MACHINE TRANSLATION FOR AFRICAN LANGUAGES: COMMUNITY CREATION OF DATASETS AND MODELS IN UGANDA

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable machine translation systems are only available for a small proportion of the world's languages, the key limitation being a shortage of training and evaluation data. We provide a case study in the creation of such resources by NLP teams who are local to the communities in which these languages are spoken. A parallel text corpus, SALT, was created for five Ugandan languages (Luganda, Runyankore, Acholi, Lugbara and Ateso) and various methods were explored to train and evaluate translation models. The resulting models were found to be effective for practical translation applications, even for those languages with no previous NLP data available, achieving mean BLEU score of 26.2 for translations to English, and 19.9 from English. The SALT dataset and models described are publicly available.

1 INTRODUCTION

People who do not speak one of the world's major languages have difficulty accessing information and resources. Reliable machine translation systems are only available for a tiny proportion of the world's roughly 7,000 languages; while there is an increasing body of work on low-resource language modelling, the main problem is a lack of data. The most important data typically needed for machine translation is a corpora of parallel texts, in which there are pairs of matching sentences in source and target languages. For systems currently in practical use, the data often includes text from secondary sources, e.g. by scraping the web. This presents a number of problems for languages in the 'long tail':

- Most of the content on the web is in English, and the amount of usable data turned up by web-scraping typically falls off rapidly after the first few tens of languages. There are languages with millions of native speakers for which web scraping has not resulted in any usable training data; in many cases, even classifiers for language identification is not feasible Caswell et al. (2020).
- The use of secondary data means that models learned are subject to any biases brought about by that data. Text from the internet can be offensive and of poor quality; but alternatively the reliance on single sources, e.g. JW300 Agić & Vulić (2019), only cover a specific range of topics.
- There is no reliable data for evaluating model quality. BLEU scores, for example, could be misleading when evaluation data is low quality and subject to bias.

In this paper we describe an alternative, community-based method for creating open-source corpora, which provides a practical blueprint for assembling the necessary data resources for the several thousand 'long tail' languages.

This result of this work is a parallel text dataset in English and five Ugandan languages (Acholi, Ateso, Luganda, Lugbara, and Runyankole) that we call SALT (Sunbird African Language Translation dataset), as well as trained multilingual benchmark models for translating to and from English to these languages. The dataset is summarised and compared with previously existing resources for Ugandan languages in Table 1. The selected languages are a representation of the main language families that are spoken in Uganda. We emphasise the collection of multi-way parallel text, because

Table 1: Comparison between our dataset and existing resources for Ugandan languages.

Dataset	Language	Quantity	Quality
CCAligned	Luganda	14k	Mostly not Luganda: unusable.
MT560	Luganda Acholi Runyankore	225k 73k 50k	Professional translations, on a single topic (religion).
FLORES-101	Luganda	3k	Professional translations on generic topics.
SALT (this work)	Luganda Acholi Runyankore Lugbara Ateso	40k 25k 25k 25k 25k	Professional translations on locally relevant topics.

Table 2: Ugandan languages represented in the SALT dataset.

Language	Family	Speakers	Region
Acholi (ach)	Western Nilotic	1.47M	North
Ateso (teo)	Eastern Nilotic	1.57M	East
Luganda (lug)	Niger-Congo-Bantu	5.56M	Central
Lugbara (lgg)	Central Sudanic	1.10M	North
Runyankole (run)	Niger-Congo-Bantu	3.22M	West

languages in geographically neighbouring areas have similar characteristics, which a model can usefully learn from. While creating the SALT dataset, we placed particular emphasis on making this locally relevant, such that the corpus covers the topics and concepts that we would wish to use for translation in a practical setting.

We also discuss training multi-lingual translation models, including combining our dataset with existing corpora in these same languages. We achieve strong practical translation performance, evaluated both quantitatively by BLEU score and qualitatively by local experts. The dataset and models described in this work are openly available[1].

