

Rochester Institute of Technology

RIT Digital Institutional Repository

Theses

8-2020

Machine-assisted translation by Human-in-the-loop Crowdsourcing for Bambara

Allahsera Auguste Tapo
aat3261@rit.edu

Follow this and additional works at: <https://repository.rit.edu/theses>

Recommended Citation

Tapo, Allahsera Auguste, "Machine-assisted translation by Human-in-the-loop Crowdsourcing for Bambara" (2020). Thesis. Rochester Institute of Technology. Accessed from

This Thesis is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.

Machine-assisted translation by Human-in-the-loop
Crowdsourcing for Bambara

by

Allahsera Auguste Tapo

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in Computing and Information Sciences

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology

August 2020

Machine-assisted translation by Human-in-the-loop
Crowdsourcing for Bambara

by

Allahsera Auguste Tapo

Submitted to the
B. Thomas Golisano College of Computing and Information Sciences
Department of Computer Science
in partial fulfillment of the requirements for the
Master of Science Degree
at the Rochester Institute of Technology

Abstract

Language is more than a tool of conveying information; it is utilized in all aspects of our lives. Yet only a small number of languages in the 7,000 languages worldwide are *highly resourced* by human language technologies (HLT) [47]. Despite African languages representing over 2,000 languages, only a few African languages are highly resourced [1], for which there exists a considerable amount of parallel digital data.

We present a novel approach to machine translation (MT) for under-resourced languages by improving the quality of the model using a paradigm called “humans in the Loop.”

This thesis describes the work carried out to create a Bambara-French MT system including data discovery, data preparation, model hyper-parameter tuning, the development of a crowdsourcing platform for humans in the loop, vocabulary sizing, and segmentation. We present a novel approach to machine translation (MT) for under-resourced languages by improving the quality of the model using a paradigm called “humans in the Loop.” We achieved a BLEU (bilingual evaluation understudy) score of 17.5. The results confirm that MT for Bambara, despite our small dataset, is viable. This work has the potential to contribute to the reduction of language barriers between the people of Sub-Saharan Africa and the rest of the world.

The thesis “Machine-assisted translation by Human-in-the-loop Crowdsourcing for Bambara” by Allahsera Auguste Tapo has been examined and approved by the following Thesis Committee:

Dr. Christopher M. Homan
Associate Professor, RIT
Thesis Committee Chair

Dr. Marcos Zampieri
Assistant Professor, RIT
Reader

Dr. Sarah Luger
Orange Silicon Valley
Observer

Michael Leventhal
Monde Moderne Mali Emergent (RobotsMali)

Dr. Julia Kreutzer
Google

Acknowledgments

This thesis has been an incredible experience, a source of learning and growing. I am grateful to the Fulbright program, International Institute for Education (IIE), Rochester Institute of Technology (RIT) for the opportunity to further my studies.

I would like to thank the following amazing people:

Hans-Peter Bischof, the program director.

Cindy Wolfer, my academic advisor, who is the best at what she does.

Christopher Homan, the chairman of my committee for the constant support and insights.

Michael Leventhal for tirelessly being involved in the project and making things move in Mali.

Julia Kreutzer for providing first hand guidance with JoeyNMT [53] and insights.

Sarah Luger for her efforts on always pushing forward.

Marcos Zampieri, the reader of my committee for his valuable inputs despite his busy schedule.

Patrick Kelly for his inputs about some aspects of the Malian culture.

Kadiatou Traore for the inputs about Voice of America (VOA).

Seydou Traore, liaison to the Academy of Malian Languages (AMALAN).

Nathan Berger for first hand guidance on the crowdsourcing platform [54].

Coleman Donaldson for his early input into the direction of our work.

Daan Van Esch for his early input into the direction of our work.

SIL-Mali for ethnographic information about Malian languages.

Masakane community for putting Africa on the NLP map.

Djeneba Traore for proofreading my thesis.

Ada Tapily for designing my presentation.

Matthew Herberger for helping to align some of the dokotoro texts.

This work is dedicated to all those who have done, continue doing, and will do work in their local languages to uplift their people.

Contents

1	Introduction	1
2	Related Work	5
2.1	MT	5
2.2	Crowdsourcing	9
2.3	Bambara	9
3	Methods	11
3.1	Data	11
3.2	MT Pipeline	17
3.3	Crowdsourcing Pipeline	18
4	Results and Discussion	23
5	Future Work	36
6	Conclusion	37
	Appendices	49
A	Additional Resources	50
A.1	Models Detail	50
A.2	Useful Links	50
B	Proposals	52
B.1	ISDB Proposals	52
B.2	Google Proposals	59

B.3 IRDC Proposals 68

List of Figures

1.1	The complete system once completed.	4
3.1	Custom aligner.	15
3.2	MT model.	17
3.3	Crowd-source pipeline.	20
3.4	Login interface.	20
3.5	Registration interface.	21
3.6	Marking.	22
3.7	Choice of mode.	22
4.1	Character level MT Transformer model.	24
4.2	Character level training loss.	25
4.3	Character level validation loss.	26
4.4	Character level validation perplexity.	27
4.5	Character level validation score.	28
4.6	Word level MT Transformer model.	29
4.7	Word level training loss.	30
4.8	Word level validation loss.	31
4.9	Word level validation perplexity.	32
4.10	Word level validation score.	32
4.11	BPE level MT Transformer model.	33
4.12	BPE level training loss.	34
4.13	BPE level validation loss.	34
4.14	BPE level validation perplexity.	35
4.15	BPE level validation score.	35

List of Tables

3.1	VOA dataset.	12
3.2	Dataset discovery information.	13
3.3	Main datasets.	14
4.1	System information.	23
4.2	Pilot study metrics.	26
4.3	Score variance standard deviation	26
A.1	Model details.	51

Chapter 1

Introduction

In this thesis, we present a state-of-the-art machine translation (MT) pipeline that automatically generates translations from Bambara text into French text and a crowdsourcing pipeline, tightly coupled to the MT component, that will recruit Malians to help collect parallel text to test, clean source data, and provide crucial expert and non-expert supervision for the MT model.

For historical reasons, French is the official language of Mali, used by the government and media for the diffusion of information. For the same reasons, the information sources that Malians have access to outside of Mali are francophone. Yet, only 35% of Malians¹ are sufficiently fluent in French to be able to understand the information from these sources. There is little information available in the native languages of Malians, known as “national languages.” One of these national languages, Bambara, is understood by 80% of Malians, some native speakers, some as an L2 language [11, 16, 24]. The government has attempted to make information available to the public in national languages, but these efforts have advanced little due to lack of resources [81]. Other non-governmental organizations (NGOs) like United States agency for international development (USAID) are active in the promotion of local languages [89] and education [28, 37, 40] throughout Africa [62]. Malians communicate with text using French, or Bambararized French. They do the same while using internet; others use Arabic and English to some extent. MT offers a path using automation to make information available to Malians in the languages that they understand. Bambara, understood by 80% of the population, is our first

¹<https://www.macrotrends.net/countries/MLI/mali/literacy-rate>

target as an MT system in this language will have the greatest impact.

We focus on translating Bambara French, since marking and post-editing French is an easier task than marking and post-editing Bambara due to the challenges in its writing, orthography, and diacritics. Our collaboration with AMALAN aims to alleviate those challenges to establish standard orthography and coinage of new words in Bambara where none existed before to name a few. The choice of pairing Bambara with French is due to the fact that our crowdworkers, while native speakers of Bambara, are most accustomed to reading and writing in French, the official language of instruction in universities in Mali.

Machine translation is a challenging problem in itself [52, 66], and although there have been some great advances made in recent years [43, 56, 76], only a few systems [32, 57, 61] have been trained to communicate in African languages. Commercial providers support high-resourced languages such as English, French, etc. Bambara, as is the case for all but a handful of African languages, is under-resourced in that it lacks a large quantity of parallel data, a base requirement for the most successful MT systems in current use. Machine translation driven by humans in the loop [5, 18, 41, 73] is an interesting and innovative approach that combines human intelligence with the computing power of a machine to improve the MT model through user feedback. Human-in-the-loop has the benefit of incorporating human contribution to a parallel corpus, whereas a one-off interaction lacks this important human component; the former gives an opportunity to the model to be fluent and robust, all of which is missing in the latter. However, this technique has not been applied to a low-resourced language, as far as we know. Thus, this work is a critical first step in an effort to address the lack of parallel data through a massive program of crowdsourcing. In addition to providing needed data, crowdsourcing is a means to engage the Malian public in a technology project capable of furthering the social and economic development of their country and the development of post-colonial identity built around Malian languages and culture [63, 67, 85].

The crowdsourcing pipeline uses Bambara texts and their corresponding French texts translated using our MT model to crowd-source the tasks of marking and post-editing the French texts. To the best of our knowledge, there is no system currently available to translate from Bambara to English/French or from English/French to Bambara.

NMT is currently the most successful approach to generalized machine

translation. Used alone, it appears to be unsuited for under-resourced languages such as Bambara due to the lack of the quantity of labeled data (parallel texts) needed. Our hypothesis is that NMT combined with human-in-the-loop crowdsourcing will generate the volume and quality of data needed for successful NMT, improving translation quality and breadth over NMT-only approaches. In order to obtain a sufficient quantity of labeled data, the project will additionally test the hypothesis that a crowdsourcing interface, coupled with process design, can enable non-professional annotators to contribute data suitable for training an NMT. An overarching hypothesis for this work is that machine translation will enable Bambara speakers to understand information in their native language, and increase literacy, which, in turn will foster social and economic development. This project will not attempt to prove this hypothesis, but it will contribute results that are needed in order to test it.

We translate from Bambara into French text a held out test dataset of 488 sentences utilising our MT model build using JoeyNMT [53]. We measure the meaningfulness and fidelity of translation of the model compared to a translation produced by a human translator using common metrics such as BLEU (bilingual evaluation understudy) [70], an automatic reference-based metric utilized to assess the quality of text which has been machine-translated from one natural language to the other. We evaluate our models before and after user feedback, and we also contribute new evaluation metrics aimed at measuring how successfully our results convey critical information to end users, via our crowdsourcing component [54]. We utilize crowdsourcing to produce human-computer-interface (HCI) knowledge about how to design crowdsourcing interfaces that best elicit translations from Bambara to French in terms of quality, quantity, and the acceptance of users of the interface. In this case, the quality is determined by how well the data collected can be used to train and evaluate an MT system, by integrating expertise from a cohort of Malian crowdworkers who understand both Bambara and French.

Our contributions include a dataset of quality (e.g., in terms of cleanness and alignment) data of French-Bambara and English-Bambara pairs, an MT pipeline to translate from Bambara into French, as seen in Figure 3.2, and a crowd-sourced pipeline, as seen in Figure 3.3, to mark and post-edit French translations and use the feedback to improve our MT model, both are discussed in detail in chapter [3]. Figure 1.1 shows the entire system as a whole design once the project is completed. **Note:** Not everything on this figure is covered

in this thesis.

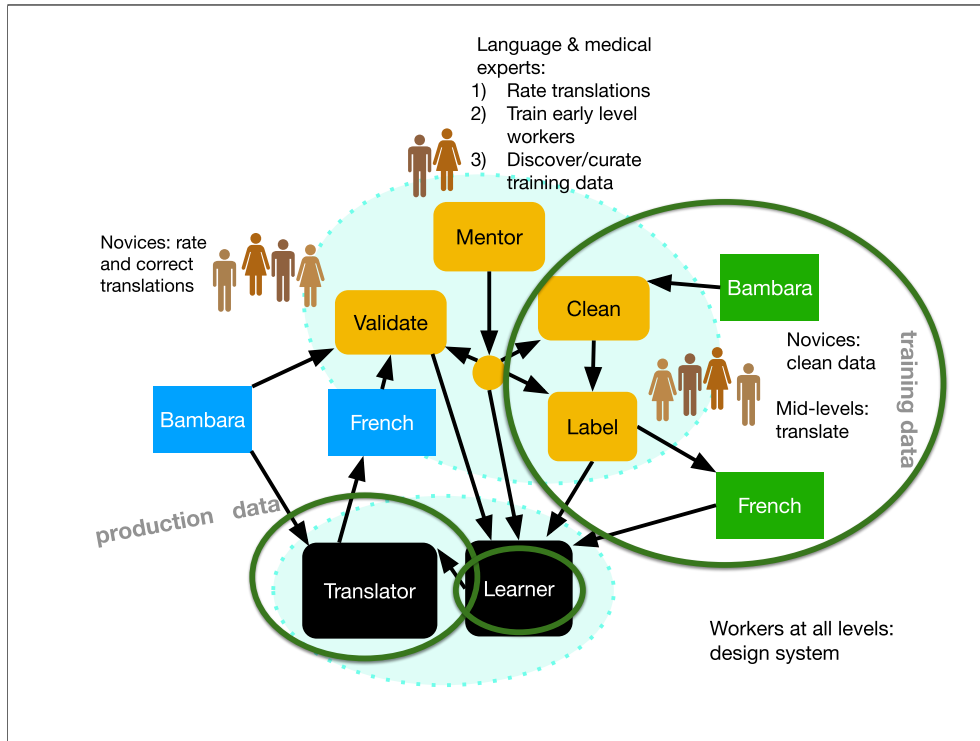


Figure 1.1: Integration diagram of the entire system as it should be once completed. This thesis however, focuses on dark green circles. The MT pipeline as shown in Figure 3.2 and Figure 3.3 component (1), and the crowdsourcing pipeline as shown in Figure 3.3 components (1), (2) and (3)

The remainder of the proposal is structured as follows:

Chapter [2] presents related work. Chapter [3] describes the technologies, methodologies, and techniques utilized in building our MT models and our crowdsourcing pipeline. Chapter [4] discusses the results. Chapter 5 presents future work. Chapter [6] concludes the discussion on lessons learned. Appendix [A] presents additional resources discussed in the research. Finally, Appendix [B] presents the grants proposals we submitted during the course of this thesis.

Chapter 2

Related Work

2.1 MT

Machine translation (MT) is an NLP task part of computational linguistics. It explores the translation of text or speech from a language to another utilizing technology. Translation involves composition, sentence structuring, semantic complexities such as intent, and requires handling of language-specific morphology, disambiguation, etc.

So far the two major issues for machine translation are ambiguity and non-standard speech [12, 51]. There exists two techniques that are applied to disambiguation, shallow and deep approaches. Shallow approaches assume no knowledge of the text, and simply apply statistical methods to the words surrounding the ambiguous word. Deep approaches presume a comprehensive knowledge of the word. So far, shallow approaches have been more successful [27]. The problem of non-standard speech, one of the major pitfalls of MT, is its inability to translate non-standard language with the same accuracy as standard language. In this work we do not directly address disambiguation or non-standard speech.

Below are some previous and related work that researchers have explored during the course of time in the field of MT.

Rule-based MT (RBMT) [39] generates translations based on linguistic information guided by the dictionaries and grammars of the bilingual corpora. One of its biggest limitations is the lack of good dictionaries in quantity, particularly for low-resource languages like Bambara, and building new ones is

expensive. These techniques may be augmented by post-processing using language statistics and statistics guided by rules [55] aiming at correcting the generated output to produce better quality and increase productivity.

Statistical MT (SMT) [50] generates translations based on statistical models whose parameters are obtained by processing the bilingual corpora addressing some of the limitations of RBMT. One of its biggest limitations is that the system depends on huge amounts of parallel texts.

Example-based MT (EBMT) [82] generates translations based on examples encountered during training from the bilingual corpora. One of its biggest limitations is that the system is unable to translate new examples that it did not encounter during training.

Hybrid MT [34] takes advantage of the strengths of both, the statistical system and the rule-based system translation methodologies. The combination of these two different systems significantly retains the limitations of each approach.

More recently, with the advancement in the field of neural MT (NMT), a new kind machine translation system is emerging that combines the benefits of rules, statistical, and neural machine translation [68]. The approach allows MT models to utilize the strengths of pre-processing and post-processing in a rule guided workflow as well as to utilize those of NMT and SMT. The downside is the inherent complexity that makes the approach suitable only for specific use cases (e.g. domain specific task).

NMT [9,92], a deep learning based approach to MT, where massive amounts of data are fed to the model and it learns by predicting the likelihood of a sequence of words, typically representing entire sentences in a single integrated model from the data of a source language (e.g., Bambara) and a target language (e.g., French). Current state-of-the-art MT systems perform well on languages for which there exists adequate quantities of aligned text:, i.e, corpora in the source and target languages, where for each sentence in the source language (e.g., Bambara) there is a translation in the target language (e.g., French). There is currently a growing literature on best practices for adapting such machine translation systems to under-resourced languages [49,64], such as Bambara, which has been passed down through oral tradition, and lacks a large volume of text aligned with any potential target language. In future work, we will avail ourselves of these methods and possibly contribute to this literature. However, we believe a more effective approach is to focus on crowdsourcing

innovations that will increase the volume of aligned text with an interface and process flow that is able to extract useful training data from the noisy data produced by non-professional translators.

Transformers [88] are based on the encoder-decoder paradigm. The encoder consists of a set of encoding layers to process the input iteratively one layer at a time, and the decoder also contains a set of decoding layers stacked one after the other that process the output of the encoder and generate one word at a time.

An encoder layer [88] has two major components: a self-attention mechanism and a feed-forward neural network. It processes its input to generate encodings, containing information about the importance of the parts of the inputs relative to other parts. The next encoder layer takes as inputs the set of encodings of the previous layer.

A decoder layer [88] has three major components: a self-attention mechanism, an attention mechanism over the encodings, and a feed-forward neural network. It reverses the processing of the encoder layers, the encodings are processed using their contextual information to generate an output sequence. The attention mechanism weighs the importance of every other input and encodes or extracts information into or from them accordingly to produce the output. The additional attention mechanism of the decoder layer extracts information from the outputs of previous decoders, before the decoder layer extracts information from the encodings.

The feed-forward neural network for both the encoder and decoder layers carry out additional processing of the outputs, and contain residual connections and layer normalization steps.

Transformers were introduced in 2017 [88]. They are designed and implemented to work with sequential data both in order during inference and in parallel during training unlike RNNs which only handle sequential data in order only both during inference and training. From their introduction, transformers have become the go-to model for many NLP related tasks, replacing older recurrent network models such as the long short-term memory (LSTM). Since transformers are parallelizable during training, they made training larger datasets possible. This led to advances of pre-trained systems such as BERT (bidirectional encoder representations from transformers) [29] and GPT (generative pre-trained transformer) that can be fine-tuned for specific NLP tasks [31, 95].

Prior to the introduction of transformers, most state-of-the-art NLP systems relied on gated RNNs [22], such as LSTMs [83] and Gated Recurrent Units (GRUs) [21], with added attention mechanisms [20].

Gated RNNs process tokens sequentially, maintaining a state vector that contains a representation of the data seen after every token [21] referred to as word embedding. To process the n^{th} token, the model takes into account the word embedding which is the state representing of the sentence up to the $(n - 1)^{\text{th}}$ token with the information of the new token to create a new state, the current state has the representation of the sentence up to token n . The model's state at the end of a long sentence often does not contain precise, extractable information about early tokens due in part to the vanishing gradient problem [44].

This problem was taken care of by the introduction of attention mechanisms [75]. They give a model attention to have context, thus enabling the model to make sense of the state at any earlier point in the sentence. The attention layer can access all previous states and weights them according to some learned measure of importance to the current token, providing clearer information about far-away important tokens. A clear example of the utility of attention is in the task of translation. In a Bambara-French translation system, the first word of the French output most probably depends heavily on the beginning of the Bambara input. However, in a classic encoder-decoder LSTM model, in order to produce the first word of the French output the model is only given the state vector of the last Bambara word. Theoretically, this vector can encode information about the whole Bambara sentence, giving the model all necessary knowledge, but in practice this information is often not well preserved. If an attention mechanism is introduced, the model can instead learn to attend to the states of early Bambara tokens when producing the beginning of the French output, giving it a better concept of what it is translating.

When added to RNNs, attention mechanisms led to large gains in performance [60, 69]. The introduction of the transformer brought to light the fact that attention mechanisms were powerful alone by themselves [84]. Furthermore, it demonstrated that sequential recurrent processing of data was not necessary to accomplish the performance gains of RNNs with attention. The transformer uses an attention mechanism but no RNN [48]. It processes input tokens at the same time and calculating attention weights among them. The fact that transformers do not rely on sequential processing, allows them to be

trained more efficiently on larger datasets.

Scaled dot-product attention units constitute the building blocks of transformers. The following equation¹ is a dot-product attention.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where Q is the query, K is the key, and V is the value. Q, K, and V are defined as matrices.

2.2 Crowdsourcing

Crowdsourcing [17,35] draws from a long line of research [30,86,93] where there exists a sourcing model in which individuals or organizations obtain goods and services (including ideas) from a variable number of relatively open and often rapidly evolving groups of participants by decomposing complex processes into simpler micro-tasks.

There are numerous researches investigating crowdsourcing in MT [6,42] and they discuss challenges like quality control [58], and partial solutions [65].

The literature demonstrates, and our experience confirms, that a crowdsourcing tool is not, by itself, a complete solution for successfully crowdsourcing data. The total solution requires careful process design (e.g., quality control, training, etc.), since you're working with humans.

2.3 Bambara

Bambara is a member of the Mande family of languages [91] spoken principally among the Bambara ethnic group in Mali. It is usually written using the Latin alphabet, though the Arabic and N'ko alphabets are also used. Literacy is a major issue in Mali, especially in rural areas. Although written literature is only slowly evolving due to the predominance of French as the "language of the educated", there is a rich oral literature, consisting largely of tales of kings and heroes mainly translated by the "Griots", who are a mixture of storytellers, praise singers, and human history books and who have studied the trade of singing and reciting from generation to generation.

¹[https://en.wikipedia.org/wiki/Transformer_\(machine_learning_model\)](https://en.wikipedia.org/wiki/Transformer_(machine_learning_model))

“A language is considered low resourced when no automated human language technology (HLT) capability exists depending on whether the language”²: is mainly spoken [2, 13, 14, 25], or lacks written or spoken data [26, 45], or mixture of those. A language could be considered low resourced when there is no publicly available data to investigate it although it might not be necessarily limited to digitized data.

This thesis explores a machine translation (MT) pipeline utilizing JoeyNMT [53], a Pytorch [72] based framework that re-implements baselines from major publications, tightly coupled with a crowdsourcing [54] pipeline to mark and post-edit French texts of translations from Bambara to French.

This work has been motivated by the potential of MT to empower an entire nation in ways that were never experienced before, by opening its people to the world, and the world to its people by bridging the gap of language barriers.

²<https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents>

Chapter 3

Methods

The link to our data source can be found at https://github.com/israaar/mt_bambara.

3.1 Data

Our initial objective was to directly build a speech-to-speech system [46] from the source language (e.g., Bambara) into the target language (e.g., French) through an automatic speech recognition (ASR) model [15]. As Bambara is mainly a spoken language, speech, rather than text, is the most accessible output to a majority of potential users. We started looking into the literature of what was done, being done, and left to be done. We realised that almost all systems used an approach where the speech from the source language is first transcribed into text, referred as speech-to-text (STT) [38], translated into the target language text, and then synthesized the translated text into the target language speech, referred as text-to-speech (TTS) [77]. It was clear, after discussion with other researchers, that we would have to do text-to-text translation anyway so we decided to start with that approach. This would also give us the opportunity to compare the performance of both approaches.

