.

4.2 Data

For **training and development**, we obtained the available English-Swahili parallel corpora we knew of: namely JW300 (Agić & Vulić 2019), TICO-19 (Anastasopoulos, Cattelan, Dou, Federico, Federmann, Genzel, Guzmán, Hu, Hughes, Koehn, Lazar, Lewis, Neubig, Niu, Öktem, Paquin, Tang & Tur 2020), and the UFAL Medical Corpus.[12]

JW300 contains texts covering 343 languages, sourced from the Jehovah's Witnesses website jw.org. The parallel English-Swahili section of the corpus features about one million sentences, which is sizable, but evidently the dataset is out-of-domain with regard to public health information. At the time of conducting our experiments, the dataset was still available from the OPUS site,[13] but subsequently has been removed for unknown reasons – meaning that a potentially valuable source of multilingual training data is no longer available. In our second experiment, we add training data from JW300 in other languages from the Volta-Congo family: Igbo and Yoruba, each of which has about 0.5M parallel sentences with English in the corpus.

TICO-19, the Translation Initiative for COVID-19, is another multilingual dataset – a collaborative effort to publish a dataset featuring scientific and Wikipedia articles about COVID-19. There are approximately 3K sentences in the training set, a modest size for neural machine translation, but of vital importance since it is in-domain and contains vocabulary and phrases particular to the COVID-19 pandemic.

The **UFAL Medical Corpus** is a collection of different corpora, again in the biomedical domain, but without Swahili texts. We make use of two EMEA corpora from the UFAL dataset: the approximately 0.5M sentences from OpenSubtitles (Lison & Tiedemann 2016) and the 'new crawl' of approximately 0.6M sentences from European Medicines Agency documents (Tiedemann 2012). We passed the English side of the German-English texts to the Google Translate API and requested Swahili translations as a way to obtain silver-standard training data.

Our **test set** comes from a crowdsourced translation initiative undertaken at the start of the COVID-19 pandemic, in which English or French health advice was translated into twenty African languages by native speakers. Seventy-one sentences were translated from English into Swahili in this way, featuring questions and answers designed to be shown online in a flashcard or infographic style. We recognise that this is a relatively small test set in the context of modern practice in machine translation, and this should be kept in mind when inspecting our experimental results below. We emphasise that this test set arises from a volunteer effort led by the Engage Africa Foundation to translate COVID-19 information across as many languages as possible, as rapidly as possible.[14] The motivation for this translation exercise was real and urgent and we are grateful to have been granted access for research purposes. Given that volunteer translation efforts are usually one-off mo-

[12] https://ufal.mff.cuni.cz/ufal_medical_corpus

[13] http://opus.nlpl.eu/JW300.php, attempted access 14 November 2021.

[14] For background, see this University of Cambridge news article: https://www.cam.ac.uk/stories/Translations-for-Africa.

bilisation activities in response to specific crises, they have the tendency to wane over time. It could be valuable for health organisations to proactively build the capacity for sustainable translation efforts through partnerships with community networks, in anticipation of emergency situations such as pandemics, which are expected to reoccur. An example from the test set is given below:

- What are the symptoms of COVID-19?
  - The most common symptoms of COVID-19 are fever, tiredness, and a dry cough.
  - Some patients may have aches, pains, nasal congestion, a runny nose, sore throat or diarrhea.

This is the test set we refer to as **PHI** (Public Health Information) in the results reported below. We also report model evaluation on the JW300 test set in order to put the PHI results in context, showing how domain effects in the training data are manifested in test results.

4.3 Methods & results

In this subsection we detail the methods and report the results of the experiments enumerated towards the end of section 4, in the same order as set out in that list.

4.3.1 Out-of-domain training data

At first we use the English-Swahili texts available from JW300 and report chrF scores on the same-domain JW300 Swahili test set, alongside those for our PHI test set (Table 1). It is apparent that translation performance on the JW300 test set is much better than it is on public health information. This is unsurprising, given that the training dataset, albeit a large one, is from a quite different domain.

Test set	chrF
JW300:sw	0.67
PHI:sw	0.37

Table 1 Experiment 1 – out-of-domain training with English-Swahili texts from JW300, evaluated on both the JW300 test set and PHI sentences.