2 Related work

2.1 Low resourced language data collection

The need to collect quality language data without extensive digital data resources is critical, and so innovative ways of obtaining this kind of data have been an important component of recent African machine translation work. Masakhane (Nekoto et al., 2020) is a pan-African network of NLP enthusiasts which has collectively been looking at more participatory methods of collecting language data in sustainable ways – for example the collection of Setswana and Sepedi language data (Marivate et al., 2020). Funding bodies including Lacuna[2] have also emerged to specifically support the systematic collection and maintenance of language data on the African continent.

Several other smaller efforts existed previously, led by individual researchers trying to make these low resourced languages noticeable. This included creating pockets of specific translation data but stopping short of making them widely accessible. In Uganda, some efforts by Nabende (2018),

[1] https://github.com/SunbirdAI/ug-language-parallel-text-dataset
[2] https://lacunafund.org/

Table 3: Overview of the dataset creation process.

(1) Prompt		**(2) Seed sentence**		**(3) Translations**	
	→	"He spends a lot of money on pesticides."	→	Luganda	Asaasanya ssente nnyingi mu kugula eddagala ly'okufuuyira ebiwuka.
				Runyankole	Nateeka sente nyingiomu mibazi y'obukooko.
				Acholi	En obalo cente mapol me wilo pesticide.
				Ateso	Itosomaenenei ngesi ikapun luipu kotoma agwel ikee luka aisik ikur.
				Lugbara	Eri sende a'yu angiri aro osasaaniri ma dria.

Nandutu & Mwebaze (2020) and Omona & Groce (2021) to collect local datasets can be cited. As with these and in much of Africa, previous machine translation efforts have hinged on religious text datasets such as JW300 Agić & Vulić (2019) which have not captured the local contextual use of the languages.

2.2 Modeling approaches

Neural Machine Translation (NMT) had a significant boost with the introduction of the encoder-decoder Sutskever et al. (2014) network and specifically the implementation of this network with RNN and Transformer architectures Vaswani et al. (2017). Transformers particularly allowed models pre-trained on large monolingual datasets to be fine-tuned to smaller tasks with limited data for example translation with low-resource languages. A drawback of deep learning based approaches to NMT is the amount of data and compute required to train these models successfully, particularly affecting low-resource languages.

The use of pre-trained models greatly impacted low-resource language modeling, because appreciable results could start to be obtained for complex tasks like machine translation. For example the mBART architecture Tang et al. (2020) implemented as a pre-trained model fine-tuned on low-resource languages including Nepali, Sinhala and Gujarati shows marked improvement in performance as measured by the BLEU score.

Further improvements in modeling have been realised in multilingual models which solve the challenge of obtaining large monolingual datasets for training. Experiments done by Tang et al. (2021) already show that even with a dataset of 10K sentences, multilingual pre-trained models are able to achieve BLEU scores significantly above the baseline.

3 Dataset creation

Uganda has 43 local/native languages used by large sections of the population. We focus on Acholi, Ateso, Luganda, Lugbara, and Runyankole. Table 2 gives an overview of the languages used in this research. The selection was based on the number of speakers and also making sure that each region of the country is represented. The numbers represent the native speakers of the language based on the Uganda Bureau of Statistics (2016).

While designing our data collection process, we were mindful of various challenges and potential sources of bias. Achieving consistency can be challenging given that low-resource languages may lack norms such as standardised orthographies and have a scarcity of professional translators. Dialects can also mean that there is no standard way of arriving at the same translation: we noted that a given concept may be expressed differently between sub-regions, even though they all nominally belong to the same language group. Translators themselves have individual biases and idioms.