Finding Bambara data to investigate our hypotheses turned out to be challenging, and finding “good” (e.g., aligned with another language such as French or English) data was more so. Table 3.2 lists some more resources we came across during our data discovery phase, but none of them ended up being a viable source of data for our current research, mainly due to factors such as

unavailability of the data in text or speech form with Bambara, or restriction on usage (e.g., query only dictionary). These dictionaries however could be of reference to the workers to acquaint themselves with new words or finds examples of usage of words.

During our data discovery phase we had identified the US agency for global media, Voice of America (VOA) to potentially collect six years worth of speech data, as seen in Table 3.1.

Title	Duration (Minutes)	Frequency per Week
Mali Kura	30	5
An Ba fo	60	1
Farafina Foli	60	1

Table 3.1: Six years worth of speech data from VOA

Another potential data resource is the Washington Correspondant News Forum in Bambara with its discussions of headline news. The timing of broadcasts is variable, but generally about 8 minutes long.

This gives a total of **14,078.583 minutes** of speech data along with the written Bambara (in Latin alphabet, but not always respecting the standard orthography nor diacritics).

URL	Description	Rate	Pros	Cons
http://www.alanwood.net/unicode/n127ko.html	Test for Unicode support in web browsers	2	character (n'ko, decimal), decimal, character (n'ko, hex), lex, name.	no audio present, no usable data.
http://www.fakoli.net/index.html	N'ko Institute of America, dedicated to the exploration of n'ko	2	accessible to English speakers.	confusing, writes n'ko using latin alphabets, no audio, and no usable data.
http://www.personal.psu.edu/ejp10/symbolcodes/bylanguage/nko.html	N'ko computing information (penn state)	2	facilitate installation of n'ko fonts and gives general information of it as a font	no usable data.
https://ipfs.io/ipfs/QmXoypzjW3WmFjNkLwCnL72vedxjQzDDPmXw6bucc/v1k1/N127Ko_a1phabet.html	N'ko alphabet	2	provides information about the alphabet from all angle possible	no audio present, no usable data.
http://www.bmanuel.org/clr/clr2.mp.html	Multilingual and parallel corpora	1	n/a	Unaccessible when we recently visited, no usable data.
https://podcasts.apple.com/us/podcast/learn-bambara-with-jeff-franzee/id45366610	podcast	2	short speeches.	not accurate, no usable data.
http://www.rfi.fr/emission/kan-jum-be-yen-mandenkan	rss/podcast	2	audio available	not transcribed, no usable data.
https://www.voabambara.com/podcasts	rss/podcast	2	audio/video available	not transcribed, no usable data.
https://www.voafrique.com/z/3551	podcast	2	audio/video available	not transcribed, no usable data.
https://www.voabambara.com/z/5098	podcast	2	audio/video available	not transcribed, no usable data.
http://ma.rfi.fr/	streaming station	2	available	not transcribed, no usable data.
http://dictionary.ankataa.com	dictionary	2	available; easily understandable	no phonetics, no usable data.
http://www.mali-pense.net/bm/lexicon/index.htm	dictionary	2	french/bambara – bambara/french	no supporting audio, no usable data.
http://command_huma-num.fr/	project of bambara de reference	3	Dictionary like	Not publicly available.

Table 3.2: Datasets for Bambara including: a URL, linking to where we found the source; a description, describing what it is about (as we understood it); a rating from 1 to 5, one is “related to our topic with no usable data”, and five is “related to our topic with usable data”; pros, describing its pluses; and cons, describing its minuses.

We currently have the following corpora datasets for the initial training of the system, as shown in Table 3.3, containing data from a dictionary dataset from SIL Mali¹ with examples of sentences used to demonstrate word usage in Spanish, French, English, and Bambara; and a tri-lingual book entitled “Where there is no doctor”²

Language		Bambara	French	English
Dictionary data	glosses (in words)	3,548	4,847	4,855
	examples (in sentences)	2,023	2,021	2,021
	combined	5,571	6,868	6,876
	aligned	2,160	2,146	2,160
Split % (in sentences)	training	75		
	validation	12.5		
	test	12.5		
Medical data	bigrams	26,430	25,746	31,412
	chapters	27		
	files	336		
	paragraphs	9,336	9,367	9,356
	stopwords	147	123	69
	trigrams	5,816	11,312	21,398
	wordlist	8,209	9,893	6,935

Table 3.3: The dictionary dataset from SIL Mali with examples in Spanish, French, English, and Bambara and the tri-lingual book “Where there is no doctor” dataset.

We pre-processed the dictionary data from a “.lift” file which was in xml [71] format using python [87] and python’s built-in xml module; then we used an aligner that we implemented, as seen in Figure 3.1, developed in python to manually align sentences and to save those sentence pairs that a human editor considered properly aligned. We considered only French, English, and Bambara examples. Furthermore, we divided the data into the following pairs: French-Bambara and English-Bambara, and both pairs have some overlaps.

We started with 2,021 sentences from the examples of the dictionary data for each language; the challenges were numerous, and implementing an automated tool to deal with the different types of errors that were found in the text was quickly found to be an exceedingly complex task.

¹<https://www.sil-mali.org/en/content/introducing-sil-mali>

²<https://gafe.dokotoro.org/>



Figure 3.1: The custom aligner we developed to manually align the dictionary dataset. The controls are as follows: for each language, “>” goes to the next item and “<” goes to the previous item; for all languages, “>>>” goes to the next items and “<<<” goes to the previous items; “Aligned B-F-E” saves to memory the alignment of all 3 languages; “Aligned B-F” saves to memory the alignment of Bambara and French items; “Aligned B-E” saves to memory the alignment of Bambara and English items; “Save” saves to a new file; “Continue Saving” continues saving the file created.

Bellow are examples of the cases we had to handle manually, going through the entire file line per line.

The case where only one language is represented.

For those cases where only one language is represented, we discarded them.

The case where multiple pronouns are separated by “/”.

Before: French - Bambara

Il/elle est né à Bamako en 1938. - A bangerà Bamakò san 1938.

After: French - Bambara

Il est né à Bamako en 1938. - A bangerà Bamakò san 1938.

Elle est né à Bamako en 1938. - A bangerà Bamakò san 1938.

The case where the meaning is explained in parentheses in the other languages while those are absent in Bambara example.

Before: French - Bambara

Un doigt ne peut pas prendre un caillou (C’est important d’aider les uns les autres). - Bolokòni kelen tɛ se ka belɛ ta.

After: French - Bambara

Un doigt ne peut pas prendre un caillou. - Bolokòni kelen tɛ se ka

bɛlɛ ta.

The case where “Proverb:” is used to indicated proverbs.

Before: French - Bambara

Proverbe: Une longue absence vaut mieux qu’un communiqué (d’un décès). - Fama ka fisa ni kɔmunike ye.

After: French - Bambara

Une longue absence vaut mieux qu’un communiqué. - Fama ka fisa ni kɔmunike ye.

We had a dataset of 2,161 sentences of Bambara-French pairs at the end of the pre-processing step. We split the data into training, validation, and test sets of 75%, 12.5% and 12.5% of the data respectively. In other words, the training set is composed of 1611 sentences, the validation set of 268 sentences, and finally, the test set of 267 sentences. Each language pair was in its own separate files as: train.bam, train.fr, dev.bam, dev.fr, test.bam, test.fr in a data directory.

There is no standard tokenizer for Bambara. Therefore, we simply apply whitespace tokenization for word-based NMT models and BLEU computation.³ As part of the pre-processing step, we segmented the data for both language pairs into subword units using subword-nmt⁴, and applied BPE (byte pair encoding) dropout to the training sets of both languages [74] with a vocabulary size of 500 and 1000. This step allows the model to mask part of words, which in turn enables the model to generalize better. Our observations are discussed in Chapter 4. Each language pair was in its own separate files as: train.bpe.bam, train.bpe.fr, dev.bpe.bam, dev.bpe.fr, test.bpe.bam, test.bpe.fr in the same data directory as for the previous step.

A great deal of effort and time went into corresponding with other institutions and researchers. Thanks to these efforts, we have obtained the cooperation of the Institut National des Langues et Civilisations Orientales (INALCO) in France, giving us access to the Corpus Bambara de Référence [90], the largest corpus on Bambara-French that we know of. Due to time constraints, this new data source will be utilized in future work. Our perspectives for future are discussed in more detail in chapter 5.

³SacreBLEU BLEU+case.mixed+numrefs.1+smooth.exp+tok.none+version.1.4.9, chrF2+case.mixed+numchars.6+numrefs.1+space.False+version.1.4.9

⁴<https://github.com/rsenrich/subword-nmt>

3.2 MT Pipeline

Our MT pipeline, as seen in Figure 3.2 and Figure 3.3 component (1), utilizes a transformer model [88], widely considered to be the best model [10] for translating highly-resourced languages-based implemented via JoeyNMT [53].

Our model is based on encoder-decoder approach [19]. Encoder and decoder are each made up of transformers consisting of six layers with four heads per layer, and a word embedding layer of 256 dimensions. The encoder transforms a source sentence to a fixed-length context vector and the decoder learns to interpret the output from the encoded vector by maximizing the probability of a correct translation given a source sentence.

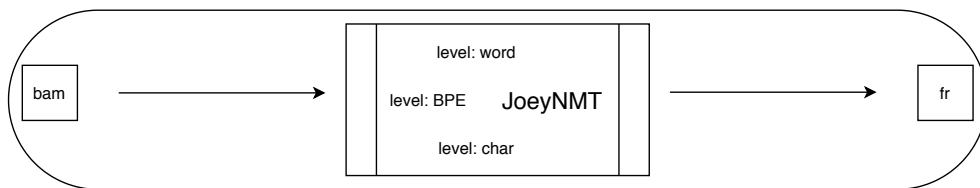


Figure 3.2: MT model to translate from Bambara to French.

Beyond this relatively well-established model, there are some key subjects of ongoing research in the under-resourced language community that we investigated. The first is the translation of multiple languages at once [4, 7]. We believe it will be beneficial to leverage such multi-language approaches whenever possible due to the small size of training set of Bambara and French aligned pairs. Another emerging approach is to train and obtain feedback from translations in both directions, referred to as back translation [33] or data augmentation [36].

We explored character level [59], word level [78], and BPE (byte pair encoding) level machine translation [80] and utilized the best performing model with our dataset to translate the test set to feed the output and the test set to our crowdsourcing pipeline.

Because there is no one right answer for translations, these methods can be used to substantially increase the number of aligned texts, from a much smaller seed set. They can also help leverage translators that may know only the target (e.g., French) or the source language (e.g., Bambara). Variations on these approaches that emerge over the course of this thesis may lead to new

innovations that could apply beyond our setting. We can use this technique for only a very limited set of the texts we have available to use, chiefly religious in nature, whose language is archaic and not relevant to the general domain [3].

3.3 Crowdsourcing Pipeline

We performed some preliminary experiments to help design this system. As an initial experiment we published *Assessing Human Translations from French to Bambara for Machine Learning: a Pilot Study* [57] at AfricaNLP workshop co-hosted with the International Conference on Learning Representations (ICLR) 2020 in Addis Abeba, Ethiopia. The conference presentation was made via Zoom due to conversion of the conference to a virtual format during the Coronavirus pandemic. The goal of this pilot study was to guide our choice in selecting and designing the kind of data to collect and the tools to build in that regard.

The crowdsourcing mechanism is set up, but the study has not been concluded, so the research question cannot be answered yet. The crowdsourcing pipeline is the one used in [54] with the language pairs substituted for Bambara-French. It has three main functions: (a) providing marking ability (see Figure 3.6) to crowdworkers to mark part or all of the machine translated text that is not correctly translated (e.g., the translation does not convey the intended meaning, or there is a better way of conveying the idea, or the translation is incorrect), (b) providing post-editing ability (see Figure 3.7) to crowdworkers to post-edit part or all of the machine translated text that is not correctly translated (e.g., the translation does not convey the intended meaning, or there is a better way of conveying the idea, or the translation is incorrect), and (c) providing the ability to take the target corrected texts (obtained through post-editing) along with the source texts and to add them to our training set, thus providing more data and better data for subsequent training cycles.

Our platform (see Figure 3.4) supports two micro-tasking components, as seen in Figure 3.3 component (2): marking, where crowdworkers take newly translated texts from Bambara to French and manually mark them in a website textarea-like environment; post editing, where crowdworkers take newly translated texts and if and when need be mark, or edit the French texts to match the Bambara texts. Additionally, the crowdsourcing pipeline offers a mode where the crowdworker can choose of those two micro-tasks.

Tying these two micro-tasks together, as seen in Figure 3.3 component (3), is the collection of the new Bambara and French pairs that can be utilised to augment our training data. We simply apply established tests for evaluating annotator quality, such as BLEU scores and inter-annotator agreements (e.g., agreement among annotators regarding marking or post-editing the same French text) [8].

We then use the tests from the initial phase to model the reliability of the workers and weigh their inputs to the machine learner, based on the classic approach of Dawid and Skene [79]. Once completed, the system will offer more experimental components that provide workers feedback and rewards to incentivize performance and help improve the fluency in written Bambara and French and the skill as translators of crowdworkers.

We will elicit feedback from all crowdworkers on how to improve the system and make it fairer and a better training experience. A major challenge is in recruiting crowdworkers who may have little familiarity with formal translation tasks. [57]. We will begin with a relatively small group of 5-20 crowdworkers recruited from students in Mali, who are more fluent in written Bambara and French than the general Malian population, and with whom we previously collaborated with on language translation tasks.

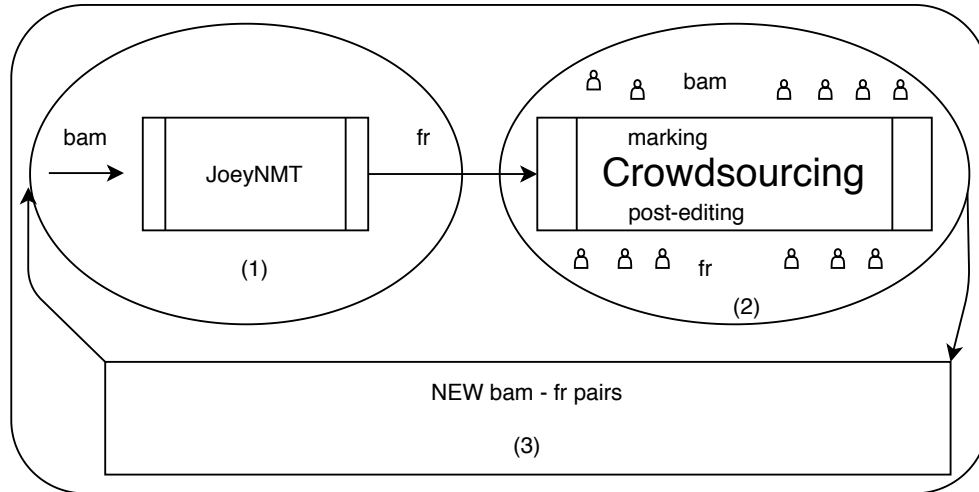


Figure 3.3: MT pipeline translating from Bambara to French, tightly coupled with the crowdsourcing pipeline. (1) the MT model that translates Bambara into French using word, character, or BPE level translation based on the best BLEU score. (2) the crowdsourcing platform that enables crowdworkers to either mark or post-edit French texts translated by component (1) from Bambara texts. (3) The collections of newly French texts marked or post-edited along with their corresponding Bambara texts that will be added to the training set to make the MT model robust.

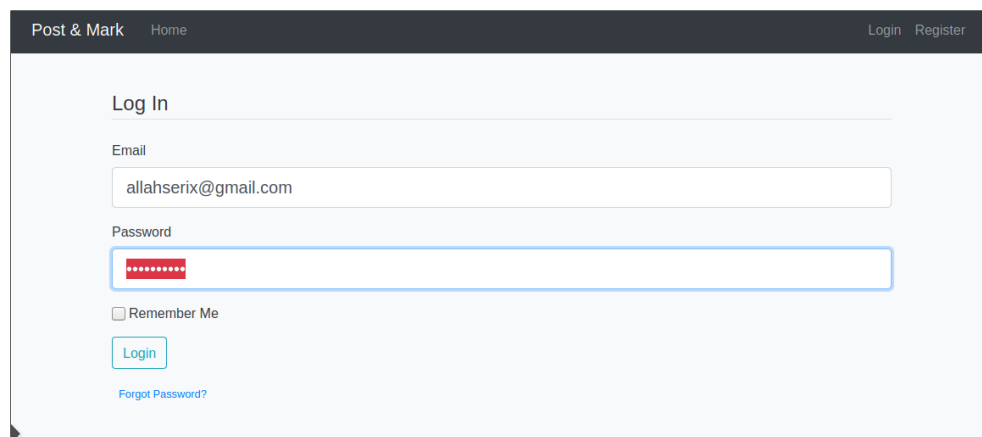
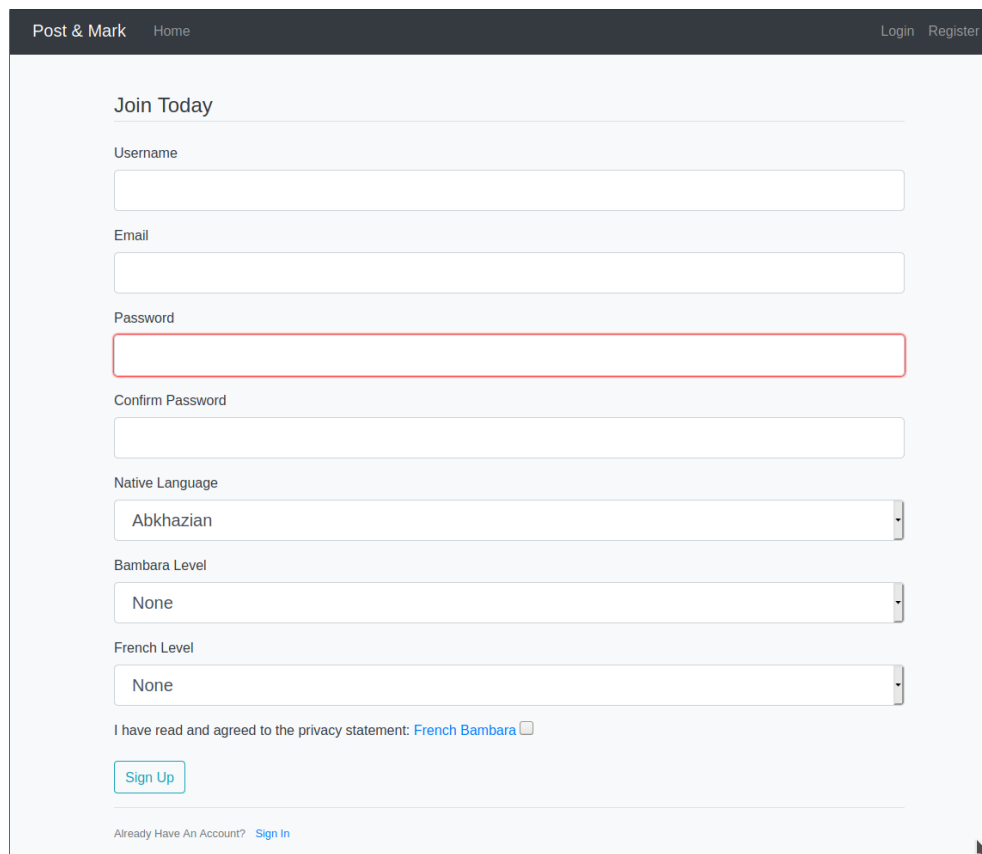


Figure 3.4: The interface shown when landing on the home page.



The image shows a registration form on a website. At the top, there is a dark navigation bar with 'Post & Mark' and 'Home' on the left, and 'Login' and 'Register' on the right. The main content area is titled 'Join Today' and contains several input fields: 'Username', 'Email', 'Password' (highlighted with a red border), 'Confirm Password', 'Native Language' (set to 'Abkhazian'), 'Bambara Level' (set to 'None'), and 'French Level' (set to 'None'). Below these fields is a checkbox for 'I have read and agreed to the privacy statement: French Bambara'. A 'Sign Up' button is located at the bottom of the form. At the very bottom, there is a link for 'Already Have An Account? Sign In'.

Figure 3.5: The interface shown when registering on the crowd-source platform.

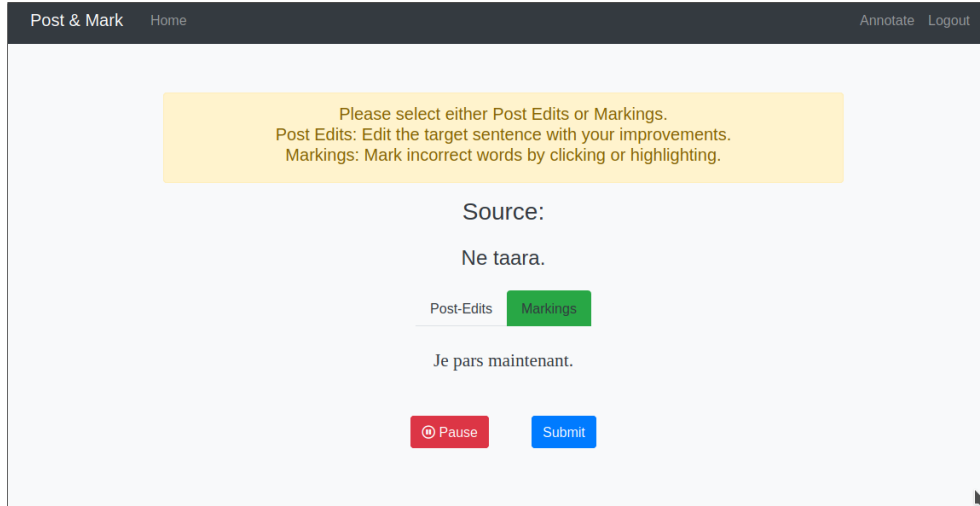


Figure 3.6: The interface shown in the user-choice mode. The user chose Markings here allowing her to mark errors in the target language (e.g., French). The interface will be the same when post-editing is chosen except that the “Post-Edits” will be highlighted.