4.3.2 Out-of-domain training with related languages

So-called 'triangular machine translation' methods have grown in popularity in recent years, and triangular MT was one of the tasks at the most recent Conference on Machine Translation (Akhbardeh, Arkhangorodsky, Biesialska, Bojar, Chatterjee, Chaudhary, Costa-jussá, na Bonet, Fan, Federmann, Freitag, Graham, Grundkiewicz,

Haddow, Harter, Heafield, Homan, Huck, Amponsah-Kaakyire, Kasai, Khashabi, Knight, Kocmi, Koehn, Lourie, Monz, Morishita, Nagata, Nagesh, Nakazawa, Negri, Pal, Tapo, Turchi, Vydrin & Zampieri 2021). One possible approach is to augment a training set with data from related languages (Pascal, Assobjio & Assiene 2021). In our case we made use of two other Volta-Congo languages, Igbo and Yoruba which are also found in JW300 in parallel texts with English, thereby approximately doubling the size of the training set.

Test set	en-sw/ig	en-sw/yo
JW300:sw	0.36	0.35
PHI:sw	0.30	0.33

Table 2 Experiment 2 – an English-Swahili machine translation model with additional training JW300 texts from Igbo and Yoruba, evaluated on both the JW300 test set and PHI sentences (chrF scores).

These results are worse than training on English-Swahili texts alone. It may be that the languages are not similar enough for the benefits of this approach to be realised, as Swahili is in the Bantoid branch of the Volta-Niger family, whereas Igbo and Yoruba are in the Igboid and Defoid branches respectively (Hammarström et al. 2021). Given time, we would also investigate other triangular MT methods reported to have worked well, such as transfer learning (Zoph, Yuret, May & Knight 2016, Kim, Petrov, Petrushkov, Khadivi & Ney 2019) or reference based MT (Li, Zhao, Wang, Utiyama & Sumita 2020).

4.3.3 In-domain training data augmented through text generation

Next we turn to the much smaller TICO-19 corpus which is *in-domain* for our PHI texts, but at only 3K sentences not big enough on its own for model training. Thus we augmented the training data through automatic generation of new texts with the open-source `nlpaug` toolkit.[15] We used the `ContextualWordEmbsAug` option with BERT (Devlin, Chang, Lee & Toutanova 2019) to double the size of the TICO-19 training set. We added the 6K real and artificial English-Swahili sentences derived from the TICO-19 corpus to the JW300 English-Swahili training set.

Test set	chrF
JW300:sw	0.51
PHI:sw	0.40

Table 3 Experiment 3 – partly in-domain training with English-Swahili texts from TICO-19 as well as JW300, artificially doubled in size through language model augmentation, evaluated on both the JW300 test set and PHI sentences.

[15] https://github.com/makcedward/nlpaug

The addition of a small amount of in-domain training data improves machine translation on the PHI test set, albeit causing a deterioration in performance on the JW300 test set compared to Table 1. This is unsurprising, because the domain of the training data has started to drift or become diluted by COVID-19 texts besides the Jehovah's Witnesses texts.

4.3.4 Silver-standard in-domain training data

Our final experiment in this section involved obtaining Swahili translations of English biomedical texts from the UFAL Medical Corpus – approximately 1.1M sentences from the OpenSubtitles and EMEA 'new crawl' datasets – using the Google Translate API.[16] We treated the resulting data as silver-standard and added it to the TICO-19 training data in place of the JW300 religious texts. Now, results on the PHI test set are reported alongside those for a held-out test set of 20% of the Google translations.

Test set	CHRF
UFAL:sw	0.56
PHI:sw	0.28

Table 4 Experiment 4 – in-domain training with English-Swahili texts from TICO-19 and automatic Google translations of English-Swahili texts from the UFAL Medical Corpus, evaluated on 20% of the UFAL data and the PHI sentences.

Here we see that the silver-standard training data does not help with model performance on our public health information test set, compared to Table 3. Instead, we think that targeted collection of in-domain, gold-standard training data is needed.