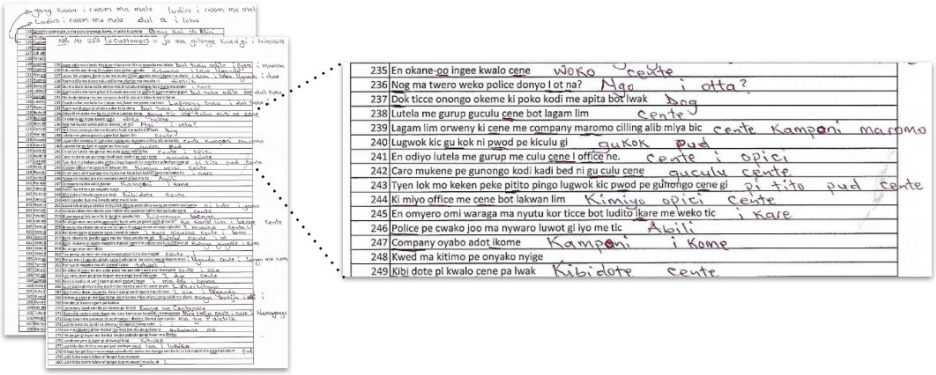

Figure 1: Validation checks and corrections were done using different methods to suit the individual needs of translators. This included annotating and scanning paper copies in the case of the elderly, who had authoritative language knowledge but difficulty with digital data entry.

All the Ugandan languages that we consider in this work have genderless pronouns: the translations of the sentences "he went to Kampala" and "she went to Kampala" are identical. During data collection we only asked for translations from English to Ugandan languages, however when using them to train translation systems in the other direction, this could be misleading.

Our methodology for dataset creation included recruitment and training of local translators (§3.1), identification and selection of source domains and "prompts" (§3.2), creation of English seed sentences (§3.3), and finally translation and validation (§3.4).

3.1 Translator recruitment

Given the diversity and geographical coverage of the selected languages, translators were recruited from different parts of the country. Translators included language experts, including professional translators as well as teachers and tutors from language schools. We placed this emphasis on experienced professionals given that the collection of parallel text corpora has a language-preservation effect; for languages with few digital data resources, any new resources carry extra weight as examples of how that language is used.

Translators were given further project-specific training to ensure consistency and quality of the translations. For example, emphasis was put on preserving meaning and context of the source in the target sentence, and to avoid "translation-ese" by giving natural translations of the seed sentences, as someone might commonly speak or write them.

The choice of data entry and translation tools can be a barrier for some language experts. Elderly translators had authoritative knowledge of the language but could find it challenging to work with digitised data entry methods such as spreadsheets. We therefore found it important to diversify the tools for translation, including not only a custom online translation management system[3] but also printing, handwriting and scanning as shown in Figure 1. Other translators had intermittent access to the internet and needed tools that made it possible to work offline, for which spreadsheets sent by email were a workable solution.

3.2 Prompt materials

Sentences in English to be translated were created using a process of prompts, for example headlines from online news sources from different regions in Uganda. We preferred not to use text from online sources directly for translation, as there can be issues both with copyright and with the formal tone of such text. A diversity of sources were used for prompts, including social media and more specific material such as legal articles[4], where although the prompt material may be technical and difficult to

[3] https://airnlp.herokuapp.com/
[4] https://ulii.org/

translate, can be an inspiration for natural sentences on those topics. Prompts were generated with a target balance of diverse topics, including but not limited to agriculture, health, and social issues.

3.3 GENERATION OF ENGLISH SEED SENTENCES.

Given a list of prompt English sentences, the data collection team was able to generate new sentences that comprised our seed English dataset. To generate a seed sentence from a prompt, one has to ensure the context is kept while rewriting the sentence to have a similar meaning to the original one. Table 3 depicts an example of the seed sentence generation process. A seed sentence is generated from a newspaper headline to reflect the gist of a comtemporary issue in Agriculture featured in the newspaper.

Given the list of prompts, these were converted to specific sentences to be translated, according to the following principles:

- Do not include personally identifying information such as names. Replace with fictitious details, or change the sentence to not include a name.
- Vary the level of formality, for example by looking at a news headline and asking 'how would I communicate this naturally to a family member?'
- Add additional variation by changing between first/third person point of view, active/passive voice, and so on.
- Check that the sentence is sufficiently different from the prompt.

The end result of this process was a list of the English sentences which would each be translated to all of the Ugandan languages.