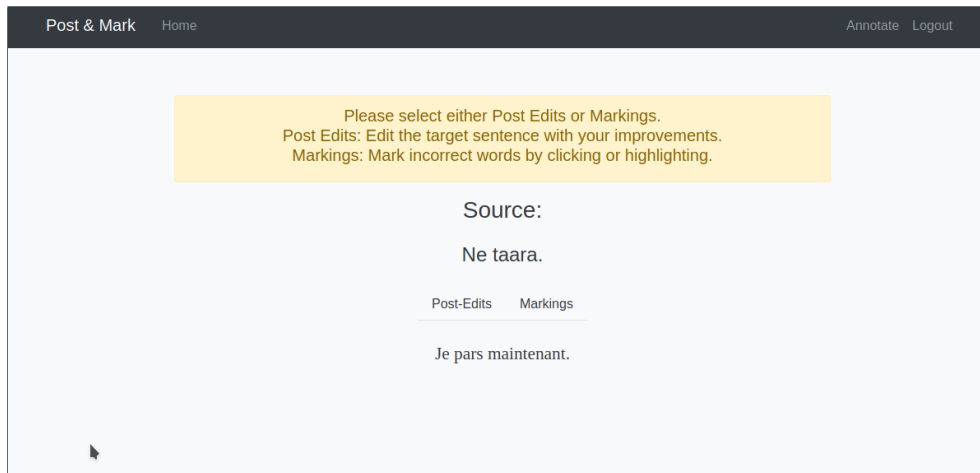


Figure 3.7: The interface shown when choosing between post-editing or marking the errors in the target language (e.g., French).

Chapter 4

Results and Discussion

We trained our models on a Dell Inspiron 15 7000 2-in-1 with GPU capability running a GNU/Linux operating system with the Kernel 5.4.0-42-generic x86-64 architecture. See Table 4.1 for the details.

Software	Version	Hardware	Amount
KDE Plasma	5.12.9	Processors	8 x Intel Core i7-8550U CPU @ 1.8GHZ
KDE Frameworks	5.47.0		
Qt	5.9.5		
Kernel Version	5.3.0-51-generic	Memory	15.4 GiB of RAM
OS Type	64-bit		

Table 4.1: Details of the system on which we trained our models.

The crowdsourcing mechanism is set up, but the study has not been concluded, so the research question cannot be answered yet.

We ran all experiments in under three hours each. As shown in Figure 4.1, 4.6, and 4.11 our best performing model is the BPE model with a BLEU score of 17.5.

Malian university students translated French texts of the Malian news broadcast and the Wikipedia article, producing either written or oral translations to Bambara. Our results suggest that similar quality can be obtained from either written or spoken translations for certain kinds of texts. They also suggest specific instructions that human translators should be given in order to improve the quality of their work.

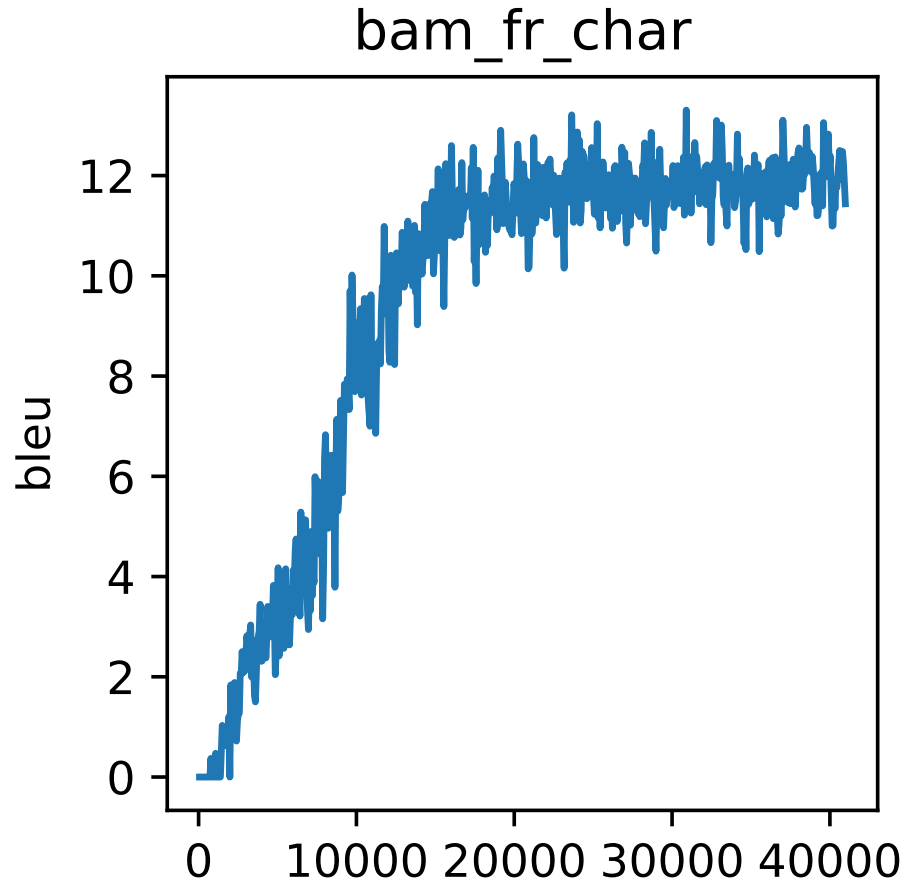


Figure 4.1: Plot of the validation data: Bambara to French Transformer model at character level.

Table 4.2 shows the results for the Malian news broadcast and the Wikipedia article. The differences in average scores between the news broadcast and the Wikipedia article, aside from the small sampling, most probably reflect the different challenges of the texts. The news broadcast is essentially an oral text and it is easier to reproduce the exact meaning with a more colloquial style.

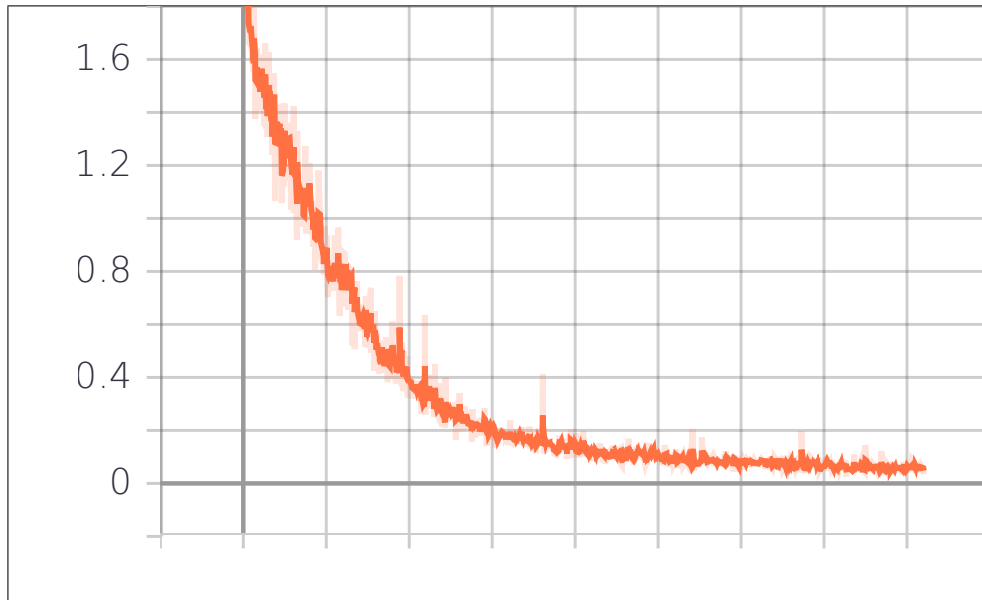


Figure 4.2: Character level: the training loss decreasing.

The Wikipedia article has long and complex sentences, making it easier to miss details and nuances while the translator hews closer to the French source and falls back on more formal standard Bambara.

The translations of the news broadcast showed a relatively limited difference in meaning and use of standard Bambara between written and oral translations, but significant difference in the literalness of the translations. The relatively large standard deviations shown in Table 4.3 indicate a wide range of quality between translators and translations, suggesting that screening translations based on basic quality metrics may be necessary and effective.

Here is an example of a written and oral translations of French text, followed by the qualitative and quantitative evaluation of quality.

French Objectif réfléchir à de nouvelles stratégies de lutte contre le terrorisme qui continue de faire des victimes dans le sahel.

English Objective to reflect on new strategies to fight terrorism which continues to claim victims in the Sahel.

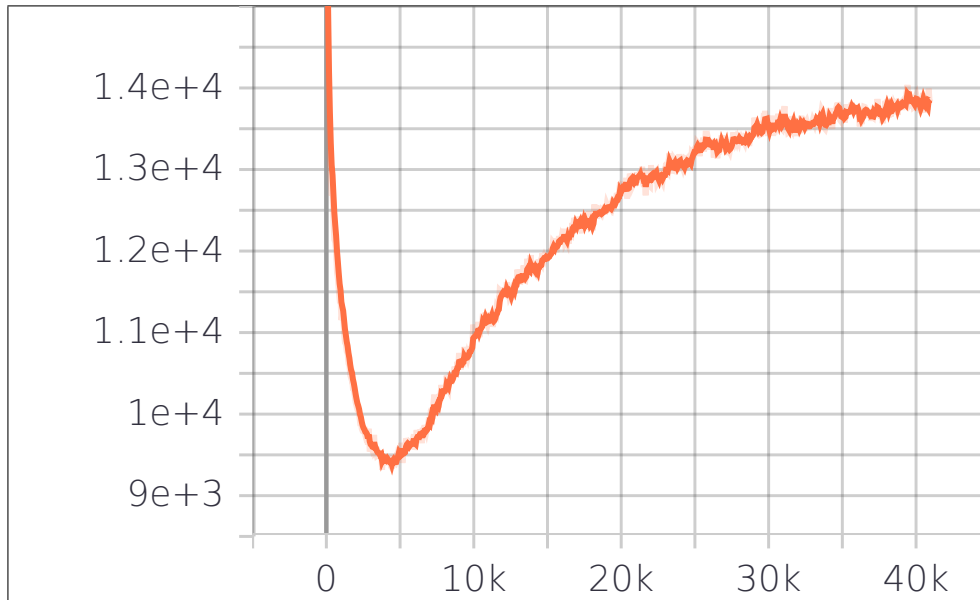


Figure 4.3: Character level: the validation loss was decreasing but start increasing probably due to over fitting.

	Malian news			Wikipedia		
	Written	Oral	Overall	Written	Oral	Overall
Exact meaning	0.830	0.770	0.840	0.870	0.530	0.730
Literalness	0.730	0.530	0.640	0.830	0.580	0.760
Standard Bambara	0.740	0.790	0.760	0.830	0.850	0.830
Highest BLEU Pair	0.408	0.363	0.408	0.645	0.377	0.645

Table 4.2: Malian news broadcast and Wikipedia article translation ratings and BLEU scores.

	Malian news			Wikipedia		
	Written	Oral	Overall	Written	Oral	Overall
Exact meaning	0.181	0.208	0.197	0.206	0.186	0.220
Literalness	0.130	0.298	0.243	0.238	0.211	0.244
Standard Bambara	0.234	0.178	0.207	0.171	0.068	0.144

Table 4.3: Score variance standard deviation

Written Laje ni kun tun ye ka hakili jakabo ke fere kuraw la

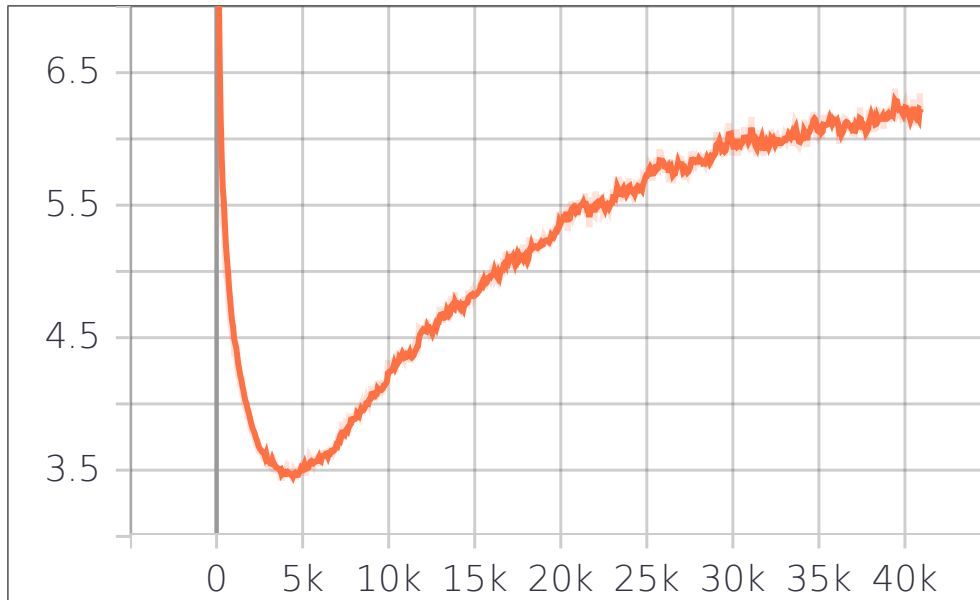


Figure 4.4: Character level: the perplexity of validation following the same trend as the validation loss.

banbaanciw juguya la miniw be ka ciyenni ke Saheli konononona la

Oral A kun tun ye ka miriya kuraw ta ka banbaanciw kelesi sira kan o mun bi ka ke sababu ye ka fagali caman ke saheli kononana

The highest scoring BLEU pairs in all but one of the aligned translations from the news source were between oral and written translation methods. In the one remaining case written-to-written and written-to-oral pairs had approximately the same high BLEU scores, the scores being the highest from among all the news source translations.

The translations of the Wikipedia article show that the meaning of the text was captured much better in the written-to-written translations. With only one exception, the highest scoring BLEU pairs were the written-to-written translations. These results suggest that written-to-written translation may be best for more complex texts while oral translations works well on simple texts.

Due to the political unrest in Mali, we unfortunately were not able to carry out with the crowdsourcing part of the thesis. We are waiting on our

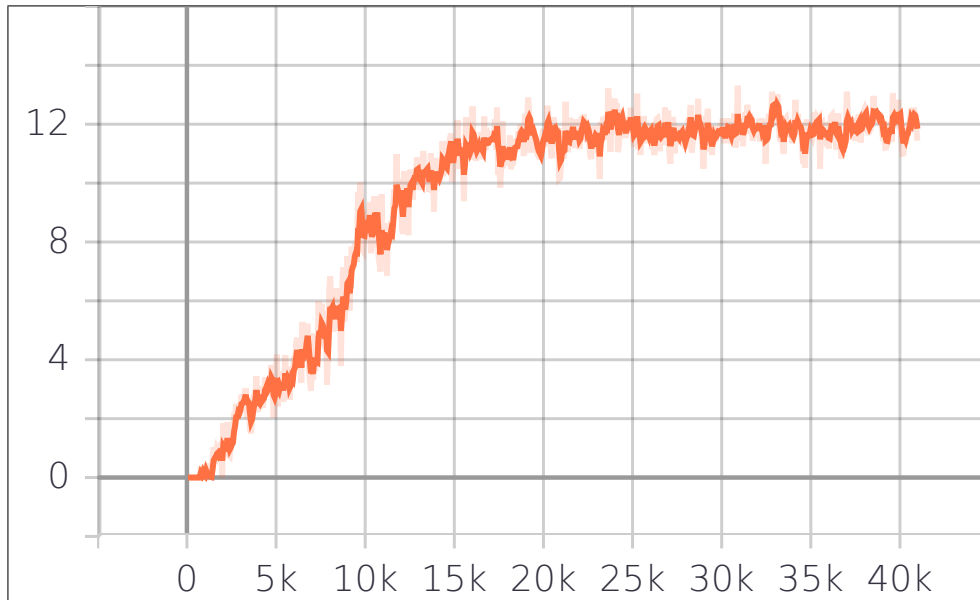


Figure 4.5: Character level: the validation score is the same as the BLEU score.

IRB (Internal Review Board) approval.

At the end of our pilot study we had 2,021 sentences that we believed were correctly aligned. We split the data into 80% and 20% for training and validation respectively. We conducted an experiment using an LSTM encoder-decoder architecture model based on Tensorflow 2.0 tutorial on neural machine translation with attention¹ where we the original languages were replaced with Bambara-English to translate from Bambara to English.

The translation quality was poor, but the generated attention plot provided important insights into the performance of the model on the dataset. It showed which parts of the source language sentence had the model's attention while translating into the target language. The results from the experiment were encouraging despite the fact that all the sentences translated with the trained model yielded a wrong translations.

Example one, the input is: A bɛ malo sɛnɛ.; the prediction is:

¹https://www.tensorflow.org/tutorials/text/nmt_with_attention

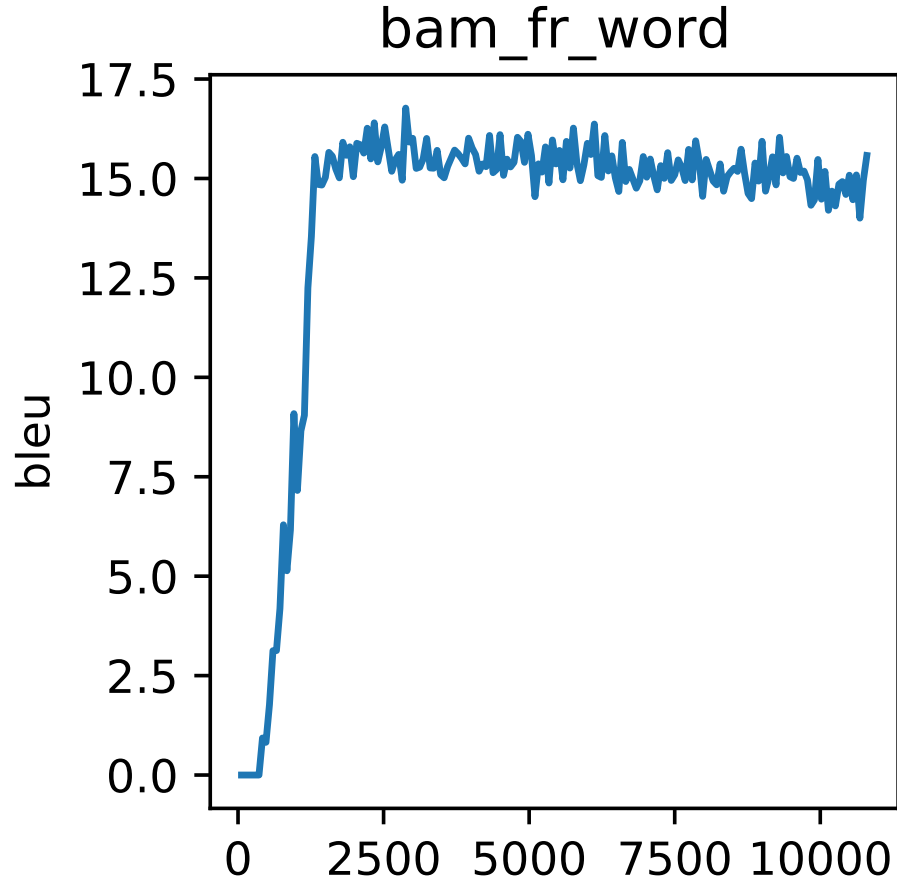


Figure 4.6: Plot of the validation data: Bambara to French Transformer model at word level.

the woman is a good idea.; the correct reference is: He farms rice.
or She farms rice.

Example two, the input is: I yɛɛ ka taa o ɲɔŋ kɛ !; the prediction is: to the child; the correct reference is: To go do the same

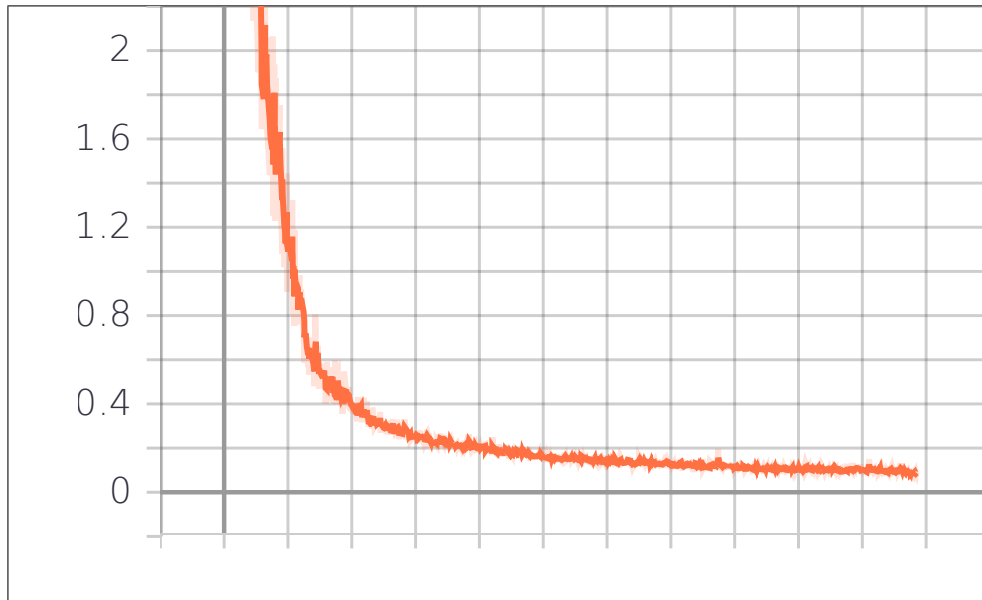


Figure 4.7: Word level: the training loss decreasing.

thing !

Example three, the input is: soso faga; the prediction is: to the child; the correct reference is: to kill mosquitoes.

Example four, the input is: layidu ta.; the prediction is: the woman is a good ideaa.; the correct reference is: make a promise

No tuning or fine tuning was conducted on the LSTM model. The training took about 40 minutes to train the model without GPU.

This experiment left some room for us to experiment further with some of the hyper-parameters of the model and fine tune it. Further digging into the data, we discovered that the data was not “good” enough to get any useful translation out of it. Additionally, in the tutorial there were no mechanisms to assess the performance of the model, to plot the results, and to log the training and validation processes. We struggled fine-tuning the model to output meaningful translations. We were unable to get satisfactory results using up to 50 epochs. With the use of a GPU, we were able to train over many epochs

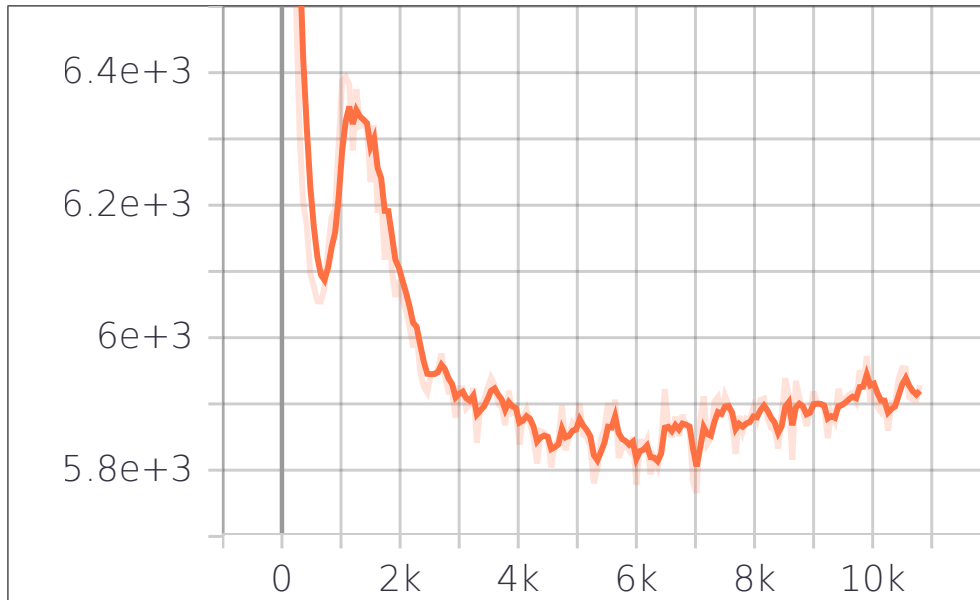


Figure 4.8: Word level: the validation loss was decreasing but start increasing probably due to over fitting.

but the results remained unsatisfactory. We pre-processed the data further as explained in section 3.1. We decided moving forward to utilise JoeyNMT [53] which has plotting mechanism already in place and logs everything during training.