4.4 Discussion

Our overall conclusion from the above set of experiments is that the most promising approach for training machine translation models to translate public health information is to obtain in-domain training data and augment it through additional generation of artificial sentences with large language models. Specifically, we doubled the size of the TICO-19 training data for Swahili through artificial augmentation. This is an oversampling technique which is costly in terms of time and the environment, so we stopped having doubled the TICO-19 data. Possible future work involves further augmentation and investigation of the effects on model performance.

In any case, the outcome of our experiments reinforces the well-known point that in-domain training data are vital, since generalisation to unseen domains is hard for machine translation models (Müller, Rios & Sennrich 2020). Trying to improve in

[16] We wish to thank Google for waiving the usual charges for this translation request, on the basis that it was a one-off research task. The cost would have been more than 1000 USD, making this method a not inexpensive one.

one domain may lead to deterioration on another one, especially when the domains are very different, as shown in our results on the religious JW300 texts and the biomedical PHI ones (Table 1 vs Table 3).

Therefore we propose that the ongoing collection of in-domain training data is needed in order to continue improving machine translations of COVID-19 information, and to prepare for future health emergencies. It may be that generic biomedical data will help, as found in the UFAL Medical Corpus: but these need to be translated to a gold-standard by humans, rather than the automatic silver-standard approach we trialled with Google Translate.

It certainly seems to be the case that COVID-19 specific data helped: our best results on the PHI test set came with the TICO-19 dataset. This is a clear and intuitive finding, but it does mean that data collection needs to adapt quickly to the topic at hand. Future health emergencies are by their nature unpredictable, and therefore likely to feature new vocabulary and phrases, or a dramatic boost in usage for existing ones – as has been seen during the COVID-19 pandemic (e.g. the terms *social distancing*, *hand hygiene*, *face covering*, *lockdown*, *anti-vaxxer*, and *COVID-19* itself).

However, it may be that there is a core vocabulary that remains useful across different situations – for instance, words and phrases commonly used regarding public health (e.g. *do*, *don't*, *if you have symptoms*, etc.). This suggests that further data collection efforts along the lines of the CMU English-Haitian Creole dataset could be useful (Lewis 2010). Collection of the dataset was prompted by the magnitude 7.0 Haitian earthquake of 2010, and contains medical domain phrases and sentences translated into Haitian Creole. For TICO-19, the authors sampled the corpus for English conversational phrases containing COVID-19-related keywords, meaning that 140 sentences were retrieved and transferred from one crisis scenario to another (Anastasopoulos et al. 2020).

In the appendix of the TICO-19 paper, the authors present translation terminology lists relating to COVID-19 and public health obtained from Facebook and Google. This again suggests a way to build public health machine translation technology, perhaps with statistical phrase-based approaches which allow for more explicit encoding of n-gram equivalents and weights, since public health information is a relatively narrow domain for which neural MT is a blunt instrument.

The Masakhane research project offers a possible model for data collection which is driven by bottom-up researcher motivation and resourceful collation of domain-specific materials (∀ 2020). In the paper the authors describe the benefits of the participatory research model for their work on African NLP – combining many people's interests and energy to work towards highly-impactful and much-needed diversification of NLP work away from mainly English and a few other languages. This includes data collection efforts for machine translation. Besides the example notebook for JW300 data we use in this paper, there are moves to incorporate other multilingual data such as the Tatoeba Challenge corpus,[17] and those datasets already included in the 'community library' offered by HuggingFace (Lhoest, Villanova del

[17] https://github.com/Helsinki-NLP/Tatoeba-Challenge

Moral, Jernite, Thakur, von Platen, Patil, Chaumond, Drame, Plu, Tunstall, Davison, Šaško, Chhablani, Malik, Brandeis, Le Scao, Sanh, Xu, Patry, McMillan-Major, Schmid, Gugger, Delangue, Matussière, Debut, Bekman, Cistac, Goehringer, Mustar, Lagunas, Rush & Wolf 2021), such as the FLORES-101 evaluation corpus (Goyal, Gao, Chaudhary, Chen, Wenzek, Ju, Krishnan, Ranzato, Guzmán & Fan 2021). New multilingual datasets in the biomedical domain could also slot into this model of open availability, ease of use, and participatory research.