3.4 TRANSLATION AND QUALITY CHECKS

Translation took place in a two step process to assure some level of quality. The English seed set of sentences was chunked out in batches of 100 or 150 sentences and given to each translator. Upon completing a batch the translator was responsible for passing this on to the nominated language expert who worked with the translator to correct any errors in grammar and translation. A batch was not cleared off until it had be signed off by the language expert and checked for consistency and completeness by the overall coordinator, at which point another batch was provided to the translator.

After the whole dataset was translated, a subset of 500 sentences in every set of 2000 translated sentences was sampled and reviewed by the team to ensure completeness in the translation of all five languages and correctness of the translations. Some judgement was required here, for example in terms of which words to leave as English and which to try to translate, such as when to attempt translation of names. A further set of checks on the whole dataset was done by the technical team packaging the data for online release and modeling, and a back and forth process of refinements was maintained until there was convergence of the dataset. Table 4 gives examples of the dataset with translations from English seed sentences to the five local languages.

4 BENCHMARK TRANSLATION MODELS FOR UGANDAN LANGUAGES

4.1 TRAINING WITH OUR DATASET

A key result of our work was creation of a parallel text dataset of key Ugandan languages. To evaluate how useful these are, we developed benchmark translation models based on this dataset. The scope of our experiments was restricted to translation between each of the Ugandan languages considered and English; we do not study here translation between Ugandan languages. We use multilingual translation models from OPUS-MT Tiedemann et al. (2020) as a starting point, utilising the MarianMT architecture[5]. We particularly focus on the `mul-en` (multiple languages to English) and `en-mul` (English to multiple languages) models as starting points for our work. Both of these models have support for Luganda, but none of the other Ugandan languages in our study.

[5] `https://huggingface.co/docs/transformers/model_doc/marian`

Table 4: Examples of parallel translated sentences.

English	Luganda	Runyankole	Acholi	Ateso	Lugbara
My uncle planted maize on four acres of land.	Kojja wange yasimba kasooli ku yiika nnya ez'ettaka.	Tatento abyire ebicoori aha hiika ina za itaka.	Nera opito anyogi I poto ma dongo angwen.	Abu mamaika iraa emudunga toma iekaasi iwongon.	Ma adroyi sa kaka nyaku eka su ma dria
Coron-avirus deaths are increasing in Uganda.	Abafa akawuka ka ssennyiga kolona beeyongera mu Uganda.	Abarikwitwa Korona beeyongyeire omu Uganda.	To macalo adwogi me two Corona tye ka medde i lobo Uganda.	Iyata atwanare na ekorona ko Uganda.	Ba odrapi azo corona niri si 'diyi ma kalafe ni tu Kari Uganda niri ma alea tutu.
I need to renew my driving permit.	Neetaaga okuzza obuggya dulayivingi pamiti yange.	Nyine kugarura busya ekihandiiko ekirinyikiriza kuvuga ebiiruka.	Amito nongo pamit manyen me dwoyo mutoka.	Akoto eong isiekikinai bobo apermit ka apak. Akot aisiteteun akapapula naka airenge.	Le ma ma oja ma vile kokobi mutukari ma ti ojizu o'dinisi ra.

Multilingual models are particularly of interest for us here, because (1) it is more practical in terms of computational resources during training and deployment to have fewer models covering multiple languages, and (2) we are interested in the possibility of cross-lingual transfer, in which translation performance in one language is actually improved by the availability of training data for different – but related – languages.

Multilingual modelling employs the basic concept described in Johnson et al. (2017), that source text in training and test data can include a token specifying the desired language of the target text. In the case of a model which supports multiple source languages but only one target language (`mul-en`) then these tokens are not needed. We denote these target-language tags using the ISO 639-3 code for each language: `>>ach<<` (Acholi), `>>lgg<<` (Lugbara), `>>lug<<` (Luganda), `>>run<<` (Runyankole) and `>>teo<<` (Ateso). These were added as special tokens to the pre-trained `en-mul` tokenizer.