Moving to JoeyNMT did solve most of the issues. Additionally, with the use of a GPU, we were able to train over many epochs and the results improved.

We believe that the data was still noisy to obtain good results. Taking time to go through the entire dataset manually and cleaning it one line at a time was rewarding. We plan to acquire a much larger quantity of data through our collaboration with INALCO to investigate if more data will increase the BLEU score. Additionally, we ran into some “out of memory” issues mainly due to hardware limitations, we will seek to use a more robust computing system.

Despite our challenging circumstances, we were able to make progress, establish collaboration and move forward.

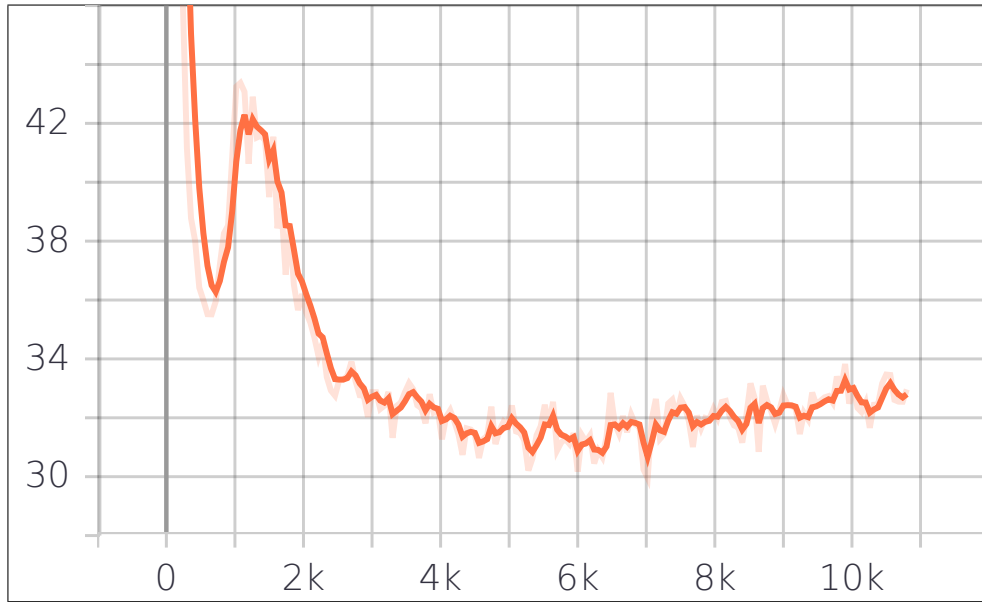


Figure 4.9: Word level: the perplexity of validation following the same trend as the validation loss.

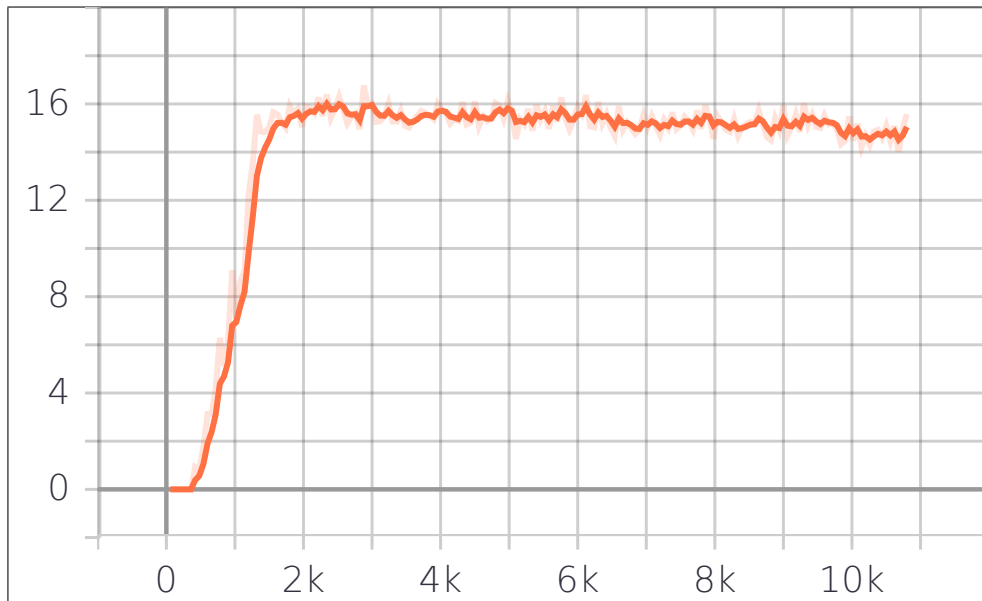


Figure 4.10: Word level: the validation score is the same as the BLEU score.

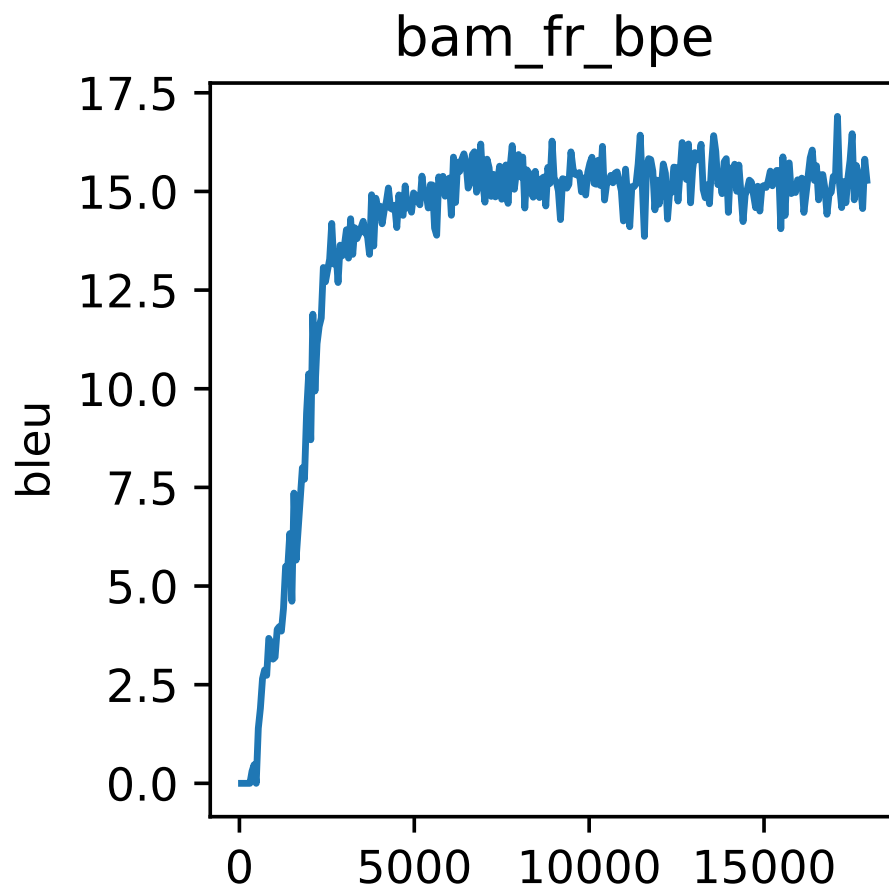


Figure 4.11: Plot of the validation data: Bambara to French Transformer model at BPE level.

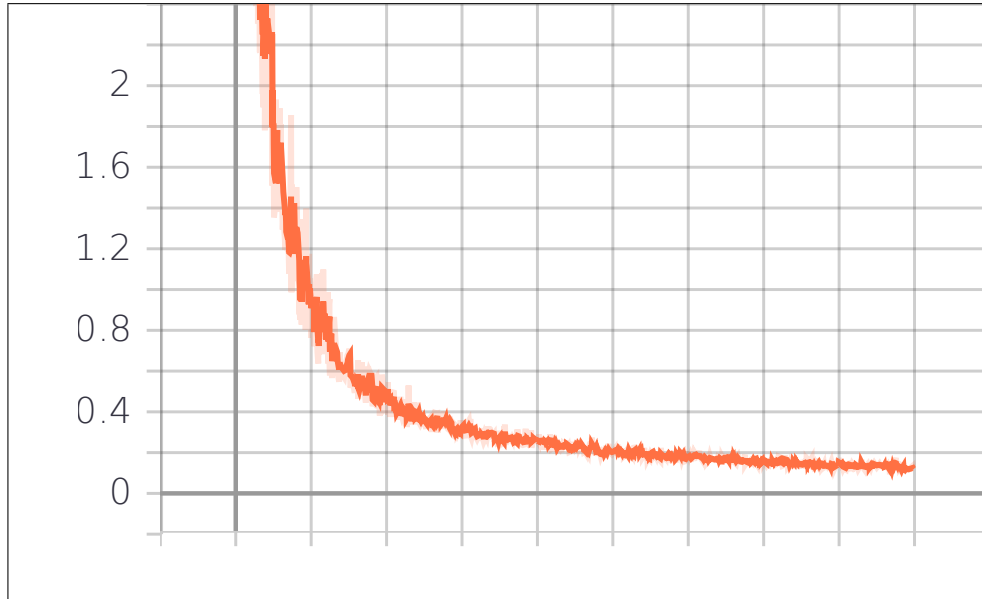


Figure 4.12: BPE level: the training loss decreasing.

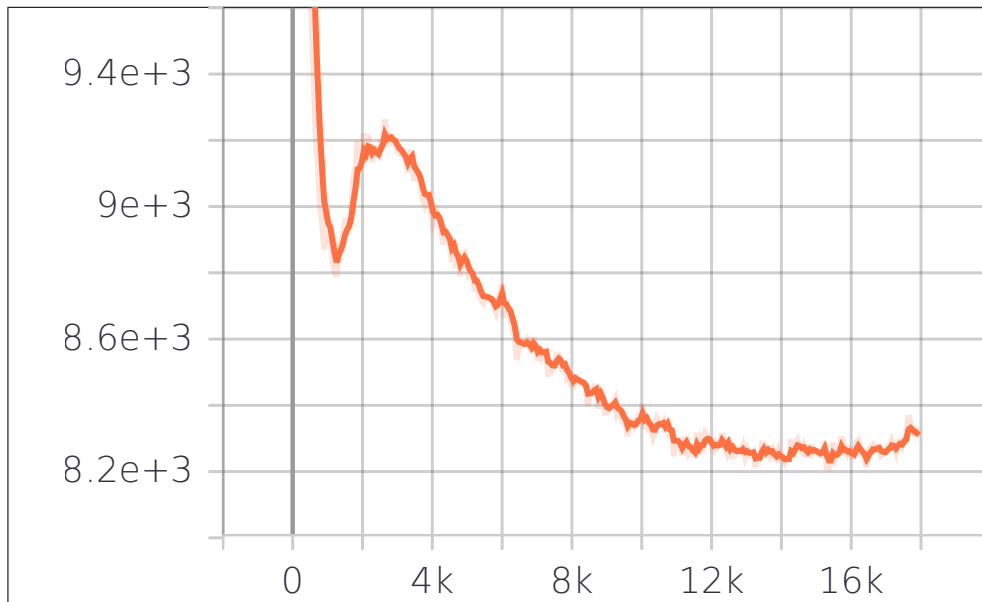


Figure 4.13: BPE level: the validation loss was decreasing but start increasing probably due to over fitting.

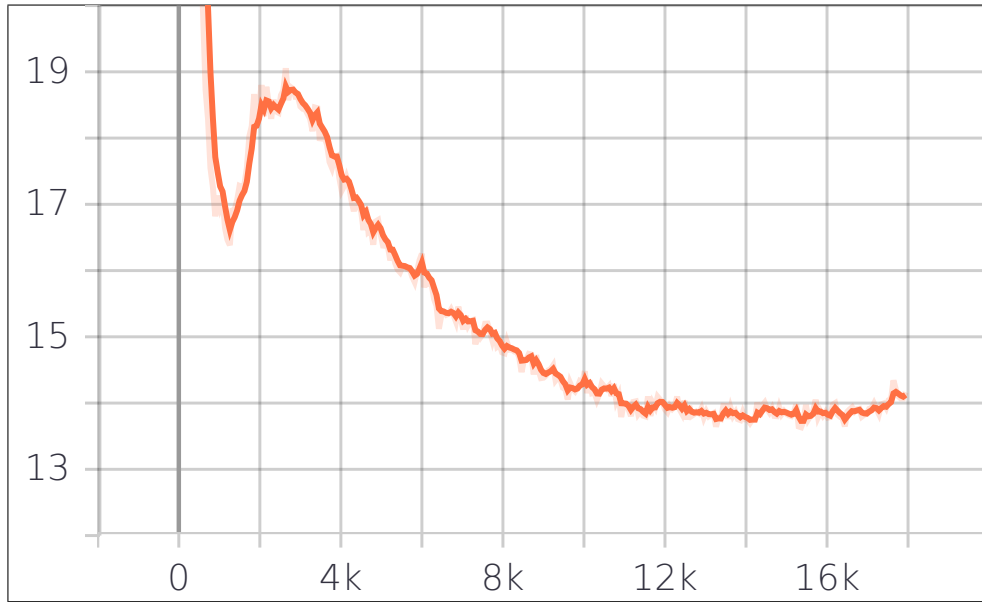


Figure 4.14: BPE level: the perplexity of validation following the same trend as the validation loss.

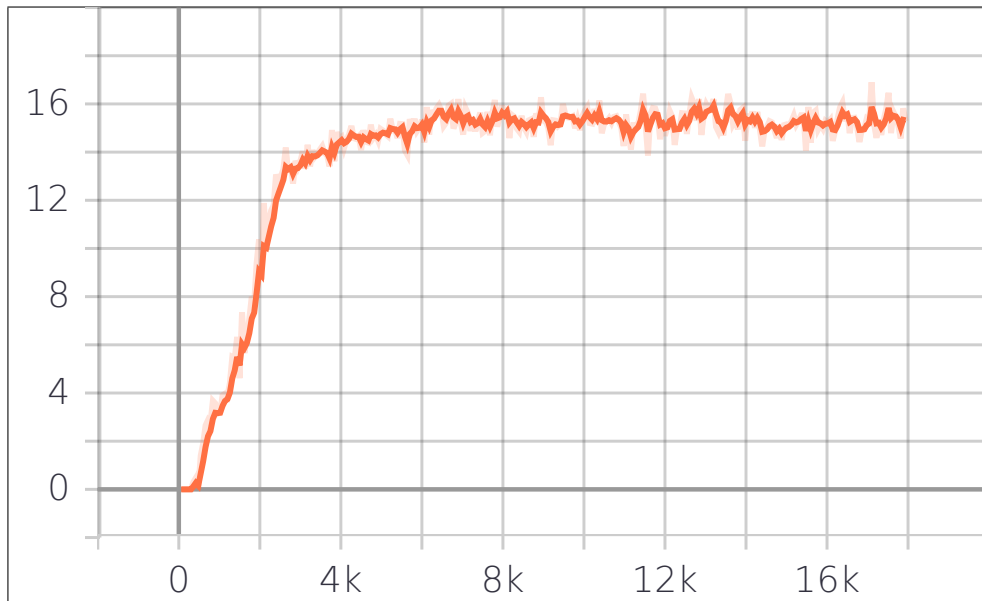


Figure 4.15: BPE level: the validation score is the same as the BLEU score.

Chapter 5

Future Work

During the course of this thesis we submitted three grant proposals in which we presented some concrete applications of how our work could prove useful in practical applications. The submissions were in the context of the COVID-19. We included them in Appendix B.

We will continue experimenting with hyper-parameters of our models. We will explore multi-lingual machine translation [4, 7] since we have English, French, and Bambara data.

We aim at exploring some specific characteristics of Bambara itself such as its tonality [23, 24], its morphology [94, 96] to utilize them in MT.

We will start the crowdsourcing part of the project as soon as we get our IRB approval.

Chapter 6

Conclusion

We undertook data discovery. We prepared the data through data preparation methods (e.g. cleaning, aligning, etc.). We fine-tuned hyper-parameters of our MT models. We set up our crowdsourcing pipeline.

Despite some initial setbacks, our results conclude that NMT for Bambara may be viable with a BLEU score of 17.5 from our best performing model, specially now that we have access to INALCO's dataset.

The future looks bright, we took the first step in the right direction with the right team of collaborators.

We are confident that the next iteration of the project with an increased focus on the data and the use of sophisticated crowdsourcing technologies will prove to be effective in collecting and preparing quality data for the training of our MT models. Finally, there is room for further experiments with our models' hyper-parameters to tailor an MT model that is well-suited to our needs or to build a model entirely from scratch.

Bibliography

- [1] Jade Z Abbott and Laura Martinus. Towards neural machine translation for african languages. *arXiv preprint arXiv:1811.05467*, 2018.
- [2] Wiehan Agenbag and Thomas Niesler. Automatic sub-word unit discovery and pronunciation lexicon induction for asr with application to under-resourced languages. *Computer Speech & Language*, 57:20–40, 2019.
- [3] 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*, pages 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics.
- [4] Roei Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. *Proceedings of the 2019 Conference of the North*, 2019.
- [5] Vicent Alabau, Luis A Leiva, Daniel Ortiz-Martínez, and Francisco Casacuberta. User evaluation of interactive machine translation systems. In *Proc. EAMT*, pages 20–23, 2012.
- [6] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).
- [7] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu.

Massively multilingual neural machine translation in the wild: Findings and challenges, 2019.

- [8] Ron Artstein. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer, 2017.
- [9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [10] Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, 2019.
- [11] Guy Belloncle. Use of the bambara language in training young people: An experiment in rural mali. *Prospects*, 10(1):107–115, 1980.
- [12] Cathy Berthouzoz. Contextual resolution of global ambiguity for mt. 1999.
- [13] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100, 2014.
- [14] Laurent Besacier, V-B Le, Christian Boitet, and Vincent Berment. Asr and translation for under-resourced languages. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.
- [15] Laurent Besacier, Viet Bac Le, C Boitet, and Vincent Berment. Asr and translation for under-resourced languages. volume 5, pages V – V, 06 2006.
- [16] Charles Bird et al. An ka bamanankan kalan: Beginning bambara. 1977.
- [17] Daren C Brabham. *Crowdsourcing*. Mit Press, 2013.

- [18] Michael Carl, Silke Gutermuth, and Silvia Hansen-Schirra. Post-editing machine translation. *Psycholinguistic and cognitive inquiries into translation and interpreting*, 115:145, 2015.
- [19] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
- [20] Heeyoul Choi, Kyunghyun Cho, and Yoshua Bengio. Fine-grained attention mechanism for neural machine translation. *Neurocomputing*, 284:171–176, 2018.
- [21] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [22] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. In *International conference on machine learning*, pages 2067–2075, 2015.
- [23] Karen Courtenay. Oli the nature of the bambara tone system. *and the African Studies Center The University of California, Los Angeles*, 5(3):303, 1974.
- [24] Christopher Culy. The complexity of the vocabulary of bambara. *Linguistics and philosophy*, 8(3):345–351, 1985.
- [25] Nic J De Vries, Marelle H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard, and Alta De Waal. A smartphone-based asr data collection tool for under-resourced languages. *Speech communication*, 56:119–131, 2014.
- [26] Barsha Deka, Joyshree Chakraborty, Abhishek Dey, Shikhamoni Nath, Priyankoo Sarmah, SR Nirmala, and Samudra Vijaya. Speech corpora of under resourced languages of north-east india. In *2018 Oriental COCOSDA-International Conference on Speech Database and Assessments*, pages 72–77. IEEE, 2018.

- [27] Rodolfo Delmonte. Deep 8: shallow linguistically based parsing parameterizing ambiguity in. *UG and External Systems: Language, brain and computation*, 75:335, 2005.
- [28] Joseph DeStefano. *Community-based primary education: Lessons learned from the Basic Education Expansion Project (BEEP) in Mali*. Number 15. USAID, Bureau for Africa, 1996.
- [29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [30] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.
- [31] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075, 2019.
- [32] Bonaventure F. P. Dossou and Chris C. Emezue. Ffr v1.0: Fon-french neural machine translation, 2020.
- [33] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- [34] Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. Using mooses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, 2008.
- [35] Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

- [36] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*, 2017.
- [37] Babacar Fall. Ict in education in mali. 2010.
- [38] Ben Foley, Alina Rakhi, Nicholas Lambourne, Nicholas Buckeridge, and Janet Wiles. Elpis, an accessible speech-to-text tool. In *INTERSPEECH*, pages 4624–4625, 2019.
- [39] Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144, 2011.
- [40] Mark Ginsburg, Don Adams, Thomas Clayton, Martha Mantilla, Judy Sylvester, and Yidan Wang. The politics of linking educational research, policy, and practice: The case of improving educational quality in ghana, guatemala and mali. *International Journal of Comparative Sociology*, 41(1):27–47, 2000.
- [41] Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254, 2012.
- [42] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Crowd-sourcing of human judgments of machine translation fluency. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 16–24, Brisbane, Australia, December 2013.
- [43] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.
- [44] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

- [45] Ann Irvine and Chris Callison-Burch. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the eighth workshop on statistical machine translation*, pages 262–270, 2013.
- [46] Ye Jia, Ron J Weiss, Fadi Biadisy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.
- [47] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*, 2020.
- [48] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyang Jiang, Masao Someki, Nelson Enrique Yalta Soplín, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019.
- [49] Tom Kocmi and Ondřej Bojar. Trivial transfer learning for low-resource neural machine translation. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018.
- [50] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.
- [51] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.
- [52] Melita Koletnik Korošec. Applicability and challenges of using machine translation in translator training. *ELOPE: English Language Overseas Perspectives and Enquiries*, 8(2):7–18, 2011.
- [53] Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. Joey NMT: A minimalist NMT toolkit for novices. *EMNLP-IJCNLP 2019: System Demonstrations*, Nov 2019.
- [54] Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. Correct me if you can: Learning from error corrections and markings, 2020.