5 Industry MT with Human Post-Editing

Finally, we investigated whether machine translation could be of use in a human-machine workflow, or would the translations be so bad as to be of no help at all? We decided to measure this by taking the best machine translations we could obtain for our English-Swahili test set of public health advice regarding COVID-19, using Google Translate. chrF for these new translations was 0.485 (compared to the best of 0.40 we achieved in section 4.3.3).

We asked volunteer English-Swahili speakers to review the original English sentences, the machine translations in Swahili, and rate the latter for accuracy, fluency and harmfulness (could the translated advice potentially cause harm?). We also asked the volunteers to produce an appropriate Swahili translation of the English source sentence, either by correcting the machine translation as it is, or starting again from scratch. Volunteers were recruited from the Masakhane research community and via social media.

We had 24 responses in total, of which 16 reported themselves to be native speakers of Swahili. The judgements of non-native speakers are not uninteresting, but since we prioritise the accuracy of health information in the target language, we decided to examine the responses of these 16 participants primarily. Each participant looked at a random selection of 24 sentences, and a control sentence for quality assurance.[18] As a result each of our 71 sentences were judged at least twice, with a median of 4 judgements per sentence and a maximum of 6.

Overall, the participants rewrote a translation from scratch twice only. However, 19 of the 71 translated sentences were judged to be harmful by at least one respondent, emphasising the importance of human correction of machine translation in this context. The relation between translation quality and amount of post-editing required is depicted in Figure 3: here we produce an overall quality score by summing fluency and accuracy scores (each with a maximum of 2), providing the translation was not judged to be harmful – in which case it would be assigned a score of zero. It is apparent that a large number of translations needed no correction at all (the jittered spread of data points at 4 on the x-axis and 0 on the y-axis), thus an important time-saver for human translators. On most occasions (93%), the human corrections amounted to less than half the translated sentence.

We also show how the quality of machine translations, measured by chrF per sentence on the x-axis, relates to the amount of post-editing needed: the difference

[18] We chose a very short sentence from the test set for which there is a single obvious translation remaining faithful to the original: *Monitor your symptoms* → *Fuatilia dalili zako.*

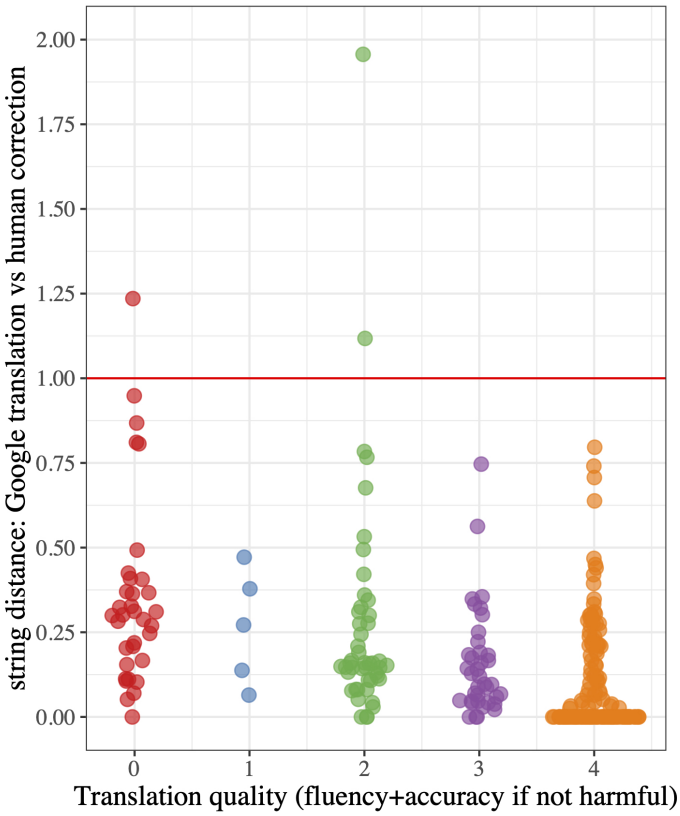

Figure 3 Translation quality (fluency plus accuracy scores (max 4), or set to zero if judged to be harmful) is on the x-axis. String difference between the Google translation and human correction is on the y-axis (number of characters, divided by number of characters in the Google translation). For instance, a difference of 1.0 indicates that as many new characters were introduced during correction as there were in the Google translation, and a difference of 0 indicates that the Google translation was fine. For visibility of mass, datapoints are randomly jittered around the vertical on the x-axis, and partly transparent to avoid over-plotting.