In line with the findings of Araabi & Monz (2020), we observe that relatively large batch sizes improve performance with limited training data. We train with a batch size of 3000 and initial learning rate of 1e-4. The number of training steps was determined by early stopping on validation loss. Training multi-lingual models took approximately four hours on a single V100 GPU. The trained models and demo are provided. [6] [7]

Table 6 provides results for the the baseline pretrained model and the multilingual models for translating from any of the local languages to English. Results of these models for the reverse translation from English to any one of these local languages is depicted in Table 7.

4.2 Augmenting our dataset with secondary data

To improve the model we augment our dataset with publicly available datasets; the FLORES-101 and the MT560 datasets that include one local language, Luganda for FLORES-101 and 3 local languages for MT560 as shown in Table 1.

We add these two datasets to our dataset and retrain both models; the (`mul-en`) and the `en-mul`. Performance of our models with a combination of these three datasets is shown in Table 6 and Table 7.

[6] `huggingface.co/Sunbird`
[7] `translate.sunbird.ai`

Table 5: Samples from supplementary training datasets for `en-mul` model training.

MT560 Source	Target
>>nyn<< "Stand firm in the faith,... grow mighty." - 1 COR.	"Muhamire omu kwikiriza,... mugume n'amaani." - 1 KOR.
>>lug<< Beware of satanic influence in entertainment	Weewale eby'okwesanyusaamu ebikubiriza endowooza za Sitaani
FLORES-101	
>>lug<< Following the race, Keselowski remains the Drivers' Championship leader with 2,250 points.	Nga ogobeledde jasibuka, Keselowski asigara ye kyampiyoni w'abavuzi akulembera nobubonero 2,250.
>>lug<< Fifteen of these rocks are attributed to the meteorite shower last July.	Kkumi na taano ku njazi zawebwayo eri ewava amazzi mu gwomusanvu gw'omwaka oguwedde.
Back-translation	
>>lug<< The chaos in the USAFI market.	Obuvuyo obuli mu katale ka USAFI.
>>ach<< The boat capsized on River Adwula and Lake Lira heading to Soroti.	Waya man okwalo igulu pii adit ame tye Adwila iyo aya i Lira woto Soroti.
Named entities	
>>teo<< Department of Natural Resources and Mines	Department of Natural Resources and Mines
>>lgg<< Catalonia	Catalonia
>>nyn<< Roberta Flack	Roberta Flack

Back-translation To extend our dataset further, we created back-translation data, using monolingual text. Bearing in mind the problems arising with web-scraped text outlined in Section 1, we carefully selected the text sources to be used for this. For example, rather than indiscriminately using English text, we curated a set of *locally relevant* English text e.g. from local news sites, containing the terms and concepts that we believed would be of interest to prospective local users of translation models. After filtering web-scraped data for obvious poor quality sentences and incorrect language, we obtained 88.6K English, 32.5K Luganda, 6.3K Acholi, and 1.1K Ateso sentences. We used our best `en-mul` model to create back-translated training data for the `mul-en` model and vice versa, for two iterations.

Named entities Early models that we trained had a tendency to fail with out-of-vocabulary (OOV) terms, particularly named entities. Models would attempt to translate proper nouns that should actually be simply passed through to the output. To help correct this tendency, we created a further dataset of named entities, in which the source and target text were identical. Named entities were sourced from the WNUT17 (Derczynski et al., 2017) and WikiGold (Balasuriya et al., 2009) databases. Supplementary training data created from these sources had identical text in the input and output, as shown at the bottom of Table 5, teaching the model to leave OOV terms unchanged in the output. This had no noticeable effect on BLEU scores, but a significant effect on subjective translation quality.

5 Results and Discussion

Our goal in building the dataset and models is to facilitate translation combining linguistic quality as well as contextual meaning. We ideally wish to evaluate this with human verification, which is

Table 6: BLEU scores: X → English.

	Training	ach	lgg	lug	nyn	teo
1	Pre-trained	-	-	13.7	-	-
2	Multilingual	13.9	13.2	30.4	23.8	17.5
3	(2) + MT560, FLORES	20.2	19.6	32.1	23.8	21.4
4	(3) + back-translation	**21.5**	**24.8**	**33.2**	**26.4**	**25.3**
5	Pairwise (including MT560, FLORES, back-translation)	19.0	18.9	31.3	23.4	20.7

Table 7: BLEU scores: English → X.