- [55] Antonio-L Lagarda, Vicent Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Díaz-de Liaño. Statistical post-editing of a rule-based machine translation system. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 217–220, 2009.
- [56] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.
- [57] Michael Leventhal, Allahsera Tapo, Sarah Luger, Marcos Zampieri, and Christopher M. Homan. Assessing human translations from french to bambara for machine learning: a pilot study, 2020.
- [58] Shasha Liao, Cheng Wu, and Juan Huerta. Evaluating human correction quality for machine translation from crowdsourcing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 598–603, Hissar, Bulgaria, September 2011. Association for Computational Linguistics.
- [59] Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*, 2015.
- [60] Tengfei Liu, Shuangyuan Yu, Baomin Xu, and Hongfeng Yin. Recurrent networks with attention and convolutional networks for sentence representation and classification. *Applied Intelligence*, 48(10):3797–3806, 2018.
- [61] Laura Martinus and Jade Z. Abbott. A focus on neural machine translation for african languages. *CoRR*, abs/1906.05685, 2019.
- [62] Yolande Miller-Grandvaux. Usaid and community schools in africa: The vision, the strategy, the commitment. In *Community Schools in Africa*, pages 133–153. Springer, 2007.
- [63] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, pages 1–26, 2020.

- [64] Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- [65] Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Anisia Katinskaia, Anabela Barreiro, Lavinia Aparaschivei, and Yaakov HaCohen-Kerner. Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 268–278, Marseille, France, May 2020. European Language Resources Association.
- [66] MD Okpor. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159, 2014.
- [67] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. Masakhane – machine translation for africa, 2020.
- [68] Martha Palmer, Owen Rambow, and Alexis Nasr. Rapid prototyping of domain-specific machine translation systems. In *Conference of the Association for Machine Translation in the Americas*, pages 95–102. Springer, 1998.
- [69] Chandra Pandey, Zina Ibrahim, Honghan Wu, Ehtesham Iqbal, and Richard Dobson. Improving rnn with attention and embedding for adverse drug reactions. In *Proceedings of the 2017 International Conference on Digital Health*, pages 67–71, 2017.
- [70] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings*

of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

- [71] XML Schema Part. 2: Datatypes, 2001.
- [72] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019.
- [73] Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. Interactive neural machine translation. *Computer Speech & Language*, 45:201–220, 2017.
- [74] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. Bpe-dropout: Simple and effective subword regularization, 2019.
- [75] Colin Raffel and Daniel PW Ellis. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756*, 2015.
- [76] Reinhard Rapp, Serge Sharoff, and Pierre Zweigenbaum. Recent advances in machine translation using comparable corpora. *Natural Language Engineering*, 22(4):501–516, 2016.
- [77] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, pages 3171–3180, 2019.
- [78] Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319, 2007.

- [79] Christian Schmidt and Lenka Zdeborová. Dense limit of the dawid–skene model for crowdsourcing and regions of sub-optimality of message passing algorithms. *Journal of Physics A: Mathematical and Theoretical*, 53(12):124001, 2020.
- [80] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
- [81] Ingse Skattum. The introduction of the national languages into the educational system of mali: objectives and consequences of the reform. *Mande Studies*, pages 95–109, 2006.
- [82] Harold Somers. Example-based machine translation. *Machine translation*, 14(2):113–157, 1999.
- [83] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.
- [84] Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424, 2018.
- [85] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):1–6, 2020.
- [86] Amazon Mechanical Turk. Amazon mechanical turk. *Retrieved August, 17:2012*, 2012.
- [87] Guido Van Rossum et al. Python programming language. In *USENIX annual technical conference*, volume 41, page 36, 2007.
- [88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

- [89] Ayesha Yaqub Vawda and Harry Anthony Patrinos. Producing educational materials in local languages: Costs from guatemala and senegal. *International Journal of Educational Development*, 19(4-5):287–299, 1999.
- [90] Valentin Vydrin, Kirill Maslinsky, Jean-Jacques Méric, and A Rovenchak. Corpus bambara de référence, 2011.
- [91] Valentin Vydrin, Andrij Rovenchak, and Kirill Maslinsky. Maninka reference corpus: A presentation. In *TALAf 2016: Traitement automatique des langues africaines (écrit et parole)*. Atelier JEP-TALN-RECITAL 2016-Paris le, 2016.
- [92] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [93] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. A survey of crowdsourcing systems. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 766–773. IEEE, 2011.
- [94] Francesco Zappa. When arabic resonates in the words of an african language: Some morphological and semantic features of arabic loanwords and calques in bambara. In *The word in Arabic*, pages 229–249. Brill, 2011.
- [95] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.
- [96] Arnold M Zwicky and Geoffrey K Pullum. Plain morphology and expressive morphology. In *Annual Meeting of the Berkeley Linguistics Society*, volume 13, pages 330–340, 1987.

Appendices

Appendix A

Additional Resources

Bellow are some additional resources related to our work.

A.1 Models Detail

Bellow are the details of the models as per JoeyNMT [53] configuration file during the training. For the three experiments we ran on the same dataset at character, word, and BPE levels. Only the parameter `cfg.data.level` changed to match the current level being considered e.g., `char`, `word`, or `bpe`.

A.2 Useful Links

1. Our repository: https://github.com/israaar/mt_bambara
2. JoeyNMT: <https://github.com/joeynmt/joeynmt>
3. subword-nmt: <https://github.com/rsennrich/subword-nmt>

cfg.name	bam-to-fr-bpe
cfg.data.src	bpe.bam
cfg.data.trg	bpe.fr
cfg.data.train	data/train
cfg.data.dev	data/dev
cfg.data.test	data/test
cfg.data.level	bpe
cfg.data.lowercase	FALSE
cfg.data.max-sent-length	100
cfg.data.src-voc-min-freq	0
cfg.data.src-voc-limit	4000
cfg.data.trg-voc-min-freq	0
cfg.data.trg-voc-limit	4000
cfg.testing.beam-size	5
cfg.testing.alpha	1
cfg.training.random-seed	42
cfg.training.optimizer	adam
cfg.training.normalization	tokens
cfg.training.adam-betas	[0.9, 0.999]
cfg.training.patience	5
cfg.training.learning-rate-factor	0.5
cfg.training.learning-rate-warmup	1000
cfg.training.decrease-factor	0.7
cfg.training.loss	crossentropy
cfg.training.learning-rate	0.0003
cfg.training.learning-rate-min	1.00E-08
cfg.training.weight-decay	0
cfg.training.label-smoothing	0.2
cfg.training.batch-size	1024
cfg.training.batch-type	token
cfg.training.eval-batch-size	256
cfg.training.eval-batch-type	token
cfg.training.batch-multiplier	1
cfg.training.early-stopping-metric	eval-metric
cfg.training.epochs	150
cfg.training.validation-freq	60
cfg.training.logging-freq	10
cfg.training.eval-metric	bleu
cfg.training.model-dir	models/bam-to-fr-bpe
cfg.training.overwrite	TRUE
cfg.training.shuffle	TRUE
cfg.training.use-cuda	TRUE
cfg.training.max-output-length	100
cfg.training.print-valid-sents	[0, 1, 2, 3]
cfg.training.keep-last-ckpt	3
cfg.model.initializer	xavier
cfg.model.bias-initializer	zeros
cfg.model.init-gain	1
cfg.model.embed-initializer	xavier
cfg.model.embed-init-gain	1
cfg.model.tied-embeddings	FALSE
cfg.model.tied-softmax	TRUE
cfg.model.encoder.type	transformer
cfg.model.encoder.num-layers	6
cfg.model.encoder.num-heads	4
cfg.model.encoder.embeddings.embedding-dim	256
cfg.model.encoder.embeddings.scale	TRUE
cfg.model.encoder.embeddings.dropout	0.2
cfg.model.encoder.hidden-size	256
cfg.model.encoder.ff-size	1024
cfg.model.encoder.dropout	0.3
cfg.model.decoder.type	transformer
cfg.model.decoder.num-layers	6
cfg.model.decoder.num-heads	4
cfg.model.decoder.embeddings.embedding-dim	256
cfg.model.decoder.embeddings.scale	TRUE
cfg.model.decoder.embeddings.dropout	0.2
cfg.model.decoder.hidden-size	256
cfg.model.decoder.ff-size	1024
cfg.model.decoder.dropout	0.3

Table A.1: Model details for BPE data level translation.

Appendix B

Proposals

B.1 ISDB Proposals

Machine translation for rapid dissemination of COVID-19-related health information

<https://www.isdb-engage.org/en/profile/dashboard>

100 word executive summary

Beyond hospital beds or ventilators, the most critical capacity for public health is information. Yet only 20% of Malians speak a highly-resourced language like French (a main language of COVID-19-related information). This leads to barriers in coordination and action, breeding misinformation.

We propose a machine translation (MT) intervention. MT requires translated texts to learn from; Malian languages lack these. We integrate state-of-the-art MT with novel crowdsourcing infrastructure that engages Malians, driving learning while providing direct economic aid to participants. This will strengthen institutional relationships between Malian and US universities and a Malian startup incubator, creating sustainable infrastructure for long-term growth.

(A) Identification of STI Capacity Building Challenges (assessments/needs): Please identify what key STI capacities already exist in a particular IsDB MC (with priority given to LDMCs) and what additional capacities will be needed (at the three levels - individual, institutional and enabling environment) to address a specific development challenge related to one of the following SDGs: No Hunger (food security), good health and well-being, quality education, clean water and sanitation, affordable & clean energy, industry, innovation and infrastructure?

This project addresses the good health and wellness SDG, with subsidiary impacts in education and industry and innovation. The Malian government has long recognized that disseminating information in a foreign language—French, mastered by only 20% of the population—is a major barrier to development in all sectors, including healthcare. This in the midst of the COVID-19 pandemic has created an information vacuum that limits positive responsiveness and breeds rumors and misinformation. From <https://theowp.org/reports/why-the-western-system-of-covid-19-response-wont-work-in-africa/>: ‘The United Nations announced a ‘Verified’ initiative to, “combat the growing scourge of COVID-19 misinformation by increasing the volume and reach of trusted, accurate information.” ... Information can be a highly powerful tool to compensate for African states’ relatively weak infrastructure and governance. The spreading of factual information can be a catalyst for local innovation once the mechanisms of viral diseases are understood.’

The capacity needs are thus, at the individual level: greater access to authoritative healthcare information in local languages; at the institutional level: power to translate much greater quantities of public health

information from higher resourced languages like French and university-level expertise in artificial intelligence; at the enabling environment level: more employment and entrepreneurial opportunities for long term growth and development—indeed, unemployment of university graduates in Mali currently exceeds 25% (pre-COVID), according to official figures.

(B) Innovativeness: Please describe your proposal, articulating clear and specific capacity building objective(s). with suggested activities.

Please clarify how these activities are expected to address the identified capacity gaps/challenges. In doing so, kindly highlight the innovation and creativity behind the proposed capacity building intervention describing how your intervention is expected to boost STI capacities in MCs to address one of the above-mentioned development challenge(s). How is this different from other capacity building efforts?

Why is this an improvement?

Objective-1 Create and deploy a practical machine translation (MT) system for public health information about COVID-19. This innovative approach to translating critical healthcare information into Bambara boosts translation capacity by a nearly unlimited factor, at a fraction of the cost of manual translation. MT, particularly for underresourced languages, is never perfect, so another innovation is to use crowdsourcing to engage a hierarchy of language experts, medical practitioners, and ordinary Malians to ensure that translations convey critical information and combat misinformation. This differs from other translation efforts because it uses machine, not human, translators.

Objective-2 Build crowdsourcing infrastructure that engages Malians across social and medical strata to provide the feedback needed to drive effective MT, while providing direct economic aid to participants. The innovation is that it binds the technical success of Objective 1 to community engagement and direct economic aid and training to this objective's beneficiaries, largely to individuals who need minimal specialized knowledge to begin.

Objective-3 Establish an academic-entrepreneurial partnership between RIT in the United States and USTTB and RobotsMali-CETACE in Mali, to provide well-trained professionals in underresourced-language MT for healthcare informatics. This would be the first collaboration at the academic level between Bamako and US AI experts. This is different from other efforts because it establishes an international channel for AI knowledge exchange. This is an improvement because effective AI systems require internationally-vetted and local-domain expertise.

Objective-4 Seed Malian entrepreneurship with MT and crowdsourcing technology that will inspire further innovations that contribute to social and economic development.

(C) Impact and outreach: Please specify key outputs and outcomes of the proposed capacity building intervention, describing its intended impact on the various elements of the STI ecosystem?

How will impact be measured?

How many people or institutions do you expect to impact directly and indirectly in the short, medium and long term?

What sort of capacity improvements would the intervention provide?

Outcome-output-1 Machine learning pipeline to translate on demand to/from written and spoken Bambara, tailored to respond to COVID-19 through native language information resources. Bambara-French translation also provides international exposure to traditional Malian healthcare.

Measurements-1 Number of organisations using it (four short-term, 10-20 medium-term, 100s long-term) quantity and quality of material produced (3-10 corpora short-term, 500-1000 medium-term, 100K-1M long-term), diffusion to health care workers (6 short-term, 50-200 medium term, 1000-10K long-term) and general population (50-1000 short-term, 44M Mande speakers medium-term, 1.2B Africans long-term). Increase in accurate public understanding of best healthcare practices, measured by surveys. Improved public health response, measured by national metrics.

Output-outcome-2 Crowdsourcing infrastructure for engaging human expertise in gathering and labeling/cleaning data, and training and evaluating machine learning performance. Learning, employment, and nation-building contribution. Upskilling opportunity for 100-200 students, via creating crowdsourcing tools and AI labeling tasks. These transferable resources will create a new industry.

Measurements-2 Youth employment and engagement measured by number of persons engaged, trained, and compensated (5-10 short-term, 100-200 medium-term, 10-100K long-term).

Output-outcome-3 Collaborative institutional partnerships between US (RIT), Malian university system (USTTB), and Malian technology and entrepreneurship agencies (RobotsMali). Training of doctoral level Malian AI researchers on crowdsourcing and entrepreneurial offshoots of the project.

Measurements-3 Placement of doctoral students from USTTB at RIT (1 short-term, 2-4 medium term) and other institutions (10s-100s long term), US-Mali knowledge sharing measured by joint publications and doctoral theses (5-15 short-term, 10-60 medium term, 50-1000s long-term).

(D) Estimated Cost & Implementation Plan: Please list the expected deliverables, activities and estimated cost for the program. How much is being requested from the IsDB and what will be the contribution of other donors?

The plan should describe how the various stakeholders in the recipient country will be engaged. Please note that IsDB support under this category will not cover the total cost of the proposed capacity building intervention, as it is considered a contribution to the total cost.

Pilot deliverables Initial data set, functional machine translation engine and crowdsourced labor site, and processes for data cleaning, validation, and crowdsourcing, based on state-of-the-art methodologies used in industry and academia.

Total funding \$175,000 (\$25,000 from IsDB). Other sources: Fulbright Scholarships, RIT, RobotsMali, and Crowdfunding.

COVID and early-deliverables Well-performing automated translator in medical and COVID domains, adequate in other areas critical for Malian social and economic development. Enough data has been collected and validated to successfully train the system. Local Malian language expert labor pool recruited and trained in data annotation/translation.

Total funding \$100,000 (\$25,000 from IsDB). Other potential sources: end users such as the Ministry of Health, Muso, Mali Health, MSF.

Full system deliverables More generalised and performant translation system deployed in cloud, desktop, and app form. Automated system supports written and oral input and output.

Total funding \$100,000 (\$75,000 from IsDB). Other potential sources: end users such as the Ministry of Scientific Research, UNESCO, UNICEF, and Orange.

Entrepreneurial development deliverables Incubation of 3 companies built around applications making use of technologies developed in this project. Crowdsourced labor companies that hire and train local language experts to perform essential AI data annotation and translation tasks have been extremely successful in other communities around the world including in India, Nigeria, and Portugal.

Total funding \$300,00 (\$25,000 from IsDB). Other potential sources: VC funds and international programs directed to entrepreneurship and technical skills development.

(E) Sustainability: Please describe how the effort will last beyond the life of this grant.

What are the special measures embedded in the design of the capacity building intervention that will ensure the continuity of its impact/results in the long term?

RobotsMali, a high-tech incubator, has successfully VC-funded 11 African startups. RobotsMali is partially funded by the Malian government and UNESCO, and worked with the World Bank on capacity-building projects in Mali and Burkina Faso. The only Sahel-based technical resource center for development, RobotsMali's participation will guarantee the resulting revenue-generating enterprises have the infrastructure necessary to support long-term innovation. This will establish a knowledge exchange

pipeline between RIT, a internationally respected US-based university, and will strengthen Malian education structures. Professor Homan of RIT is an expert in building information systems that engage and learn from populations, particularly those who are underrepresented (such as Malians) in conventional information systems.

Malian and US-based collaborators ensuring impact continuity:

Assétou Foné Samaké Migan, PhD, former Minister of Higher Education and Scientific Research in Mali, Biologist, National Coordinator of the Malian Society of Applied Sciences, University of Sciences, Technology and Technical Education of Bamako (USTTB). Will liaise with the government and NGOs ensuring results are used by the government and international organisations in COVID response.

Amadou Koné, PhD, Researcher, USTTB. Head of Mali's only P3 laboratory, playing a leading role in the COVID-19 response. Will liaise between USTTB and the Malian COVID-19 response and RIT; ensuring the cooperation with RIT builds Malian scientific and educational capacity.

The Malian government's Academy of Languages (AMALAN), will ensure national language policy conformity.

Dr. Sarah Luger is an AI crowdsourced data expert and research scientist at OSV, the strategic innovation office of Orange.

(F) Scalability: Please describe how the effort will efficiently grow over time. How is this embedded in the plan?

This project's interdisciplinary team ensures the project aligns with long-standing national objectives, the right and necessity that Malians can use their national languages in all spheres of life, providing an automation technology to achieve that which has heretofore been unachievable. Based on its technology, innovation, and breadth, this project will penetrate every corner of Malian society, achieving, at a national level, the highest scaling possible throughout the institutions and entrepreneurial ecosystem of Mali.

The broad penetration of 4G and smartphones at all levels of Malian society also supports growth and community engagement. Thus, the infrastructure for nationwide diffusion is already in place.

This project addresses a world-wide challenge faced by many people: that of digital enfranchisement of populations speaking one of the world's 6000 under-resourced languages. This includes nearly the entire population of the African continent for whom official languages like English, French, and Portuguese are not native languages. Thus, the potential market and scale of the technology developed in the project is truly vast because this novel technology is subsequently applicable to other languages; an emergent AI technique called "transfer learning" provides strong evidence that they can be. The novel crowd-sourcing methodology developed by this project will also provide a framework reemployable for other languages. We established a channel for scaling our work beyond Malian borders through our participation in the

cross-African consortium of NLP projects, Masakhane. Through Masakhane we are sharing data, experience, and code across 17 countries representing 29 major, underresourced, African languages.

B.2 Google Proposals

Step 2

Proposal Description

In this section provide details of your proposal

Project Title *

Bayelemabaga (translation: translator)

Project Description *

Max 800 characters

This diversity-aimed project will focus on cleaning, aligning, and evaluating several found datasets that contain a mixture of Bambara, French, and English language text. With Google's assistance, crowdsourcing infrastructure that engages Malians can drive machine learning while providing direct economic aid to participants and valuable worker education and training, additionally enriching the community. We are particularly interested in translating health-related information from English and French to Bambara and preventing the spread of misinformation. This project will additionally strengthen institutional relationships between Malian and US universities, a Malian startup incubator, and government and non-government agencies, creating sustainable infrastructure for long-term growth.

Bayelemabaga is a collaborative initiative for people speaking the Bambara language from the Mande language family with the objective of making essential information accessible to development specialists in all fields through the use of Natural Language Processing, (NLP), technologies.

In the wake of COVID-19, By focusing on the area of public health, the aim of this project is to utilize languages that already have robust NLP resources and automated translation such as French and English to leverage translation between the most widely spoken language in Mali, Bambara. The initial objective is the creation of a high-quality written and oral Bambara-French-English translation system for health information and the effective deployment of this system by health organizations in Mali.

Given that all national languages of Mali, like the majority of the world's languages, suffer from a lack of digital linguistic resources necessary for implementing deep learning, a current state-of-the-art NLP technique, it is expected that the scientific derivatives of this project will be useful for many other language inclusion projects.

We propose to build machine learning tools with humans-in-the-loop to automatically translate such texts, thus impacting global health, particularly in low development regions, where improvements to healthcare are most critical. Machine translation of under-resourced languages is a hard problem and a rapidly-growing area in computational linguistics. This project will seek to exploit several unique features in this setting, namely (1) WWW text is embedded in xml, a

rich source of metadata; (2) medical discourse is relatively circumscribed, with a relatively small vocabulary and an abundance of technical, domain-specific, untranslated words; (3) perfect translation is not necessary, only that the critical healthcare information is conveyed. Then, we will leverage our team's human computation experience to develop novel methods that incorporate the insights of frontline health practitioners in evaluating the translations. Next, the evaluation will be based on the key standard of communicating the critical healthcare information and will feed subsequent quality evaluation data into the training loop.

Research Question *

Max 500 characters. Outline the main research question which will be guiding the research in this pilot project

Can crowdsourcing engage Malians with novice-to-expert levels of language and domain knowledge to build and evaluate translations that augment an automated MT system that is both a communication and development game changer for an entire nation? Bambara is a low resourced language, and any increase in digital resources in this essential, native language will facilitate health communication and combat misinformation. Moreover, these efforts may have cascading financial and development benefits.

- What is the best approach to dramatically increase translation capacity (the number of people who can participate in crowdsourcing translations and the speed of their work)?
- How can we best leverage community-driven crowdsourcing to drive effective MT, while providing direct economic growth to the community?
- How can community-driven crowdsourcing assure knowledge and technology transfer including worker upskilling?
- Technical domains such as healthcare present specific challenges to machine learning and often require translators with expert knowledge. How can we integrate such support in a crowdsourcing platform to best leverage that knowledge?
- What are the best evaluation methods for, respectively: assessing absolute fidelity of translations; assessing how well health-related translations convey the specific information they are intended to convey; and driving machine learning?

Step 3

Impact Criteria

Projects will be evaluated against four main criteria for real-world impact and social relevance, innovation, feasibility and openness:

Real-world Impact and Social Relevance *

Max 500 characters. Projects should demonstrate potential for significant positive impact in the real world, specifically focusing on delivering social relevance, community building, AI data for good, etc

If successful, this project will provide:

Increased capacity for rapidly translating public health information about public health crises and combating misinformation.

Malians direct economic support and skills training, community building opportunities, and a new route to access the digitally-savvy world.

Increased academic and entrepreneurial collaboration between Mali and the rest of the world.

Crowdsourcing is a method of combining small “microtasks” performed by individuals into a greater whole. This combines subtle, but crucial, human-decision making, such as word and phrase translation, with technological processes that stitch these human judgments together. Crowdsourcing is also a metaphor for community-building as the final MT algorithms are inherently built upon the decisions of many dozens to thousands of people. The process of learning to translate, annotate, and contribute helps workers gain new skills in human-computer interaction that may be used for other technology and AI projects. Human judgment data is a valuable resource in general and an even more valuable resource when produced for a specific domain that is poorly serviced by current technology. Crowdsourced human judgment data provided for and by Malians is a unique solution that could provide a watershed moment for digital accessibility and misinformation reduction. This project

Innovation *

Max 500 characters. Preferred projects will be innovative (use of technology in hybrid settings with the crowd, novel annotation model, novel quality control approaches, new incentives and engagement approaches, etc.) and transformative for a specific research field or application and its users.