in number of characters between machine and corrected translations on the y-axis (Figure 4). We find that many sentences (30.5%) do not need editing, regardless of their chrF score, but that the translations with lower scores do tend to be corrected by adding additional text.

Examples of machine errors we observed from this post-editing exercise included the phrase *do not feel unwell* being transformed to *do not feel well* in Swahili – clearly a problematic change in polarity – and advice to *block communication* rather than *avoid contact*, mis-translated advice which carries potential social harm if followed. Note that by capturing the post-edits on machine translation, we can either update a model of our own making (Figure 1), or develop a post-editing model which would fit into the workflow after translations had been obtained from an API (Figure 2).

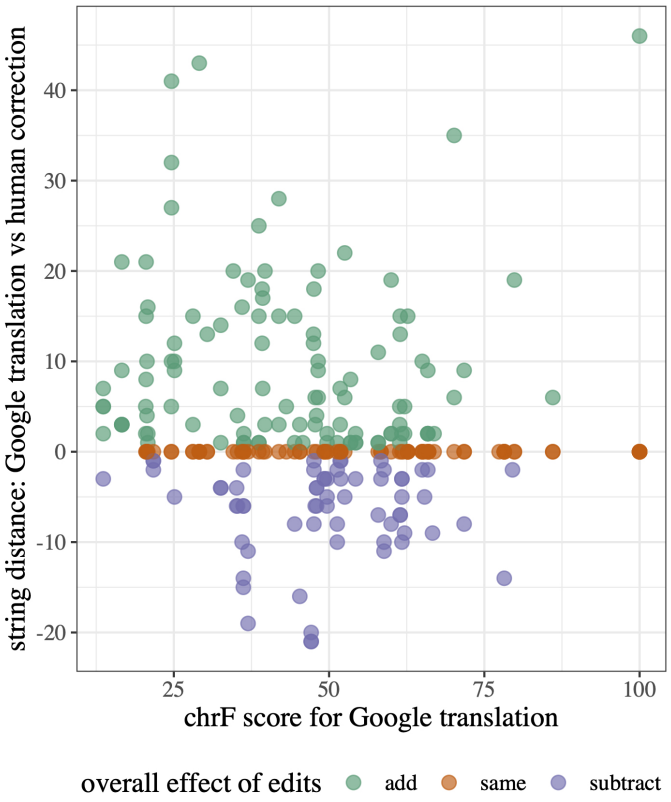

Figure 4 chrF scores for Google translations on the x-axis, string distance between Google translations and human corrections on the y-axis (number of characters). Datapoints are classified into 3 types: depending whether the Google translations have been added to, subtracted from, or are unchanged. Datapoints are partly transparent to avoid over-plotting.

6 Conclusion

We have proposed a joint human-machine approach for translation of public health information: either developing models from scratch with available training data, or harnessing the strides made by industrial machine translation teams by using industrial translation APIs as the starting point. A network of bilingual speakers should be assembled for the correction of machine translations, so that potentially harmful, erroneous advice is not disseminated. This is a lightweight architecture which could be set up at short notice in case of future pandemics; by including humans in the loop, it is more robust to the problem of dealing with the new concepts and new vocabulary which will accompany new emergencies.

Our initial experiments on English-Swahili suggest that more in-domain training data is needed for translation of public health information, but that offering good quality automatic translations to human editors can be both a time-saver and a way to obtain more training data to further improve the machine models in the

pipeline. Note that in future work other domain-adaptation translation techniques may be attempted (Saunders 2021), and that a larger test set is needed for more robust evaluation of our translation models. Our intention in this paper was to present preliminary experimental results and argue for closer integration of human and machine translation activities to prepare for future health crises.

