BIG-C: a Multimodal Multi-Purpose Dataset for Bemba

Claytone Sikasote[1], **Eunice Mukonde**[2], **Md Mahfuz Ibn Alam**[3], **Antonios Anastasopoulos**[3]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department of Literature and Languages, University of Zambia, Zambia
[3]Department of Computer Science, George Mason University, USA
`claytone.sikasote@cs.unza.zm`, `antonis@gmu.edu`

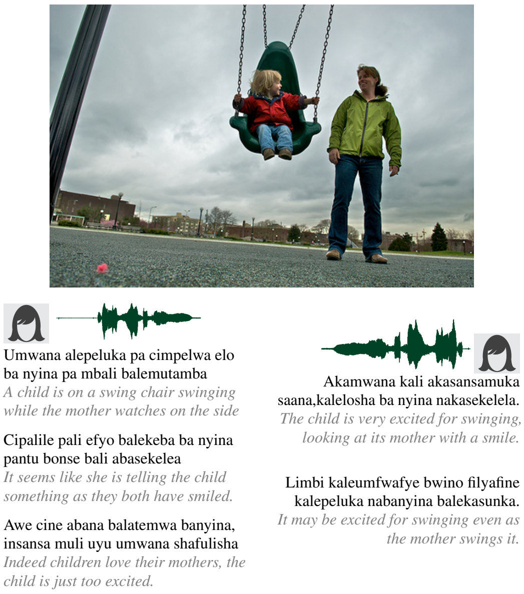
Abstract

We present BIG-C (Bemba Image Grounded Conversations), a large multimodal dataset for Bemba. While Bemba is the most populous language of Zambia, it exhibits a dearth of resources which render the development of language technologies or language processing research almost impossible. The dataset is comprised of multi-turn dialogues between Bemba speakers based on images, transcribed and translated into English. There are more than 92,000 utterances/sentences, amounting to more than 180 hours of audio data with corresponding transcriptions and English translations. We also provide baselines on speech recognition (ASR), machine translation (MT) and speech translation (ST) tasks, and sketch out other potential future multimodal uses of our dataset. We hope that by making the dataset available to the research community,[1] this work will foster research and encourage collaboration across the language, speech, and vision communities especially for languages outside the "traditionally" used high-resourced ones.

1 Introduction

The Bemba language, spoken by over 10 million people in Zambia and other parts of Africa, is a rich and vibrant language with a unique cultural heritage. However, despite its significance, Bemba is a dramatically under-resourced language, lacking in high-quality language data and resources for natural language processing (NLP) experiments and for the development of language technologies. With this work, we address this issue by creating a new multimodal dataset for Bemba. Our goal is to improve the accuracy and effectiveness of NLP systems for speakers of Bemba and support research in this under-served language.

While most datasets are constructed with a specific task in mind and tailored to its characteristics, we aim to provide a path towards building *multi-purpose* datasets. Under a limited budget, we hypothesize that the ideal scenario is to create datasets that can be useful for developing multiple language technologies for both practical applications and also facilitate cutting-edge NLP research on many dimensions. Our hope is that such datasets will aid in bridging the ever-widening language divide both in terms of data availability (Joshi et al., 2020) and NLP research (Blasi et al., 2022), and make language technologies more accessible for speakers of Bemba.



Umwana alepeluka pa cimpelwa elo ba nyina pa mbali balemutamba
A child is on a swing chair swinging while the mother watches on the side

Akamwana kali akasansamuka saana,kalelosha ba nyina nakasekelela.
The child is very excited for swinging, looking at its mother with a smile.

Cipalile pali efyo balekeba ba nyina pantu bonse bali abasekelea
It seems like she is telling the child something as they both have smiled.

Limbi kaleumfwafye bwino filyafine kalepeluka nabanyina balekasunka.
It may be excited for swinging even as the mother swings it.

Awe cine abana balatemwa banyina, insansa muli uyu umwana shafulisha
Indeed children love their mothers, the child is just too excited.

Figure 1: Example of the data included in BIG-C. The grounding image (top) and the ensuing Bemba dialogue transcribed and translated in English.