Our mixture of language and experts and ordinary Malians provides an exciting domain to explore novel, complex crowdsourcing pipelines..

Technical success is tied to direct economic aid and training, yielding new opportunities to develop humane and effective incentives.

We will systematically explore new evaluations methods for: absolute fidelity of translations; how well health-related translations convey the specific information they are intended to convey; and driving machine learning.

Feasibility *

Max 500 characters. Project should demonstrate clear indicators and metrics, and set out key risks and mitigation steps.

Machine learning pipeline to/from Bambara via number of organisations using it, quantity and quality of translations. Engagement of human crowdworkers via number of persons engaged and compensated. Collaborative partnerships between US and Malian entities, training of doctoral level Malian researchers via number of doctoral students placed, number of publications.

Risks: political instability, low literacy, power cuts. Mitigations: US and Malian partnerships to increase stability and engagement.

Openness *

Max 500 characters. identify the ways the project will share the lessons learned, how will be the code and data open sourced, and how will it contribute to the improvement of the volunteer crowdsourcing ecosystem overall.

Social media posts that share the collaborative journey and corresponding results.

Github repository of code and documentation through Masakhane, an open organization and arguably the leading one dedicated to machine translation for African languages.

Corpora of aligned text following FAIR practices with metadata provided in scheme.org.

Word-of-mouth experiences from crowdsourced workers.

Vigorous participation in academic conferences and venues, including workshops, tutorials, and publications.

- We have already presented one paper at an academic workshop and hope to have several more ready for strong conferences by late Fall.
- We will leverage existing relationships with Malian government and education officials to widely disseminate translated news about Covid-19 and over public health crises to combat misinformation and promote positive healthcare practices.

Additionally, Our participation in their code repositories and corpora and wordbanks strengthens them and both amplifies our voices and directs it to those who have the most to benefit from our contributions.

Finally, the training received and the skill sets developed by the workers will enable them to be active members of the community while increasing the supply of local domain experts. Self-translation will open up the community to the world and the world to the community and dramatically improve the education system.

Step 4

Crowds & Data

In this section describe what crowds do you plan to engage, in what annotation task and with what data

Target Crowd(s) *

Max 800 characters. Outline the characteristics of your target crowd, e.g. location, size, language group. Briefly describe how will you engage the crowd in the task, e.g. if it is an existing Crowdsourcing community, or if it is an external one, what campaigns do you foresee for the engagement and on boarding.

Maliens in Mali: estimated to 20M, of which 80% speak Bambara. Maliens abroad: estimated to 6M, of which 75% speak Bambara. We will begin with a small crowd of workers recruited by Mali-based coPI Leventhal. We presented a pilot study at the AfricaNLP workshop at ICLR on our engagement efforts. Leventhal has highly-placed contacts at the Malian National Education Center for Robotics and Artificial Intelligence, the Malian Ministry of Education, Higher Education and Scientific Research, and AMALAN (Malian Academy of Languages). MS student Allahsera is active in the Malian diaspora community in the US. From there, we hope to use their networks to broaden participation and Google resources to coordinate this engagement.

Target Data Collection *

Max 1600 characters. Briefly describe the task that you are aiming for, e.g. task description, sketch

Our primary task is to clean and align our existing data sets, and to provide a functional machine translation engine and integrated crowdsourcing site, and processes for continual data cleaning and validation, based on state-of-the-art methods used in industry and academia.

Beyond that, we will establish a well-performing (in terms of how informative the outputs are) automated translator in the healthcare domains (with an initial focus on COVID-19), that performs adequately in other areas critical for Malian social and economic development. We would additionally like to explore the use of Google resources to discover, collect, clean, and

align, new texts, and to help recruit and train a local pool of Malian language experts and nonexperts for data annotation, translation, and evaluation.

In focusing on the domain of public health, we aim to utilize languages that already have robust NLP resources and automated translation resources, such as French and English, to leverage translation to and from Bambara, the most widely spoken language in Mali. Machine translation of under-resourced languages is a hard problem and a rapidly-growing area in NLP. This project will seek to exploit several unique features in this setting, namely (1) WWW text is embedded in xml, a rich source of metadata; (2) medical discourse is relatively circumscribed, with a relatively small vocabulary and an abundance of technical, domain-specific, untranslated words; (3) perfect translation is not necessary, only that the critical healthcare information is conveyed.

What is the source data you will use in the project? *

Max 500 characters. All the data used in the pilot projects should be publicly available.

Please describe below what is the type and source of the data, e.g. images, text, annotation categories, etc

We currently have the following datasets:

Language	Medical data						
	bigrams	chapters	files	paragraphs	stopwords	trigrams	wordlist
Bambara	26,430	27	336	9,336	147	5,816	8,209
French	25,746	27	336	9,367	123	11,312	9,893
English	31,412	27	336	9,356	69	21,398	6,935

Language	Dictionary data			Aligned
	glosses	examples	combined	
Bambara	3,548	2,023	5,571	2,160
French	4,847	2,021	6,868	2,146
English	4,855	2,021	6,876	2,160

Link to an existing Crowdsourcing task *

Please specify to which existing Crowdsourcing Task are you linking to; If you are proposing a new task, select other. Also identify whether it is a web or mobile based task. Check out <http://crowdsourcing.google.com> for Crowdsourcing on the web, and download the Android app at

<https://play.google.com/store/apps/details?id=com.google.android.apps.village.boond>

Image Label Verification

Image Caption

Handwriting Recognition
Landmarks
Facial Expressions
Translation
Translation Validation
Image Capture
Sentiment Evaluation
Other:

Step 5

Project Category

The Open Crowdsourcing initiative seeks proposals for pilot projects that (1) propose tasks that have an impact in the real world, i.e. among others, the data collected will be in the public interest, useful for tackling a specific bias, bridging a specific diversity gap, or contributing to fairness or transparency of systems, (2) address human computation research challenges related to data collection, dataset quality, e.g. replication and reliability of results, and (3) work with open data and aim for open publishing of the project results

Select the Project Category for your proposal *
Please select all that apply to your project proposal.

Diversity aimed project
Dataset aimed project
Innovative Tasks project

Step 6

Optional Material

In this section you can provide any additional sketches or illustration of the crowdsourcing task that can help understand better the proposal. This is NOT intended to provide more textual description of your project. Please, limit only to visual and their brief explanations

Optional illustrative material

Provide any additional sketches or illustration of the crowdsourcing task that is proposed in this project. Please do not use this document to provide more textual description, and limit these files only to visual and their brief explanations. You can submit in either PDF, presentation, drawing or an image format

Step 7

Applicant

Provide contact information for the corresponding author of this application. By providing your contact information below, you consent to Google contacting you via email in regards to your application and proposal.

Academic Project Team *

Max 500 characters. Please include the lead PI for the team, list all team members and their background & roles

Christopher M. Homan, lead PI, Associate Professor, RIT, PHT 180 Initiative, coordinates all efforts and provides direct student mentorship.

Sarah Luger, Orange Silicon Valley, expert support on novel crowdsourcing for natural language.

Allahsera Auguste Tapo, RIT, Fulbright Scholar, prospective Ph.D. student, driving force behind this project.

Michael Leventhal, RobotsMali, worker recruitment, engagement, and training.

B.3 IRDC Proposals

Website:

<https://www.idrc.ca/en/funding/covid-19-global-south-artificial-intelligence-and-data-innovation-program>

Related:

<https://www.idrc.ca/en/news/webinar-series-gender-covid-19-and-work>

<https://www.idrc.ca/en/news/mobilizing-african-research-response-covid-19-pandemic>

1.3. If applying as a consortium, please list the name of all organizations and their country location. Note that, as outlined in the call document, IDRC expects the lead applicant to lead finances and administration.

Please provide text. [max. 250 words]

Rochester Institute of Technology (RIT), United States (LEAD APPLICANT)
University of Rochester Medical Center (URMC), United States
Université des Sciences, des Techniques, et des Technologies de Bamako (USTTB), Mali
Académie Malienne des Langues (AMALAN), Mali
RobotsMali, Le Centre National Collaboratif de l'Éducation en Robotique et en Intelligence Artificielle, Mali

2.1. Project Title

Enter a project title. (max. 250 characters)

Reaching People in the Language They Understand : Effective COVID-19 Outreach Using Automated Translation for Low-Resource Languages

2.2. Project Abstract

Provide a project abstract. (max. 400 words)

In a major public health crisis, information is a critical resource. Lack of accurate, trusted, and accessible information about COVID-19 may render ineffective policy and programs designed to stop the spread of the virus and to get the appropriate health care to the afflicted. Major barriers to public outreach exist in our focus country, the West African nation of Mali. Language is the most significant cause for this barrier. For historical reasons, Mali has an official language used by the government to communicate information in which the majority (80% according to

research) of the population is not functionally literate. While the use of the vehicular language of Mali, Bambara, as well as other national languages, is promoted by national law, policy and programs, public and private agencies lack the resources to make critical information available in these languages at scale. Our research aims to discover, develop, and provide methods and tools to address this problem for Bambara, and to create a base for applying our results to other national languages of Mali as well as other language communities in similar circumstances throughout the world.

We propose machine translation (MT) as a central element to the problem of getting critical information to Malians in a language that they understand. MT, however, requires translated texts to learn from and this is exactly what is lacking for the under-resourced languages of Mali. We integrate state-of-the-art MT with novel crowdsourcing infrastructure that engages Malians, driving learning while providing direct economic aid to participants. Our activities will, as broader impacts, strengthen institutional relationships between US universities and Malian institutions, creating sustainable infrastructure for long-term growth.

3.2. Please provide an overview of the research problem (s) you seek to address, and justification of how your proposed research might seek to address these challenges.

The research problem should consider the magnitude of local problems, contexts and needs around COVID-19, and the importance of the project for vulnerable populations (and on gender equality). It should identify key research questions.

The justification should outline how the proposed solution(s) will contribute to addressing the problems, indicate timelines, and highlight needs and challenges around building up research capacity. This section should also indicate if the proposed research builds on existing initiatives, or where it links to rapid response activities, including policies and implementation.

The research problem this project tackles is how to train an AI system to translate domain-specific information to a language which is severely under-resourced. The domain chosen for this project is COVID-19 and related health information as this information will have a short-term impact in supporting public initiatives to control spread of the virus and long-term impact on healthcare, one of the pervasive development issues in the target country, Mali. The target language, Bambara, is an under-resourced language for which acquiring training data represents the major, unsolved challenge. The project team combines US and Malian scientific expertise with a community-based approach to domain-specific crowd-sourcing on a massive scale.

Rumors, misinformation, and distrust of official sources are particularly widespread in Mali. The fact that the information is in French, a language of the elite, increases this distrust. From

<https://theowp.org/reports/why-the-western-system-of-covid-19-response-wont-work-in-africa/>: ‘The United Nations announced a “Verified” initiative to, “combat the growing scourge of COVID-19 misinformation by increasing the volume and reach of trusted, accurate information.” ... Information can be a highly powerful tool to compensate for African states’ relatively weak infrastructure and governance. The spreading of factual information can be a catalyst for local innovation once the mechanisms of viral diseases are understood.’ However, to be truly understood, trusted, and effective, that information must be in a language that Malians understand.

The Malian government has long recognized that disseminating information in a foreign language—French, mastered by only 20% of the population—is a major barrier to development in all sectors, especially healthcare. The 80% of the population that does not understand French consists disproportionately of rural and disenfranchised populations and of women who for economic and cultural reasons have significantly less access to education and opportunities to acquire literacy in French.

Official government policy mandates, for this reason, increased use of Malian languages in the public sphere, but little capacity has been developed due to lack of resources. This in the midst of the COVID-19 pandemic has created an information vacuum that limits positive responsiveness and breeds rumors and misinformation.

The capacity needs are thus, at the individual level: greater access to authoritative healthcare information in local languages; at the institutional level: power to translate much greater quantities of public health information from higher resourced languages like French.

RQ (research question) 1: How much can machine translation improve capacity for delivering actionable COVID-related public health information directly to Malians in a language they can understand?

In this project, we strive to prevent such abject consequences for Malian women by providing avenues for accurate Covid-related information *in a language they understand*, so that they can be empowered to make informed decisions about their health and that of their family members. Low female literacy rates, combined with the fact that 80% of the Malian population do not understand French, the dissemination language of the Malian government, leaves women most vulnerable to Covid-19 misinformation. This limits women’s ability to make autonomous decisions about their own bodies, the health of their children, and of their families. If women are limited in their knowledge of viral transmission modes, use of masks to minimize personal exposure, and implementation of movement restrictions, quarantine measures, and social distancing protocols to reduce community exposure, then efforts to curb the spread of the virus in some of the most vulnerable and disenfranchised Malian communities will be rendered moot.

The fact that a high percentage of Malian women do not understand health information disseminated in French has a wide impact in Malian society as matriarchs make critical decisions on the health care of family members. Only two out of every 10 Malian girls complete primary education (US grades K-6). One in 10 finish secondary school (US grades 7-12), and by the age of 18, over half of all Malian girls, 5 in 10, are married off.

(<https://www.worldbank.org/en/news/press-release/2018/12/10/rapport-reduire-les-inegalites-de-genre-au-mali-tchad-niger-et-guinee>).

Although education costs and geographic distance from schools may explain these trends, social norms related to the role of girls and early marriage are important drivers. When a girl leaves school to get married, her low educational status produces significant and lifelong effects, in particular on her income and the education of her future children.

(<https://www.worldbank.org/en/news/press-release/2018/12/10/rapport-reduire-les-inegalites-de-genre-au-mali-tchad-niger-et-guinee>).

Across all nations, the Covid-19 pandemic has exposed, amongst other things, the vulnerabilities within populations, and weaknesses in infrastructures that have been inadequately designed to address the most basic public health issues facing global populations. The pandemic has further aggravated structural inequalities that systematically disadvantage women, specifically in the care economy (both paid and unpaid) (<https://womensempowerment.lab.mcgill.ca/seminars/care-economy-global-south/>). According to a policy brief from the United Nations (UN), before the global pandemic, women were doing three times as much unpaid care and domestic work as men

(<https://www.medicalnewstoday.com/articles/how-covid-19-affects-womens-sexual-and-reproductive-health#Womens-mental-health-under-strain>) This imbalance has increased significantly since the pandemic, with children out-of-school, family members in quarantine, and movement restrictions imposed.

(<https://womensempowerment.lab.mcgill.ca/seminars/care-economy-global-south/>). Access to reproductive health services for women has become a special concern, as health systems have diverted resources away from women's health in favor of infectious disease initiatives to stave off the spread of the virus.

(<https://www.medicalnewstoday.com/articles/how-covid-19-affects-womens-sexual-and-reproductive-health#Lessons-from-previous-pandemics?>). In a Lancet journal report about the gendered impact of the COVID-19 outbreak, the authors draw parallels between the Covid-19 pandemic and the Ebola outbreak. Citing examples from Sierra Leone, the authors state that "...resources for [women's] reproductive and sexual health were diverted to Ebola emergency response systems, contributing to a rise in maternal mortality in a region with one of the highest rates in the world.

(<https://www.medicalnewstoday.com/articles/how-covid-19-affects-womens-sexual-and-reproductive-health#Lessons-from-previous-pandemics?>) A 10% decline in the provision of pregnancy-related and

newborn healthcare could result in an additional 28,000 maternal deaths and 168,000 newborn deaths (<https://www.medicalnewstoday.com/articles/how-covid-19-affects-womens-sexual-and-reproductive-health#Lessons-from-previous-pandemics?>).

RQ 2: How much can machine translation improve capacity for delivering actionable information supporting women's health and well-being in the wake of COVID-19?

Timeline (note that all times are from the start of the project) for RQs 1 and 2

0 months: A pilot machine learning system up and running.

6 months: Specific information needs identified, and French sources identified, pilot humans-in-loop training system ready.

12 months: First phase of training data translated by crowdsourced Malians, first major test of machine learning system completed.

15 months: Begin second phase humans-in-loop translations and testing.

22 months: Testing finishes and project wrap up begins.

24 months: Final reports finished.

The project is building on a research project already initiated cooperatively in Mali and by Rochester Institute of Technology to explore the problems of automated translation for under-resourced African languages, focusing on Bambara. This project has also benefited from cooperation with another project to promote NLP for African Languages, Masakhane <https://www.masakhane.io/>. Masakhane promotes open-data, data sharing, open source, and collaboration among NLP researchers working on under-resourced African languages. Another project, not currently linked with our consortium, but provided a proof-of-concept demonstration of our application is the NGO Translators Without Borders automatic translation project Gamayun <https://translatorswithoutborders.org/> <https://translatorswithoutborders.org/translators-without-borders-scales-program-to-develop-machine-translation-for-marginalized-languages/>, funded by the Cisco Foundation. Translators Without Borders has managed the translation of over 2,500,000 words for the COVID-19 response, though very few of those to African languages. Their work is a good illustration of the need for translation in a global health crisis, as well as good illustration that the Global South is dramatically underserved. Machine translation will address both the resource need and the resource gap.

The project aims to create significant, functional, domain-specific capacity in two years, with intermediate results that will enhance capacity in the first year of the project as it builds on work already initiated. The capacity enhancement will be permanent and it will enable further capacity building with

the construction of a general system of translation and efficient systems of information diffusion and communication. The project is structured as well to train Malian students in AI and will result in strengthening of the Malian education system to teach AI and to use AI for the development objectives of the nation.

RQ 3: To what extent does crowdsourcing improve the performance of machine translation systems at translating COVID-19-related public health information between French and Bambara?

The Malian consortium members are already actively engaged in battling COVID-19 in Mali as members of the research, scientific and innovation communities responding to the pandemic. The relation of the project members and the project to rapid response and policy setting in Mali is detailed in the response to question 5.2.

Mali confronts a massive and persistent youth unemployment problem, with, nationwide, 24.45% unemployment in this category in 2019 (ILO). Studies (for example, Bouton) have established that youth unemployment is twice as high in urban areas than in rural areas, with only 35% holding a job and only 52% economically active. Mali is also a country which is upside-down in respect to the effect of higher education on employment: university education produces higher rates of unemployment due to a mismatch between the skills taught in Malian universities and the needs of industry. The problem of “insertion”, that is, the transition from university to work, is exceptionally severe in Mali. The “chômeur-diplômé”, that is, the unemployed university graduate is so common that the term has come to refer to the ordinary situation of one who has completed university studies. It is also considered normal that the graduates remain in this status for several years before, for the lucky ones, landing their first job.

The significant Malian participation in AI research, education, and in crowd-sourcing and funding going to Mali by this project will make a contribution to addressing these problems. While AI has become in just the last few years a tremendous motor of economic development, the field has barely gained a foothold in Africa and even less in francophone African countries such as Mali. The government has recognized the contribution AI could make to Mali’s development, having created the first center for AI education in francophone Africa, yet, AI is still not offered in regular university curriculums or as a field of study due to the lack of qualified Professors. It is anticipated that this project will enable researchers in related fields at USTTB to gain experience in AI methods and produce Malian graduates with degrees in AI capable of teaching AI in Mali. The establishment of Mali’s (and potentially Africa’s) first degree program in AI at a national university is a desired outcome. This can contribute to increasing the relevance of university studies in Mali to the economic development of the country and, therefore, the employability of university graduates.

The crowd-sourcing aspect of the project may lead to the establishment of a new sector of economic activity in Mali, as the expansion of AI and supervised learning techniques into numerous sectors has

created a market for data-labelling services. It is an interesting sector for Mali, as its relatively isolated geographical situation is not a handicap and the country's low wage structure could allow it to be competitive. In addition to learning the mechanisms of data annotation, participants will develop translation skills and familiarity with the workflow of an AI project. These skills will give participants meaningful work experience and may help with their insertion into the workforce. For practical and ethical reasons, the project will pay crowd-sourcers for their labor. As a low-income country, the great majority of Malians cope with economic precarity on a daily basis, and students no less than any other group. The majority of Malian students are only able to attend school because they receive an allocation from the government of approximately \$48 per month. They pay for food, transportation, school supplies, communication and internet connectivity and, sometimes, for housing with this. Many also must contribute financially to their families. Opportunities for part-time jobs are extremely rare in Mali. Payment for the work done by the crowd-sourcers is necessary to ensure that the students and their families do not suffer because of time lost or costs incurred and provides some measure of economic inclusion, allowing students with modest means to participate. As an extreme low-income country, the direct economic contribution to the local economy of the project will be non-negligible.

RQ 4: To what extent can humans-in-the-loop AI systems with Mali-specific crowdsourcing tasks such as the one we are building directly impact Malians economically and through skill- and literacy-building, and indirectly through increasing capacity for AI research, development, and entrepreneurship?

Timeline for RQs 3 & 4:

0 months: Basic crowdsourcing UI, that registers users, obtains consent, tracks work down, and presents a simple editing task complete.

3 months: Pilot testing with collaborators complete. Interface expanded to include domain-specific rating system and data cleaning interfaces.

6 months: Pilot humans-in-the loop system ready, first group of local annotators recruited.

15 months: Begin peer-to-peer recruitment phase.

22 months: Testing finishes and project wrap up begins.

24 months: Final reports finished.

3.3. Research Objectives

Please highlight the general objective of the research, and specific objectives. The general objective should state the development goal being pursued by the research. The specific objectives should indicate the specific types of knowledge to be produced, the outputs anticipated, and the audiences to be reached, and forms of capacity to be reinforced. These are the objectives against which the success of the project will be judged. Use only active verbs (no passive).