In principle our proposal applies to any language pair, but it remains a matter for future work whether our conclusions apply equally strongly to high-resource language pairs such as English-German, or languages with even fewer resources than Swahili, more closely-related language pairs, or target languages in other language families.

Ethical considerations

For the most part, the data referred to in this paper came from pre-existing datasets. The crowdsourced test set arose from a voluntary community effort organised by researchers and grassroots organisations. We received ethics approval from our Institutional Review Board for the human evaluation on Google translations which we describe in section 5. The main risk with machine translation of public health information is that inaccurate translations might cause the spread of harmful advice: in this paper we make it clear that MT should not be deployed without a human post-editing step to ensure accuracy and fidelity to the original text. We do not release the Google translations or post-edit corrections with this paper, but propose that such a data release would be valuable in future, as it would help accelerate the human-machine process of translating public health information. Any data release should be handled with care, of course, for instance with an accompanying data statement (Bender & Friedman 2018).

Acknowledgements

This work was supported by the Global Challenges Research Fund, University of Cambridge, which was funded by Research England. We are grateful for the existence and support of the Cambridge Global Challenges Strategic Research Initiative[19] and the Cambridge Language Sciences Interdisciplinary Research Centre,[20] and in particular thank Dr Sara Serradas O'Holleran and Ms Jane Walsh for their encouragement and advice. We thank Professor Paula Buttery and Professor Bill Byrne for giving up their time to mentor and guide our research in advance of and throughout the project. We received warm encouragement and welcome advice from members of the Masakhane research community,[21] and especially wish to acknowledge the help of Jade Abbott, Colin Leong and Julia Kreutzer. We thank professional services staff in the Computer Laboratory who supported this work, and finally, we thank the NVIDIA Corporation for the donation of the Titan X Pascal GPU used in this research. This work would not have been possible without the

[19] https://www.gci.cam.ac.uk/
[20] https://www.languagesciences.cam.ac.uk/
[21] https://www.masakhane.io/

rapid translation response of COVID-19 information by members and volunteers from the Engage Africa Foundation[22] – **Core team**: Ebele Mọgọ, Dara Oloyede, Tola Olufemi, Aghedo Osazemen, Cyril Tata. **Translation project-specific volunteers**: Susie Monyo, Chika Jones, Joanna Mogo, Tunji Sarumi, Kehinde Akinsola, Esther Olawuyi, Ibrahim Muhammad Shamsuddin, Roselyne Orji, Gigah Eleanor Visas, Annet Akai, Harriet Brefo-Mensah, Petrina Akor, Christian Nkanga, Adaobi Ugwu, Ngozi Osuji, Medadi Ssentanda, Abdulrahman Atta, Morenike Akinyemi, Mashkur Isa, Maximillian Petzold, Tope Akinsipe, Diana Adu-Gyimah, Abas Ibekwe, Usaini Sani Adamu, Mariamawit Yeshak, Divya Bhagtani, Best Agofure, Tania Bishola, Jedah Mayberry, Aimable Uwimana, Nosiphiwo Lawrence, Vanessa Lum N. Sab, Patrick Kanampiu and Babelos Ltd.

References

Agić, Ž. & I. Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 3204–3210. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1310. https://aclanthology.org/P19-1310.

Akhbardeh, F., A. Arkhangorodsky, M. Biesialska, O. Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussá, C. E. na Bonet, A. Fan, C. Federmann, M. Freitag, Y. Graham, R. Grundkiewicz, B. Haddow, L. Harter, K. Heafield, C. Homan, M. Huck, K. Amponsah-Kaakyire, J. Kasai, D. Khashabi, K. Knight, T. Kocmi, P. Koehn, N. Lourie, C. Monz, M. Morishita, M. Nagata, A. Nagesh, T. Nakazawa, M. Negri, S. Pal, A. A. Tapo, M. Turchi, V. Vydrin & M. Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the sixth conference on machine translation (wmt*, http://www.statmt.org/wmt21/pdf/2021.wmt-1.1.pdf.