In this work, we present our methodology and results of creating a new multimodal dataset for Bemba, and demonstrate the potential of this dataset to develop NLP systems and support NLP research. Our dataset will fill multiple roles: enable development of fundamental tools such as speech recognition, speech and text translation systems for Bemba; serve as a benchmark for academic and in-

[1] All data and code are publicly available: `https://github.com/csikasote/bigc`.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2062–2078
July 9-14, 2023 ©2023 Association for Computational Linguistics

Dataset	Images (#unique)	Text (turns)	Audio (hours)	Languages(s)	Parallel
Task: Image Captioning					
MSCOCO (Lin et al., 2015)	330K	1.5M	-	Eng	NA
Flickr8K Audio (Harwath and Glass, 2016)	8K	40K	65	Eng	NA
Flickr30K (Plummer et al., 2015)	30K	158K	-	Eng	NA
Pascal Sentences (Funaki and Nakayama, 2015)	1K	10K	-	Eng, Jap	Partial
IAPR TC-12 (Grubinger et al., 2006)	1K	10K	-	Eng, Deu, Spa	No
Multi30K (Elliott et al., 2016, 2017; Barrault et al., 2018)	30K	155K	-	Eng, Deu, Fra, Ces	Yes
WIT (Srinivasan et al., 2021)	11.5M	37.6M	-	108 langs	Partial
HaVG (Abdulmumin et al., 2022)	30K	30K	-	Eng, Hau	Yes
BAN-Cap (Khan et al., 2022)	8K	40K	-	Eng, Ben	Yes
Bloom Library (Leong et al., 2022)	90K	110K	428	363 langs	NA
Task: Dialogues over Images					
IGC (Mostafazadeh et al., 2017)	4.2K	25K	-	Eng	NA
Image-Chat (Shuster et al., 2020)	202K	202k	–	Eng	NA
BIG-C	16K	90K	185	Bem, Eng	Yes

Table 1: BIG-C and related datasets. BIG-C is the only *multi-purpose* dataset in an under-served language.

dustry research even as NLP for low-resource and under-represented African languages gets developed; facilitate research in language grounding and multimodal model development, or building context-based dialogue agents, among other possible use cases. To our knowledge this is the first such dataset of its kind for any Zambian and possibly African language. We hope that it will provide an example of how to create a *multi-purpose* dataset in an under-served language to facilitate its coverage by multiple technologies.

The rest of the paper is structured as follows: in Section 2, we briefly introduce the Bemba language discussing any currently available resources. In Section 3, we summarise work related to multimodal tasks and existing datasets. In Section 4, we provide a description of the BIG-C dataset and the methodology used, and in Section 5, we provide baseline experiments for some NLP tasks.

2 The Bemba Language

Bemba, also known as *IciBemba* or *Cibemba*, is a Bantu language native to Luapula, Muchinga and Northern provinces of Zambia. It is also spoken in other urban parts of the country like Copperbelt, Central and Lusaka provinces. It is estimated that Bemba is spoken by over 30% of the population of Zambia as either the first or second language, making it the language with the most speakers in the country (Kapambwe, 2018). A map of Bemba usage in Zambia is provided in Appendix Figure 3.

The Bemba language has a number of dialects and the main varieties are: Standard Bemba also Central Bemba, Aushi, Bisa, Chishinga, Lamba, Lala, Luunda, Ngumbo, Swaka, Tabwa and Unga. These dialects show minor differences in phonology, morphology and vocabulary(Spitulnik and Kashoki, 2001; Spitulnik and Kashoki., 2014). In this work, we focus on the Standard Bemba dialect, i.e., the one spoken in urban centers around the country.

Datasets for Bemba For ASR, to the best of our knowledge, there is only a single dataset publicly available for Bemba, BembaSpeech (Sikasote and Anastasopoulos, 2022). It contains 24 hours of read-styled speech data recorded from text mainly sourced from various source but mainly literature books. The low resource nature of the BembaSpeech (Sikasote and Anastasopoulos, 2022) dataset makes it difficult to build usable ASR system for Bemba. For machine translation (text-to-text), there is not a single dedicated dataset for Bemba. However, there exist some parallel text-to-text data in multilingual datasets such as JW300 (Željko Agic and Vulic, 2020) and in evaluation benchmarks such as NTREX-128 (Federmann et al., 2022) and FLORES-200 (NLLB Team et al., 2022). The text in the JW300 (Željko Agic and Vulic, 2020) is mostly religious as it is derived from the Bible text. For speech translation (speech-to-text; ST), to our knowledge, no prior work or Bemba dataset exists. This essentially renders it impossible to build a ST system where Bemba is a source or target language. The same is true for multimodal and dialogue datasets: there is no multimodal or dialogue-related dataset for any Zambian language that would enable development of multimodal systems. Our work aims to fill these gaps.