The general objective of this research is to deliver a service for automatically translating, on demand and in large volumes, French and Bambara news and information about COVID-19, the most widely-used language in Mali, using a machine learning approach driven by humans in the loop. The entire premise of this project is that most information coming to Mali about COVID-19 is in French and that there is a demand for it to be translated into Bambara. It composes the following specific objectives:

Specific Objective 1. A state of the art machine learning pipeline that will automatically translate, on demand actionably public health information related to COVID-19 between French and Bambara. We will produce computer science knowledge about the best machine learning algorithms for translating French to Bambara, focused on the specific domain of public health response to the COVID-19 challenge, and tailored to the specific needs of the local population, with considerations for Malian women as specific end-user targets. We anticipate this will contribute specific technical innovations around loss functions that focus on the actionable value of the translations, rather than their absolute fidelity. The primary audience for this knowledge will be computer scientists interested in machine translation of under-resourced languages, as well as public health officials who work in regions where under-resourced languages are predominant, and where automated machine translation may help alleviate a dearth of public health information, particularly in times of crisis. This objective will reinforce capacity for translation services, specifically for public health information, and capacity for actionable healthcare data. This objective will deliver Outcomes-1 and -4 (see section 3.5)

Specific Objective 2. A crowdsourcing pipeline, tightly coupled to the machine learning component, that will recruit Malians to help collect parallel tests, clean source data, and provide crucial expert and non-expert supervision for the machine learning model. We will produce human-computer-interface (HCI) knowledge about how to design crowdsourcing interfaces that best, in terms of quality, quantity, and providing a beneficial work environment, elicit translations from French to Bambara (and other Malian languages), where quality is determined by how well the that can be used to train and evaluate a

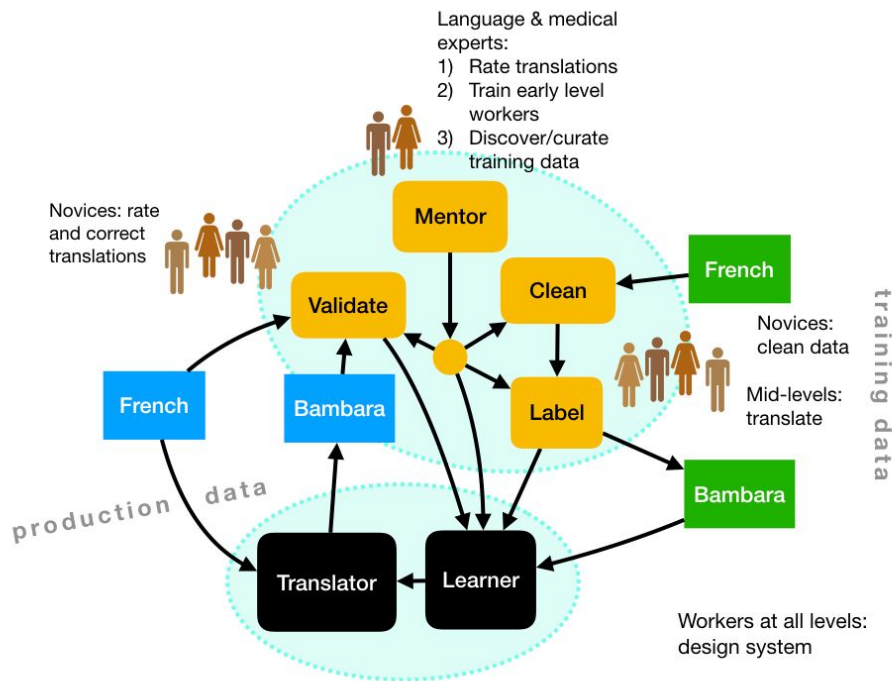
machine translation system, by integrating expertise from any Malian who understands both French and a Malian language with public health and disease experts. The primary audience for this knowledge will be HCI researchers with an interest in integrating heterogeneous expertise to improve machine translation through humans-in-the-loop machine learning. This objective will reinforce capacity for rapidly obtaining the parallel texts needed to train machine learning models. This objective will deliver Outcome-2 (see section 3.5)

Specific Objective 3. A system for recruiting and training workers into the crowdsourcing platform we will build for Outcome 2 a skilled population of Malian translators, who will help clean and label training data, our machine translation service will use to learn how to translate from French to Bambara, We will produce crowdsourcing knowledge on how to recruit a body of crowdworkers from across Malian society, specifically one that provides equal representation across gender, to translate French to Bambara (and other Malian languages), and retain and impart transferable skills in language translation to these workers. This will reinforce capacity for human translators, both to support the production of parallel texts for machine translation, and to provide interpersonal translation services for any reason. The intended audience are Malians of all backgrounds and genders, who have a degree of literacy in French and Bambara. It will also provide capacity for training Malians in an in-demand skill and pay them while doing so, thus reinforcing capacity in direct economic aid to Malians of all genders. This objective will deliver Outcome-3 (see section 3.5)

Specific Objective 4. A system for identifying and collecting French texts or French-Bambara parallel texts, which crowdsourcers will rate translate themselves (if French-only) or rate for quality and edit (if French-Bambara parallel), that will most efficiently improve the quality of our machine translation system. The primary audience will be linguists (computational and otherwise) with an interest in French-Bambara corpora for studying translation and other comparative language problems, as well as the Bambara-speaking Malian public at large, who will benefit from the information translated in these texts. This objective will reinforce capacity for rapidly obtaining the parallel texts needed to train machine learning models. This objective will deliver Outcome-4.

3.4. Research Methodologies

Explain how each specific objective will be achieved in enough detail to enable an independent assessment of the proposal. This section should show how the research questions will be answered, including articulating relevant conceptual and theoretical frameworks. For proposals suggesting AI and data science innovations, proposals should highlight if the proposed work will be applying existing methodologies and/or developing novel techniques.



Current state-of-the-art machine translation systems perform adequately well on languages for which there exists adequate quantities of aligned text: i.e., corpora in the source and target languages, where for each sentence in the source language (e.g., French) there is a translation in the target language (e.g., Bambara). There is currently a growing literature on best practices for adapting such machine translation systems to under-resourced languages, such as Bambara, which has been passed down through oral tradition, and lacks a large volume of text aligned with any potential source language, and we will avail ourselves of these methods and possibly contribute to this literature. However, we believe a more effective approach is to focus on innovations that will increase the volume of aligned text whilst directly providing paid work and training to Malians who will provide this text.

We are already using a transformer model—widely considered to be the best model for translating highly-resourced languages—based on JoeyNMT, and comes to us from our extended partner, Google scientist Julia Kreutzer, through our joint participation in Masakhane, an African-based organization dedicated toward pooling machine translation code and data resources throughout Africa. We do not anticipate being able to beat this as the main component of **Specific Objective 1**, however it will require extensive tuning and continual maintenance throughout its lifespan. Our Ph.D. student and programmer will initially (during the first year of the project) perform these functions, but we anticipate by project's end—after we have a better understanding of the problems specific to our learning environment—that the bulk of this work can be performed by undergraduate- and master's-level computer science students,

ideally based in Mali, providing them with a unique, practical, learning experience. However, RIT will provide remote guidance and direct assistance as needed.

Beyond this relatively well-established model, there are some key subjects of ongoing research in the under-resourced language community that we will investigate. The first is to translate multiple languages at once. We can use this technique for only a very limited set of the texts we have available to use, chiefly religious in nature, whose language is archaic and not relevant to the domain of pandemic response. However we do have several examples of French, English, and Bambara texts available to us and we will seek to leverage such multi-language approaches whenever possible. Another emerging approach is to train and obtain feedback from translations in both directions. Because there is no one right answer for translations, these methods can be used to substantially increase the number of aligned texts, from a much smaller seed set. They can also help leverage translators that may know only the target or source language. Variations on these approaches that emerge over the course of this project may lead to new innovations that could apply beyond our setting, and we will seek to publish those at top natural language processing venues if they arise.

To evaluate our models, we will use common metrics such as BLEU score, but we also expect to contribute new evaluation metrics aimed at measuring how successful our results convey critical public health information to end users, via our crowdsourcing module. We will discuss those in more detail below, but from the perspective of innovation, evaluation metrics often coincide with loss functions. However, as loss functions must drive the learning process, they are subject to constraints that evaluation metrics are not (for instance, they must be piecewise differentiable). And so there is always active research around new loss functions, whenever new evaluation metrics are proposed, that can translate as faithfully as possible the performance details the evaluation metrics are intended to capture, into a form that can be used to drive machine learning.

We will track who uses our system and require them to answer a brief survey before they use it, stating what organization(s) they represent and how they intend to use their translations, and providing us with an email address or phone number to reach them by. Immediately after the system produces their translations, it will send them a brief automated survey on their level of satisfaction with the system and invite them to upload an edited translation, if they decide to make changes. We will follow up six months later with a final survey asking them what they did with the translations and whether or not they were effective. We will evaluate this objective by analyzing the responses of these surveys and tracking the number of documents translated.

The crowdsourcing module has three main functions: (a) provide cleaned and labeled data, (b) evaluate machine translation performance (and use intermediate evaluations to improve the transformer model through an online, continuous feedback loop), and (c) recruit, and engage, economically support and train Malians in the discipline of translation. We discuss (c) below, as **Specific Objective 3**, which is about the

social and economic contributions of crowdsourcing. **Specific Objective 2** is about functions (a) and (b), which focus crowdsourcing's contributions to machine learning.

Functions (a) and (b) draw from a long line of research on crowdsourcing platforms that decompose complex processes into simpler *microtasks* that participants (*crowdworkers* in common parlance) with different skill sets work on. Our platform composes four microtasks: *cleaning*, where inexperienced crowdworkers take newly found texts (French only in some cases, French and Bambara in others) and manually clean them in a text editor like environment: remove garbled data; split long passages into individual sentences, align sentences in the case of parallel French Bambara texts (actually, much of this will be automated; however, experience has shown us that such automation is not perfectly reliable and that manually checking and fixing any errors dramatically improves performance when the amount of data available is rather limited, as is the case with under-resourced languages like Bambara); *labeling*, where relatively more crowdworkers with some fluency in written French and Bambara either fix language errors from the labeling phase or outright translate French into Bambara sentence-by-sentence; *validating*, where crowdworkers with more experience than cleaners, but less than labelers, rate the quality of the outputs (Bambara text) of the machine translator, generating new translations in the process, and feed these reports back to the machine learner to help it improve; finally, *mentoring*, where domain experts in Bambara and public health epidemiology review and rate the crowdsourced work and provide feedback to the other crowdworkers.

Tying these four microtasks together is the *manager* (the yellow circle in the middle of Figure 1). In its initial phase it will simply apply established tests for evaluating annotator quality, such as BLEU score and interannotator agreement. Then, we will use the tests from the initial phase to model the reliability of the workers and weigh their inputs to the machine learner, based on the classic approach of Dawid and Skene. Finally we will add more experimental components that provide workers feedback and rewards to incentivize performance and help improve their fluency in written French and Bambara and their skill as translators.

We will evaluate **Specific Objective 2** by tracking over time the improvement in the volume and quality of the training data produced, using worker feedback obtained from the validation microtask. We will work with our crowdworkers and domain experts on a variety of approaches to each of these subtasks.

Specific Objective 3 focuses on the human resources aspect of crowdsourcing. All crowdworkers will be paid for their work, for typically very small amounts of work at a time, e.g., translating or evaluating 10 sentences, after which they can perform additional microtasks or continue later. Workers can perform these tasks on their smartphones or anywhere else they have internet access. We will run pilot tests with paid participants to determine the amount of time required to complete the tasks, what the optimal workload is, and how much to pay workers in order to ensure that they are fairly compensated and to ensure worker well-being and task quality. We will test all new workers on each of the three lower experience microtasks (*validating*, *cleaning*, and *translating*) and assign them based on skill level and need. As workers gain experience and their performance improves, we will allow them to advance to more complex tasks and eventually share the mentoring microtask with our domain experts. We will additionally provide bonuses to workers who are exceptionally productive. We will elicit feedback from all workers on how to improve the system and make it fairer and a better training experience.

A major challenge is in recruiting crowdworkers to join the system. We will begin with a relatively small group of 20-100 crowdworkers recruited from students at USTTB, who are more fluent in written Bambara and French than the general Malian population, and with whom PI Leventhal has previously collaborated with on language translation tasks. We will then recruit new workers through word of mouth and peer recruitment methods, with which PIs Homan from RIT and Anto-Ocrah from URMC have experience, independently and in collaboration with each other.

We will evaluate **Specific Objective 3** by tracking over time the number of workers recruited and retained (and for how long) and use the same metrics we use for evaluating the improvement of training data quality for **Specific Objective 2**, but instead we track the performance of individual crowd workers over time.

Specific Objective 4 is about the raw material for training, which we expect will primarily be French texts related to COVID-19, but may also include some parallel French-Bambara and Bambara-only texts. We already have a French-Bambara dictionary with over 2000 translated sentences between French and Bambara, in addition to the word translations. We also have the full text of the pamphlet "Where There Is No Doctor: A Village Health Care Handbook" translated into French, English, and Bambara. Recent research suggests that training on simultaneous translations between more than one pair of language can greatly improve machine translation performance, especially when one of the languages is under-resourced.

Our approach will be: (1) identify key French and Bambara sources of COVID-19 related health information, and information to support health or well-being problems that may be impacted due to COVID-19 lockdowns (such as intimate partner violence or unemployment) (2) vet them with our team of language, health, and machine learning experts. (3) Repeat. In doing so, we will develop a specific set of best practices for discovering new sources of training data and criteria for vetting the quality of this data. In general however, in choosing which texts to send to our crowdsourcing system for translating and editing (and, eventually, as training examples for the machine learning pipeline) we need to balance several objectives: (a) Is the information accurate? (b) Would the information, if properly translated, benefit Malians, particularly the most vulnerable? If so, which Malians? (c) Is there a channel or agency for disseminating the information (d) How easy will it be for crowdworkers to translate them, within the limits of training we can provide them? (r) How much would the information, if added to our training data set, likely improve the performance of our machine learning pipeline?

We will evaluate this objective by tracking the performance of our machine learning system as new training data is added over time.

3.5. Expected outputs and outcomes of the research

Succinctly describe the concrete outputs and outcomes of the proposed research.

Explain how gender and inclusion considerations will be integrated into the outputs and outcomes, and the expected impacts of the work.

Outcome-1. A web-based art machine learning service to translate on demand to/from written and spoken Bambara, tailored to respond to COVID-19 through native language information resources.

Bambara-French translation also provides international exposure to traditional Malian healthcare. Number of organisations using it (four short-term, 10-20 medium-term, 100s long-term) quantity and quality of material produced (3-10 corpora short-term, 500-1000 medium-term, 100K-1M long-term), diffusion to health care workers (6 short-term, 50-200 medium term, 1000-10K long-term) and general population (50-1000 short-term, 44M Mande speakers medium-term, 1.2B Africans long-term). Increase in accurate public understanding of best healthcare practices, measured by surveys. Improved public health response, measured by national metrics.

Outcome-2. A web-based data annotation interface that will recruit, pay, and train Malians to help collect parallel texts, clean source data, and provide crucial expert and non-expert supervision for the machine learning model. We anticipate this will contribute specific technical innovations around methods that elicit human feedback on both the fidelity of the translations and, independently, the quality of the information they convey. This outcome is a crucial supporting component to Outcomes 1 and 3, but the knowledge we acquire in building this will produce scientific knowledge that will be of independent interest to those design complex crowdsourcing systems that integrate human knowledge across multiple domains of expertise for human annotation tasks.

Output-3. A skilled population of Malian translators, who we will recruit and train through the crowdsourcing system will build for Outcome 2 who will help clean and label training data, our machine translation service will use to learn how to translate from French to Bambara, evaluate the quality of the translations it provides, thus evaluating Specific Objective 1 while providing additional training data, and iteratively redesigning the crowdsourcing interface itself, thus evaluating Specific Objective 2.

Crowdworkers will also be peer evaluated by mentors, thus evaluating Specific Objective 3. We will screen by gender to help ensure that women are hired in equal proportion to men. Learning, employment, and nation-building contribution. Upskilling opportunity for 100-200 students, plus additional volunteers across the country, via creating crowdsourcing tools and AI labeling tasks. These transferable resources will create a new industry.

Output-4. A collection of texts that our service will translate from French to Bambara, which crowdsources will edit and rate for quality, that are related to encouraging public health behaviours to lessen the spread of disease, specifically COVID-19, in Mali, and with special consideration to the needs of the Malian population, particularly women. We have an expert on women's health disparities, PI Anto-Ocrah, and will additionally seek input through a community participatory advisory board, who will help to seek and identify information that is particularly critical for the health and well-being of Malians, particularly the most vulnerable. These texts will also form the basis of a corpus for training and evaluating the systems. This is the output of Outcome-1 and thus shares its impacts. Additionally, the

corpora produced will provide linguists and machine translation specialists with a useful set of data for training, experimenting, and evaluating their research.

Outcome-5 Collaborative institutional partnerships between US (RIT), Malian university system (USTTB), and Malian technology and entrepreneurship agencies (RobotsMali). Training of doctoral level Malian AI researchers on crowdsourcing and entrepreneurial offshoots of the project. We will seek to hire women in the equal proportion to men in all hiring decisions. Placement of doctoral students from USTTB at RIT (1 short-term, 2-4 medium term) and other institutions (10s-100s long term), US-Mali knowledge sharing measured by joint publications and doctoral theses (5-15 short-term, 10-60 medium term, 50-1000s long-term).

4.1. What strategies will you take to ensure your proposal is using a responsible approach (Rights-based, Inclusive, Ethical and Sustainable) to research and implementation?

What strategies or research questions will you undertake to ensure your proposed research takes human rights into consideration, is inclusive and aware of the needs of vulnerable and marginalized populations, and minimize harms? Please also provide a brief explanation of how you will design and integrate oversight or transparency mechanisms in your research to ensure quality, accuracy, and minimizing harms. Please provide text. (max. 1000 words)

You may also upload documents and/or PDFs at the end of the application.

This project was initiated by a Malian (Tapo Allahsera, currently a master's student at RIT) specifically for the purpose of providing Malians who cannot speak French great access to world information resources. PIs Homan and Anto-Ocrah and consultant Luger all conduct research on responsible use of science to assist underrepresented communities. We have several mechanisms to ensure human rights are considered in this project and to keep us informed of the need of vulnerable communities: our three IRBs, a community participatory group that RobotsMali will host and all investigators (plus consultant Luger) will participate in, the use of experts in research on underrepresented and vulnerable communities ("fairness experts"), built-in mechanisms in our crowdsourcing platform, and. We discuss our IRB plans in section 4.3, and the three mechanisms below.

We will host bimonthly community participatory advisory group meetings through advertisements on the RobotsMali website and through word-of-mouth and through the PIs local to Bamako, inviting any community members to participate. We will provide dinner (if hosted live) and/or a small honoraria. These approximately 90 min meetings we feature progress reports on our project and elicit feedback from community participants on what information related to health and well being is most needed by them and their peers. We will use this feedback in searching and selecting texts to

translate and to address any other ethical or human rights concerns raised that are within our purview to address.

Our fairness experts will play an especially important role in the acquisition of text to target for human and machine learning (Output-4). In section 3.4, in our explanation of Specific Objective 4, we discuss the broad criteria for searching for and selecting text. One of these criteria is whether the information, if properly translated, would benefit Malians? Here we will rely on PI Anto-Ocrah to ensure that we address the impact of COVID-19 on women's health and well being, and our other fairness experts will ensure that feedback from our participatory group is conveyed to our medical experts. We will also invite members of the participatory board to directly contribute to this process and compensate them for their time.

Finally, our crowdsourcing platform will administer open-ended questionnaires from time-to-time to assess workers' health and well-being and how their work impacts them. They will be paid for the time it takes them to answer these. We will use our performance metrics, such as time to complete tasks and quality of annotations to reach out to and adjust workloads and compensation to ensure workers are fairly treated.

4.2. How will your project address and integrate gender considerations? Please give a brief explanation of your strategy to integrate gender questions and considerations in your research questions, design and implementation, and offer examples.

Please provide text and/or up to five links here to material that demonstrates your strategy and research. (max. 1000 words)

We will involve at least 50% female Malian participants and stakeholders at the various stages of the project as follows:

Crowd-Sourcing and Translation: A significant portion of this project consists of using crowd-sourcing data, guided, annotated and validated by domain experts across the social and medical strata. This will include Malian research participants, who represent user communities across the country. At least 50% of our crowdsourcing approaches will involve matriarchs who are the major decision makers in their households, those in rural and disenfranchised settings with limited access to accurate health information, and women who are native speakers of Bambara; the language they understand. We will actively engage female professional translators, who are currently responsible for producing official translations from French to Bambara, in the design and execution of the data validation pipeline to ensure that the results i) are contextually appropriate and acceptable to the research participants, ii) incorporate the needs of the women in the community, and iii) are well aligned with the language policy objectives of

the Malian government, which is the right and necessity of Malians to access and use their native languages in all spheres of life; particularly their health and well being.

Project Scale: Due to language limitations, the Covid-19 pandemic has created an information vacuum that breeds rumors and misinformation; limiting positive responsiveness, and increasing distrust of the “elitist” government. By the end of the two-year project duration, we aim to have constructed a system of translation and information diffusion that is functional, domain-specific, culturally and contextually appropriate, and permanent. Through the use of machine translation (MT) systems, Malian public agencies and health organisations will have automated tools to rapidly produce and distribute health information in the language that 80% of Malians understand, a capability that does not exist today. The target end users will be mostly women from disenfranchised communities, who are seeking accurate covid-specific information in non-French languages. This represents over half of the Malian population, who are spread over expanses of rural and Sahel geographies.

Our innovative approach of combining MT systems with human validation of the data will ensure that language translations convey critical information and combat misinformation for the people who matter the most.

You may also upload documents and/or PDFs at the end of the application.

4.3. Please provide a brief assessment of known ethical challenges and review pertaining to your proposed research, particularly as it pertains to human subjects.

What approaches and methods will you use to collect and manage data, and how will you manage ethical questions related to confidentiality and privacy? Please share if you have access to an Independent Review Board (IRB) for reviewing or managing any ethical challenges that might arise. If you do not, please document how you intend to manage any ethical challenges that might arise. Please provide text and/or up to five links here that demonstrate ethical oversight for your project. (max. 1000 words)

You may also upload documents and/or PDFs at the end of the application.

USTTB, URMIC, and RIT all have IRBs and all three will review this project before it involves human subjects. The data that we use for machine translation training, development, and production will all be publicly available and intended for broad public dissemination, and as such will likely be seen as exempt from human subjects review.

However, we will conduct research on our crowdworkers, and this research will likely count as human subjects research, though no more than minimal risk. The benefits to these participants include compensation for their involvement and training in language translation. The risks include some risks to privacy, as participants must reveal enough about themselves to be paid for their work. The crowdworker training will include a paid onboarding period where the participants will be taught not only how to use the language technology and translation tools but also how their contributions will be incorporated into the larger project.

There is also some risk of stress and burn-out due to the cognitive load of translating sometimes rather technical language that sometimes recounts traumatic or painful stories about other people. We have conducted some research on these risks and will take measures to elicit the mental health of participants and prevent burnout. During our pilot study phase of the humans-in-loop crowdsourcing interface, we will time how long various tasks take and use that to gauge a fair wage for crowdsourcing works, we will also conduct surveys designed to elicit whether they believe the wages are fair. In addition, we will pay above the local legal Malian minimum wage and ensure that this payment is more than fair and actually desirable due to the thought-provoking nature of the labor.

In order to ensure that unexpected consequences do not exceed our capacity to respond to them, the rollout of the crowdsourcing system will have three phases: first, only our PIs and consultants will initially use and pilot test the system. Second, we will recruit primarily students from USTTB and others from Bamako, so that we can meet with them face-to-face in case there are any crises engendering collaboration and receptiveness to the crowdworkers' work experience. Once we are assured that there appear to be no significant unmanageable consequences will we commence the peer-recruitment phase.

The COVID-19 Global South AI and Data Innovation Program

Section 5. Policy Relevance, Uptake and Scale

5.1. Describe how your proposal connects or aligns with sub-national, national, and/or multi-national responses to the COVID-19 in the countries of proposed impact? Please describe the kinds of engagement your organization/consortium has had or will have with national and sub-national governing institutions (e.g.

government departments, bodies and agencies as well as regulators and other public-sector institutions) on topics related to your proposal.

Please provide text and/or up to five links here to material that demonstrates your strategy and research. (max. 1000 words)

Mali has a comprehensive approach to COVID-19 strongly coordinated by the government in a multi-agency approach. All three Malian consortium members are organisations under the aegis of the Ministry of Higher Education and Scientific Research, the Ministry responsible for leading the scientific and medical response to COVID-19. USTTB is home to Mali's only P3 laboratory responsible for COVID-19 testing M. Koné, co-PI is researcher in this lab and a Malian authority on COVID-19.