Anastasopoulos, A., A. Cattelan, Z. Dou, M. Federico, C. Federmann, D. Genzel, F. Guzmán, J. Hu, M. Hughes, P. Koehn, R. Lazar, W. Lewis, G. Neubig, M. Niu, A. Öktem, E. Paquin, G. Tang & S. Tur. 2020. TICO-19: the translation initiative for Covid-19. *CoRR* abs/2007.01788. https://arxiv.org/abs/2007.01788.

Bartusevičius, H., A. Bor, F. Jørgensen & M. B. Petersen. 2021. The psychological burden of the COVID-19 pandemic is associated with antisystemic attitudes and political violence. *Psychological Science* 32(9). 1391–1403. doi:10.1177/09567976211031847.

Bender, E. M. & B. Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6. 587–604. doi:10.1162/tacl_a_00041. https://aclanthology.org/Q18-1041.

Buguzi, S. 2021. Covid-19: Counting the cost of denial in Tanzania. *BMJ* 373. doi:10.1136/bmj.n1052. https://www.bmj.com/content/373/bmj.n1052.

Bunyi, G. & L. Schroeder. 2017. Bilingual education in Sub-Saharan Africa: Policies and practice. In O. García, A. M. Y. Lin & S. May (eds.), *Bilingual and multilingual*

[22] https://www.engageafricafoundation.org/

education, 311–328. Springer International Publishing. doi:10.1007/978-3-319-02258-1_13.

Caines, A., C. Bentz, D. Alikaniotis, F. Katushemererwe & P. Buttery. 2016. The Glottolog data explorer: Mapping the world's languages. In *Proceedings of vislr ii: Visualization as added value in the development, use and evaluation of language resources*, https://cainesap.shinyapps.io/langmap/.

Chatterjee, R., M. Weller, M. Negri & M. Turchi. 2015. Exploring the planet of the APEs: a comparative study of state-of-the-art methods for MT automatic post-editing. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, 156–161. Beijing, China: Association for Computational Linguistics. doi:10.3115/v1/P15-2026. https://aclanthology.org/P15-2026.

Cinelli, M., W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo & A. Scala. 2020. The COVID-19 social media infodemic. *Scientific Reports* 10. doi:10.1038/s41598-020-73510-5.

Dash, S., A. A. Parray, L. De Freitas, M. I. H. Mithu, M. M. Rahman, A. Ramasamy & A. K. Pandya. 2021. Combating the COVID-19 infodemic: a three-level approach for low and middle-income countries. *BMJ Global Health* 6(1). doi:10.1136/bmjgh-2020-004671. https://gh.bmj.com/content/6/1/e004671.

Devlin, J., M.-W. Chang, K. Lee & K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Firat, O., B. Sankaran, Y. Al-onaizan, F. T. Yarman Vural & K. Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 268–277. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1026. https://aclanthology.org/D16-1026.

∀, W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. Muhammad, S. Kabongo Kabenamualu, S. Osei, F. Sackey, R. A. Niyongabo, R. Macharm, P. Ogayo, O. Ahia, M. M. Berhe, M. Adeyemi, M. Mokgesi-Selinga, L. Okegbemi, L. Martinus, K. Tajudeen, K. Degila, K. Ogueji, K. Siminyu, J. Kreutzer, J. Webster, J. T. Ali, J. Abbott, I. Orife, I. Ezeani, I. A. Dangana, H. Kamper, H. Elsahar, G. Duru, G. Kioko, M. Espoir, E. van Biljon, D. Whitenack, C. Onyefuluchi, C. C. Emezue, B. F. P. Dossou, B. Sibanda, B. Bassey, A. Olabiyi, A. Ramkilowan, A. Öktem, A. Akinfaderin & A. Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the association for computational linguistics: Emnlp 2020*, 2144–2160. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.195. https://aclanthology.org/2020.findings-emnlp.195.

Goyal, N., C. Gao, V. Chaudhary, P. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán & A. Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR* abs/2106.03193. https://arxiv.org/abs/2106.03193.

Hammarström, H., R. Forkel, M. Haspelmath & S. Bank. 2021. *Glottolog 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org.

Kim, Y., P. Petrov, P. Petrushkov, S. Khadivi & H. Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, 866–876. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1080. https://aclanthology.org/D19-1080.