	Training	ach	lgg	lug	nyn	teo
1	Pre-trained	-	-	7.9	-	-
2	Multilingual	16.1	17.9	24.8	14.5	19.0
3	(2) + MT560, FLORES	**17.9**	19.0	25.9	**15.5**	19.9
4	(3) + back-translation	17.5	**19.9**	**26.7**	15.3	19.9
5	Pairwise (including MT560, FLORES, back-translation)	17.4	18.3	25.7	14.6	**20.0**

underway at the time of writing this paper. We therefore present statistical results using Bi-Lingual Evaluation Understudy (BLEU) score metric, computed on a 10% held out test set. While BLEU is not a perfect measure it provides a convenient indication of translation quality.

Results from our experiments are in Table 6, which depicts results of all our models with the different training data subsets for the (`mul-en`) translation task and Table 7 for the `en-mul` task. We only study translation to and from English in this study, and do not present any results on translation between local languages – though we view this as an interesting future direction.

We first note that all our results were strongly superior to the pre-trained OPUS baseline for Luganda. Performance is highest generally for models where the largest amount of data was used to train the model, i.e. our dataset with FLORES-101, MT560 and back-translation data. Back-translation was more effective for `mul-en` translation, since more English text was available and with a greater degree of consistency. Back-translation slightly deteriorates the performance of English to Acholi and English to Runyankore.

We also show the performance achieved when training a series of pairwise models (e.g. `en-ach`, `en-lgg`, `en-lug`, `en-nyn` and `en-teo` rather than one `en-mul`). In all but one case, the performance of multilingual model training is better than pairwise, particularly for translation to English where we observe BLEU score improvements of +1.9 to +5.9. This provides some evidence that cross-lingual transfer can occur for these languages, so that the model can benefit from simultaneous training of multiple different-but-related languages. We hope to continue to benefit from this effect by expanding the dataset with other languages in future, and merging it with other African language translation datasets.

6 Conclusion

Africa has very high linguistic diversity but a paucity of digital data resources. Public research has focused on datasets with limited scope, such as religious text, whereas data for proprietary translation systems are commonly developed by third party vendors who do not have access to the models under development. We opted to take a community-based approach to the collecting of more contextual language data. This allowed us to create a dataset and models which we believe to have high local relevance.

We have collected a multi-way dataset with at least 25k sentences in each of five Ugandan languages, and trained a set of effective, publicly available benchmark models. For some of the languages in the study, these are the first existing machine translation resources available, to our knowledge. Our data

collection framework also makes it easily extensible to other local languages, since translation would begin from the seed English text already available. This fits into our goal of creating sustainable and easily maintainable language datasets for Africa.

REFERENCES

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1310. URL https://aclanthology.org/P19-1310.

Ali Araabi and Christof Monz. Optimizing transformer for low-resource neural machine translation. *arXiv preprint arXiv:2011.02266*, 2020.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. Named entity recognition in wikipedia. In *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources (People's Web)*, pp. 10–18, 2009.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. *arXiv preprint arXiv:2010.14571*, 2020.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 140–147, 2017.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho B. Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *CoRR*, abs/2003.04986, 2020. URL https://arxiv.org/abs/2003.04986.

Peter Nabende. Towards data-driven machine translation for lumasaaba. In *The 2018 International Conference on Digital Science*, pp. 3–11. Springer, 2018.

Irene Nandutu and Ernest Mwebaze. Luganda text-to-speech machine. *arXiv preprint arXiv:2005.05447*, 2020.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2144–2160, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL https://aclanthology.org/2020.findings-emnlp.195.

Julius Omona and Nora Groce. Translation and research outcomes of the bridging the gap project: A case of the luo language, spoken in northern uganda. *Translation Studies*, 14(3):282–297, 2021.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3450–3466, 2021.

Jörg Tiedemann, Santhosh Thottingal, et al. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 2020.

Uganda Bureau of Statistics. The national population and housing census 2014–main report, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.