[\(https://fmos.usttb.edu.ml/index.php/2020/04/09/laboratoire-de-lucrc-plus-de-400-tests-du-covid-19-effectues/\)](https://fmos.usttb.edu.ml/index.php/2020/04/09/laboratoire-de-lucrc-plus-de-400-tests-du-covid-19-effectues/) USTTB schools of Medicine and Public Health, key players in Mali's medical and scientific response will contribute their expertise to the project. One of the primary objectives of the government policy is to inform the public and to gain acceptance of measures necessary to prevent the spread of the virus and, in most respects, this is the most critical and most challenging aspect of the national response. This activity is led by the Ministry of Health. While French is the official language of Mali used by government structures, only 20% of the Malian population are fluent in this language. The Ministry of Health uses the Academy of Malian Languages (AMALAN), the second Malian partner in our consortium, to translate materials into the national languages of Mali, the mother tongues of all Malians. AMALAN is equally responsible for the implementation of the national law giving Malians the right to education and services in national languages. In the context of the COVID crisis, this has become not just a right but a key to stemming the spread of the virus. The third Malian participant, RobotsMali: The National Collaborative Center for Education in Robotics and Artificial Intelligence, the third Malian participant, works to realise the national objective of making education and the use of advanced technologies such as a robotics and AI, a motor for the social and economic development of the country

<https://robotsmali.org/fr/developpement-des-competences-en-matiere-dapprentissage-automatique-en-afrique-dans-la-region-francophone-du-sahel/> RobotsMali, in addition to its primary mission, is a member of a consortium of Fablabs and technology centers in Mali, COVID-MALI

<https://www.covidmali.org/> developing, in cooperation with USTTB and the Ministry of Health, improvised manufacturing techniques to respond to health and medical equipment shortages. This effort has led to needs assessments with public agencies and NGOs that are providing front-line COVID-19 services to the community in cooperation with the Ministry of Health. These organisations include women's community health provider Muso (Woman)

<https://www.musohealth.org/>, the police, and community youth organisations working to heighten awareness in the population, It was RobotsMali that formed the Malian contingent of the proposal consortium and also initiated a precursor to the project, Bayeɛmabaga (Translator)

<https://robotsmali.org/fr/projets/bayelemabaga/> , training Malian students in NLP and crowd-sourcing

preliminary data, to further national policies promoting education and use of AI and the strengthening of native languages.

The US-based participants have significant experience in engaging with issues relevant to Malian COVID-19 response and, in general, with the Malian and West African situation. RIT has hosted the first Malian student, a Fulbright scholar, in the US to have studied AI and Sarah Lugar of Orange Silicon Valley and PI Homan of RIT have been working extensively with RobotsMali in the initial phase of the project. The team has direct experience in Senegal, Ghana and in other African countries apart from Mali. This is a team that brings together the Global South and the Global North in an equitable partnership with the profile necessary to contribute capacity and a sustained impact to the COVID-19 response in Mali.

5.2. How will you manage use and mobilization of the research? Please give a brief explanation of your strategy and offer examples.

Articulate who the end users are, if they participated in the design of the project, whether they will participate in the project, and how they will be engaged in the implementation of project results. (max. 1000 words)

This research project has direct and broad practical outcomes for Mali's COVID-19 response and for its long term capacity to deal with public health emergencies. The central outcome is that public agencies and health organisations will have tools to rapidly produce and distribute health information in the language that 80% of Malians understand, a capability that does not exist today. The end users, health information users and disseminators, include native Malians particularly women who tend to have the primary decision-making responsibilities in family units and communities, the Malian medical community including researchers, medical staff, and community health workers, the government, including the Ministry of Health and the nationwide network of community health centers, health and emergency-response organisations working in Mali, official, professional and non-professional translators and other social and security agents with frontline duties in responding to the pandemic.

A significant part of this project consists of crowd-sourcing data, guided, annotated and validated by domain experts, with the Malian research participants representing each of the user communities listed above. Malian scientists leading the nation's COVID-19 response in the fields of molecular biology, epidemiology, medicine and community and rural health will be responsible for framing the key knowledge in their fields relevant to the pandemic response, providing direction on translation sources

and target domains and for validating results for accuracy and utility. This is the community most responsible for developing the strategy of pandemic response and providing source material used to educate medical providers, community health workers, and the public. The project will use crowd-sourcing cohorts identified from the end user communities of medical professionals, community health workers, and other frontline workers who already have the responsibility of communicating health information provided in French to the Bambara-speaking population. Professional translators who are currently responsible for producing official translations to Bambara will also be engaged in the design and execution of the data validation pipeline to ensure that acceptable results are produced consonant with national language policy objectives. These communities are the very communities that will engage in the mobilization of the research and will provide a direct pipeline of feedback from the public that they serve as the efficacy of the research product.

The production of the health manual “Where There is No Doctor” provides an illustration of a related process that brought together a user community for the development of a resource with a practical impact on health response in developing countries. “Where There is No Doctor” was a field health manual intended for use in rural settings in developing countries by community health workers. It has been translated into 100 languages and modified to fit local circumstances. The organization “The Dokotoro Project”, currently collaborating with our project consortium, brought together community health experts, Malian health workers, and Malian translators to produce a Mali-specific version in Bambara: <https://dokotoro.org/>, <https://gafe.dokotoro.org/>. Our project will bring together a broader user community, implicating its participants directly in the data collection and evaluation of the research product, with an end result that accelerates the production of critical health information and creating a capacity for continuous automated production.

5.3. What are the potential opportunities to scale the research innovations, or potential for uptake and policy adoption?

How will you take into account how your proposal works, why, for whom, to what extent and in what contexts? What technical, social, and institutional factors will need to be addressed to scale your intervention, and how does your research design integrate thinking about these factors? Please provide text. (max. 1000 words)

This project’s interdisciplinary team ensures the project aligns with long-standing national objectives, the right and necessity that Malians can use their national languages in all spheres of life, providing an automation technology to achieve that which has heretofore been unachievable. Based on its technology, innovation, and breadth, this project will penetrate every

corner of Malian society, achieving, at a national level, scaling through the participation of public institutions and social actors

The broad penetration of 4G and smartphones at all levels of Malian society also supports growth and community engagement. Thus, the infrastructure for nationwide diffusion is already in place.

This project addresses a world-wide challenge faced by many people: that of digital enfranchisement of populations speaking one of the world's 6000 under-resourced languages. For marginalized populations in the African region, government languages like English, French, and Portuguese, borrowed from the colonial era, are not native languages and are often inaccessible. Thus, the potential long-term scale and impact of the research will be applicable to other languages; an emergent AI technique called "transfer learning" provides strong evidence that they can be. The novel crowd-sourcing methodology developed by this project will also provide a framework reemployable for other languages. We established a channel for scaling our work beyond Malian borders through our participation in the cross-African consortium of NLP projects, Masakhane. Through Masakhane we are sharing data, experience, and code across 17 countries representing 29 major, underresourced, African languages.

6.1. Project Management and Collaboration

Please provide an overview of how the organizational matters of the proposed research will be managed, including roles and responsibilities. If applying as a consortium, indicate responsibilities and proposed management of the projects, as well as the history of collaboration between organizations. How will the approach share ownership for the outcomes among the different organizations? Please also include considerations about the anticipated time to implement rapid response mechanisms, and strategies that will be undertaken to minimize the impact of social distancing activities. (max. 1000 words)

RobotsMali

Coordination of all activities in Mali, including external Malian participants, AMALAN and USTTB. Recruitment of crowd-sourcing managers, crowdworkers, and management of data pipeline including collection and stages of validation of data, coordination with Malian experts consulting on language and biology of Covid-19, public health, issues, rural issues, and epidemiology. Co-supervision of USTTB students participating in project. Primary interface to US participants and research liaison. Responsible for Malian contributions to research reports, scientific articles and presentations.

Participate in weekly meetings. Co-mentor students in Rochester and Mali. Crowdsourcing evaluation, cleaning, and labeling of machine learning data.

USTTB

Expert support from bilingual (Bambara, French) Malian researchers in the following areas: COVID-19, Medicine, Epidemiology, Community and Rural Health. Coordination of participation of Computer Science department and Master's students in project working on NMT and data collection and validation.

Identify source information and source materials for translation relevant to COVID-19 response in expert's domain and individuals capable of contributing to crowd-sourced translations. Work with linguists to establish Bambara terminology for translation. Review translations for accuracy and medical efficacy. Promote use of automatic translation to Bambara in public agencies and government, participating in workshops and conferences. Contribute to project reports and scientific publications and conference participation.

AMALAN.

Expert authority on Bambara language translation and speech transcription. Ensure compliance with national language policy and promotion of automatic translation within official government channels.

Develop translation standards, train crowd-sourcing managers, develop criteria for validating and scoring Bambara translations, manually validate translations, contribute to project reports and scientific articles and presentations. Select sources and transcribe Bambara speech for training speech recognition systems.

RIT. Though nominally the lead institution, except for F&A and the travel budget, the rest of the funding will go directly to students from global South nations. While USTTB is Mali's foremost research university with a strong track-record of successful, internationally-funded projects, they have neither a core discipline in AI nor experienced researchers in this or related domains. A major objective of the RIT-USTTB relationship is to help USTTB create an academic program in AI. RIT's primary role is to provide logistical and educational support for the technical aspects of this project. We will seek to recruit Malian students and we currently have one (male) who is committed to serving as a programmer in the first year of this project and then join the team as a Ph.D. student in the second year. We will seek to fill in the open second year Ph.D. slot with a woman from Mali. Investigator Homan is an expert in humans-in-the-loop machine-assisted analyzing unstructured language in the health domain. He is the director of the Lab for Population Intelligence, which seeks to build AI systems that are inclusive of all members of the populations they serve. He also serves on the leadership team of the Center for Personalized Healthcare Technology (PHT180).

URMC. As population health researchers on the [translational science spectrum](#), Dr. Martina Anto-Ocrah, the epidemiologist on the consortium, strives to give representation to vulnerable populations in under-resourced settings. At the core of the lab is a dedication to gender

disparities. Director Anto-Ocrah will fill several roles in this project, primarily focused on ensuring gender disparities are fairly represented in all aspects of this community-focused project.

The approach we are taking is highly participatory, and will depend on the work of as diverse a body of translators as possible. Obtaining domain-specific source texts requires expert advice on what source texts to seek. Anto-Ocrah will ensure that this effort remains focused on women's health as we seek those texts. There are countless examples where machine learning exhibits racial or gender-based biases because the training data comes from the experiences of white men. Anto-Ocrah's expertise and leadership will help to ensure that this will not happen in this project. Beyond that, she will provide guidelines on participant recruitment for the crowdsourcing and translation components of the project. As an epidemiologist, her expertise will be essential for this Covid-19 centered project.

Finally, as we are interested in broadening our reach to other Global South nations, her perspective as a Ghanaian, and expertise with research with Ghanaian, Malawian and Rwandese populations gives us a much-needed degree of perspective for this future development.

Tapo Allahsera is a Malian who is currently a Fullbright scholar enrolled in the master's degree program in the RIT department of computer science, he expects to graduate this summer, and his masters thesis forms the body of much of our preliminary work and he is the originator and driving force behind this project. If funded, he will serve as a programmer on the project during its first year as he applies for PhD programs for the second year. With luck he will attend RIT as a PhD student in the second year and continue working on the project there.

Prono Bono Consults: Sarah Luger, an AI/ML/NLP research scientist from Orange Silicon Valley, USA is an expert in natural language, crowdsourcing, and responsible AI, and will consult on all matters related to these aspects of the work.

Julia Kreutzer is a research scientist at Google, based in Montreal, Canada, and is an contributor to Masakhane. She is an expert in machine translation and human labeling, and has shared with us her extensive machine translation and crowdsourcing codebase. She has also met with us for approximately one hour per week for the past several months and will continue to do so for the foreseeable future.

Matthew Herberger is the cofounder of The Dokotoro Project, <https://dokotoro.org/>, publisher and translator of the Bambara/French/English manual, "Where There Is No Doctor: A Village Health Care Handbook." A former US PeaceCorps Volunteer, he is fluent in all three languages, and has a passion and interest in computer-assisted machine translation. He has and will continue to help us test translation interfaces.

Past collaborations

Homan (RIT), Leventhal (RobotsMali), Luger (Orange Silicon Valley), and Kreutzer (Google) are all members of Allahsera's thesis committee (Homan serves as chair). Homan, Leventhal, and

Luger (along with Marcos Zampieri) published preliminary work on Allahsera's thesis and the first AfricaNLP workshop at The Eighth Annual Conference on Learning Representations: "Assessing Human Translations from French to Bambara for Machine Learning: a Pilot Study."

Until last year, Homan (RIT) was an adjunct faculty member and URM. His collaborations with Anto-Ocrah are relatively recent. They have authored one grant proposal together, and have one published paper, Anto-Ocrah, Martina, et al. "Public knowledge and attitudes towards bystander cardiopulmonary resuscitation (CPR) in Ghana, West Africa." *International Journal of Emergency Medicine* 13.1 (2020): 1-12. This paper used peer recruitment methods of the sort we expect to use in this project, during the second phase of crowdsourcing.

Leventhal (RobotsMali) and Koné (USTTB) have cooperated for many years on various projects related to strengthening university education, research, and scientific culture in Mali. USTTB is one of the founding members of the RobotsMali Association, having participated in and financed many of its activities in robotics and artificial intelligence and provided a pipeline of students who have been trained at RobotsMali's facility. Leventhal and Koné have served on the national committee for the organization of "The Festival of Science" for the last 5 years, the primary event in Mali for the vulgarisation of scientific research and the promotion of scientific culture. USTTB and RobotsMali have cooperated on another project related to the COVID-19 response involving the use of additive and improvised technologies to produce locally manufactured PPE.

Leventhal (RobotsMali) and Traoré (AMALAN) have been cooperating since the beginning of the year on an earlier phase of the translation project, having identified data sources for monolingual Bambara texts, aligned bilingual texts, and other collaborators in the field of Bambara linguistics and language development. They have also worked on methods for evaluating Bambara translators and translations.

Shared ownership

RIT will be responsible for managing all human data. All other data and software will be open source and freely distributed. We will openly collaborate and share joint authorship on papers.

Rapid response/social distancing

Nearly all the work performed thus far has been done remotely, across two continents and three time zones with a seven hour total time distance, and we have been very productive. All products will exist on the cloud. We do anticipate that social distancing may impact who will be able and interested to participate in crowdsourcing and community advising, but we are confident that it will not substantially impact the number of people.

6.2. Relevant Experience and Related Work

Please share a short summary and/or up to 5 links to related research by your organization or, if applying as a consortium, by the consortium members on topics related to this call. (max. 1000 words)

Rochester Institute of Technology

Dr. Homan (PhD) is Associate Professor of Computer Science at the Rochester Institute of Technology, serves on the leadership committee of the RIT Center for Personalized Healthcare Technology, whose mission is to advance healthcare research across the university, and directs the Lab for Population Intelligence (LPI). The mission of LPI is to use AI to solve problems collectively. The lab designs and investigates algorithms that engage, model, and learn from human populations, creating dynamic systems that represent diverse values and beliefs, build trust, and advance community values in public policy and decision making.

His research centers on the gap between human-computer interaction (HCI) and machine learning (ML), specifically on mechanisms for human- and machine- based intelligence to learn from each other in computer-mediated settings, especially via social media such as Twitter, Facebook, and Reddit, and crowdsourcing platforms like Amazon Mechanical Turk (AMT). As a PI and coPI on NSF and NIH awards, he established methods for extracting mental-health related signals from social media and mobile data, including discourse on distress, work and employment-seeking, and suicide, that this project will extend. A particular focus has been on methods for engaging with and collecting data from underrepresented populations, particularly among LGBTQ communities and urban neighborhoods with high percentages of black and Latino residents, and ensuring that minority perspectives are preserved in machine learning with humans in the loop. These methods sometimes involve community-based participatory research (CBPR) methods for establishing research programs, novel annotation schemes, and generally ensuring that our analyses are meaningful to the communities who produce the data in the first place.

Selected publications:

<https://www.aclweb.org/anthology/W14-3213.pdf>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5421990/pdf/nihms822437.pdf>

<https://www.aclweb.org/anthology/D15-1309.pdf>

https://kilthub.cmu.edu/articles/Architecting_Real-Time_Crowd-Powered_Systems/6469835/files/11898389.pdf

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5082699/pdf/nihms822435.pdf>

University of Rochester

Dr Martina Anto-Ocrah (MPH, PhD, MT(ASCP) is Assistant Professor of Obstetrics and Gynecology, Emergency Medicine and Neurology at the University of Rochester School of Medicine and Dentistry in Rochester, NY. As an epidemiologist, Dr Anto-Ocrah's research

interests are focused mainly on under-resourced settings and vulnerable populations, with particular emphasis on gender disparities in populations in sub-Saharan Africa. Her research endeavors have spanned access to mobile phone technology and predictive health behaviors amongst pregnant women in rural Malawi, men's experiences as fathers in Ghana, and knowledge and attitudes towards bystander cardiopulmonary resuscitation amongst Ghanaian social media users.

In the wake of the Covid-19 crisis, Dr. Anto-Ocrah's research has moved deeper into the psycho-social domain. She is currently collaborating with partners in Rwanda, Ghana and the US to better understand the global impact of social isolation, social distancing, immobility and overall social and economic wellbeing of affected communities. In Ghana the team is evaluating the psychological impact of the pandemic on nurses at the front-line of their pandemic.

Research in Rwanda is addressing how communities will establish mechanisms for resilience to mitigate the adverse impacts of the pandemic, and we are doing a cross-cultural comparison of pregnancy experiences of American and Ghanaian women in the face of the global pandemic.

Each of these projects take a Team Science approach with multi-disciplinary collaborators whose expertise include Medicine (ObGyn, Emergency Medicine), Computer Science, Anthropology, Public Health, Statistics and Data Science (to name a few).

Select publications: <https://pubmed.ncbi.nlm.nih.gov/32522144/>

<https://pubmed.ncbi.nlm.nih.gov/31931748/>

<https://pubmed.ncbi.nlm.nih.gov/31929541/>

<https://pubmed.ncbi.nlm.nih.gov/30522936/>

Link to lab <https://www.urmc.rochester.edu/labs/anto-ocrah.aspx>

RobotsMali, National Collaborative Center for Robotics and Artificial Intelligence Education
Michael Leventhal is Founder and Director of Mali's first national center for robotics and AI education. A bi-national citizen of Mali and the United States, he has worked closely with the government of Mali to include artificial intelligence research as an element of the national policy to promote scientific culture and the use of advanced technologies to address the social and economic development of the country.

Diarra, Haby Sanou, & Leventhal, Michael, Developing Machine Learning Competence in Africa in the Francophone Sahel Region, ICLR, Workshop on Practical Machine Learning for Developing Countries, April 26, 2020.

https://robotsmali.org/wp-content/uploads/2020/05/DiarraLeventhal_ICLR-PML4DC_47.pdf

Leventhal worked in Silicon Valley as a technologist for over 30 years, the last 15 in developing novel hardware and software for AI and AI-related applications.

P. Dlugosch, D. Brown, P. Glendenning, M. Leventhal and H. Noyes, "An Efficient and Scalable Semiconductor Architecture for Parallel Automata Processing," in IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 12, pp. 3088-3098, Dec. 2014, doi: 10.1109/TPDS.2014.8.

<https://doi.org/10.1109/TPDS.2014.8>

USTTB, University of Sciences, Technical Methods, and Technology of Bamako

Amadou Koné, PhD, is a Researcher and Instructor in Molecular Biology at Mali's national research university, USTTB. His centers of expertise include molecular biology, cellular biology,

immunology, bioinformatics, cytometry, electron and photon microscopy, enzymatic detection by fluorescence, applied biochemistry, autoclave operation, research project planning and management, evaluation of teaching, P3 processes and certification. As a member of Mali's only P3 laboratory he is directly engaged COVID testing and is directly engaged in Mali's scientific response to the pandemic. He is President of Mali's national commission for the annual Festival of Sciences, one of the primary initiatives of the Malian government to promote the nation's scientific culture and to showcase its research. He has published extensively on HIV and other infectious diseases.

Diarra B, Safronetz D, Sarro YD, Kone A, Sanogo M, Tounkara S, Togo AC, Daou F, Maiga AI, Dao S, Rosenke K, Falzarano D, Doumbia S, Zoon KC, Polis M, Siddiqui S, Sow S, Schwan TG, Feldmann H, Diallo S, Koita OA. J Infect Dis. 2016 Oct 15;214(suppl 3):S164-S168. Laboratory Response to 2014 Ebola Virus Outbreak in Mali.

<https://pubmed.ncbi.nlm.nih.gov/27707892/>

Diarra B, Goita D, Tounkara S, Sanogo M, Baya B, Togo AC, Maiga M, Sarro YS, Kone A, Kone B, M'Baye O, Coulibaly N, Kassambara H, Cisse A, Belson M, Polis MA, Otu J, Gehre F, Antonio M, Dao S, Siddiqui S, Murphy RL, de Jong BC, Diallo S. BMC Infect Dis. 2016 Nov 28;16(1):714. Tuberculosis drug resistance in Bamako, Mali, from 2006 to 2014.

<https://pubmed.ncbi.nlm.nih.gov/27894266/>

AMALAN, Malian Academy of Languages

Seydou Traoré is a Researcher and Interpreter at the Academy for Malian Languages, the governmental body responsible for the implementation of Mali's national language policy. His experience has included projects related to the development of national language policy and for the promotion and preservation of Mali's national languages including the creation of an alphabet for Hasanya and participation in the founding of the Institute of Higher Studies and Islamic Research Institute – Ahmed Baba of Timbuktu (IHERI-ABT) and work on its Timbuktu Manuscript Project.

Orange Silicon Valley, California, United States

Sarah Luger is an Artificial Intelligence, Machine Learning, and Natural Language Processing Research Scientist at Orange Silicon Valley, a wholly owned subsidiary of Orange. Dr. Sarah Luger received her Masters and PhD in Artificial Intelligence and Natural Language Processing from the University of Edinburgh. She worked at IBM's T.J. Watson Research Labs on the Watson Jeopardy Challenge and since her return to the Bay Area has built NLP teams at numerous start-ups specializing in language technology and human-in-the-loop AI. She runs the AAI Rigorous Evaluation of AI Systems Workshop held at both the Human Computation (HCOMP) conference and AAI, holds patents in this domain, and most recently published research supporting Machine Translation for African languages. At Orange Silicon Valley she is a member of the Big Data and AI experts group and leads the Responsible AI project.

Extras:

that are related to encouraging public health behaviours to lessen the spread of disease, specifically COVID-19, in Mali, and with special consideration to the needs of the Malian population, particularly women, that may be exacerbating by COVID lockdowns. We will produce public health knowledge about how to acquire texts that are most effective as training examples for machine learning for machine translation in the area of public health, particularly in