Kreutzer, J., J. Bastings & S. Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp): System demonstrations*, 109–114. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-3019. https://aclanthology.org/D19-3019.

Lakew, S. M., A. Erofeeva, M. Negri, M. Federico & M. Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. *CoRR* abs/1811.01137. http://arxiv.org/abs/1811.01137.

Lewis, W. 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*, Saint Raphaël, France: European Association for Machine Translation. https://aclanthology.org/2010.eamt-1.37.

Lhoest, Q., A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush & T. Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations*, 175–184. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-demo.21.

Li, Z., H. Zhao, R. Wang, M. Utiyama & E. Sumita. 2020. Reference language based unsupervised neural machine translation. In *Findings of the association for computational linguistics: Emnlp 2020*, 4151–4162. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.371. https://aclanthology.org/2020.findings-emnlp.371.

Lindholt, M. F., F. Jørgensen, A. Bor & M. B. Petersen. 2021. Public acceptance of covid-19 vaccines: cross-national evidence on levels and individual-level predictors using observational data. *BMJ Open* 11(6). doi:10.1136/bmjopen-2020-048172. https://bmjopen.bmj.com/content/11/6/e048172.

Lison, P. & J. Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 923–929. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1147.

Makoni, M. 2021. Tanzania refuses COVID-19 vaccines. *The Lancet* 397. doi:10.1016/S0140-6736(21)00362-7.

Mfinanga, S. G., N. P. Mnyambwa, D. T. Minja, N. E. Ntinginya, E. Ngadaya, J. Makani & A. N. Makubi. 2021. Tanzania's position on the COVID-19 pandemic. *The Lancet* 397. doi:10.1016/S0140-6736(21)00678-4.

Müller, M., A. Rios & R. Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th conference of the association for machine translation in the americas (volume 1: Research track)*, 151–164. Virtual: Association for Machine Translation in the Americas. https://aclanthology.org/2020.amta-research.14.

Papineni, K., S. Roukos, T. Ward & W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. https://aclanthology.org/P02-1040.

Parton, K., N. Habash, K. McKeown, G. Iglesias & A. de Gispert. 2012. Can automatic post-editing make MT more meaningful. In *Proceedings of the 16th annual conference of the european association for machine translation*, 111–118. Trento, Italy: European Association for Machine Translation. https://aclanthology.org/2012.eamt-1.34.

Pascal, T. N., B. Y. N. Assobjio & J. Assiene. 2021. On the use of linguistic similarities to improve neural machine translation for African languages. https://openreview.net/forum?id=Q5ZxoD2LqcI.

Petersen, M. B., M. S. Rasmussen, M. F. Lindholt & F. J. Jorgensen. 2021. Pandemic fatigue and populism: The development of pandemic fatigue during the COVID-19 pandemic and how it fuels political discontent across eight western democracies. *PsyArXiv* doi:10.31234/osf.io/y6wm4.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, 392–395. Lisbon, Portugal: Association for Computational Linguistics. doi:10.18653/v1/W15-3049. https://aclanthology.org/W15-3049.

Post, M. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation: Research papers*, 186–191. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/W18-6319. https://aclanthology.org/W18-6319.

Saunders, D. 2021. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *CoRR* abs/2104.06951. https://arxiv.org/abs/2104.06951.

Strubell, E., A. Ganesh & A. McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting*

of the association for computational linguistics, 3645–3650. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1355. https://aclanthology.org/P19-1355.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (eds.), *Proceedings of the eight international conference on language resources and evaluation (lrec'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).

Zoph, B., D. Yuret, J. May & K. Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1568–1575. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1163. https://aclanthology.org/D16-1163.

Dimitra Stasinou
Computer Laboratory
University of Cambridge
stasinoudim26@gmail.com

Theresa Biberauer
Theoretical and Applied Linguistics
University of Cambridge
mtb23@cam.ac.uk

Ebele Mọgọ
MRC Epidemiology Unit
School of Clinical Medicine
University of Cambridge
ebele.mogo@mrc-epid.cam.ac.uk

Andrew Caines
Computer Laboratory
University of Cambridge
andrew.caines@cl.cam.ac.uk