3 Related Work

In the recent years, NLP, speech processing (SP) and computer vision (CV) fields have rapidly advanced, with computational models' performance achieving new heights on a wide range of downstream tasks. This, to some degree, can be attributed to factors such as the emergence of pre-trained models leveraging self-supervised learning, the availability of large-scale datasets, and increased large-scale computational infrastructure (Hirschberg and Manning, 2015). In NLP, language models like BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), GPT3 (Brown et al., 2020) and XLM-R (Conneau et al., 2020), pre-trained on massive text datasets such as C4 (Raffel et al., 2020), mC4 (Xue et al., 2021) and BooksCorpus (Zhu et al., 2015) among others, have lead to significant performance improvements on several language understanding and generation downstream tasks. Likewise, for speech processing, the unsupervised pretraining of models like wav2vec2.0 (Baevski et al., 2020) or XLS-R (Babu et al., 2021) – having been pretrained on publicly available speech datasets such as VoxPopuli (Wang et al., 2021), MLS (Pratap et al., 2020), Commonvoice (Ardila et al., 2020), BABEL (Punnakkal et al., 2021) among others, have led to advances on speech downstream tasks like ASR (Babu et al., 2021) and ST. In computer vision, deep learning models like DeepCNN (Simonyan and Zisserman, 2015; He et al., 2016) have become the *de facto* solution for standard vision problems like object recognition (He et al., 2016), image classification (Krizhevsky et al., 2017), or semantic segmentation (Shelhamer et al., 2017).

Since these neural models are conceptually (and architecturally) quite similar they have also enabled the integration of multiple modalities, with models such as ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020), Unicoder-VL (Huang et al., 2019) able to jointly model the relationship between text and image modalities resulting into breakthroughs across a myriad of tasks such as image-text retrieval/search (Frome et al., 2013; Huang et al., 2020), image or video captioning (Biten et al., 2019), and vision-question answering (VQA; Agrawal et al., 2017; Nam et al., 2017). A crucial necessary component for all of the above, of course, is the availability of relevant datasets. Below we discuss works that go beyond the collection of raw datasets that are used for self-supervised learning.

Dialogue In the recent past, a lot of work has been focused on dialogue datasets. On one hand there exist goal-oriented dialogue datasets, such as the case of the Ubuntu dialogue corpus (Lowe et al., 2015), the largest corpus of dialogues (almost 1 million mainly 3-turn dialogues in English) for the specific topic of troubleshooting Ubuntu problems. On the other hand, open ended conversations, such as those on the CALLHOME/CALLFRIEND (Canavan et al., 1997) or Fisher corpora (Cieri et al., 2004), often leads to uninteresting conversations. Grounding the dialogue to event-centric images and potentially a specific scenario constrains the topic of conversation to event-rich and contentful utterances.

Multimodality Multimodal works combining visual and language information typically focus on image captioning and visual question answering (Antol et al., 2015). For example, the IAPR TC-12 dataset (Grubinger et al., 2006) provides images with titles and descriptions (mostly in English, German, and Spanish), as do commonly used datasets like MSCOCO (Lin et al., 2015) and Flickr30K (Plummer et al., 2015). Flickr8K Audio (Harwath and Glass, 2016) extended a subset of the Flickr images with audio, by crowdsourcing readings of the English captions, while Multi30K (Elliott et al., 2016) further extended Flickr30K with German translations and annotations. Wikipedia-based Image Text (WIT) Dataset (Srinivasan et al., 2021) provided large multilingual coverage (108 languages) based on 11.5M images and captions from Wikipedia. More recent, Hausa Visual Genome (HaVG; Abdulmumin et al., 2022) provided over 30K parallel descriptions in English and Hausa of images from the Hindi Visual Genome (HVG; Parida et al., 2019). The dataset was created by automatically translating the English descriptions of the images in the HVG to Hausa using Google Translate[2] and post-edited by crowd-sourced Hausa volunteers. Similarly, BAN-Cap (Khan et al., 2022) provides over 40K human-annotated parallel English-Bangla image description pairs based on 8,091 images from Flickr8K (Harwath and Glass, 2016). Lastly, the Bloom Library (Leong et al., 2022) provides a set of multilingual datasets for language modeling, image captioning and visual-story telling tasks containing more than 110K image captions for over 90K images in 351 languages. It also provides a

[2] https://translate.google.com/

speech dataset with 428 hours of speech data for speech synthesis/recognition tasks covering 56 languages.

Beyond captioning tasks, the dialog component was first explored by Das et al. (2017), who extended the VQA scenario collecting sequential questions grounded on images. Mostafazadeh et al. (2017) went beyond goal-oriented dialogue to collect image-grounded conversations (contrasting this to open-ended dialogue research). More recently, the Image-Chat dataset (Shuster et al., 2020) collected open-ended conversations grounded in images with a focus on engagement, by assigning desired style traits to the speaker.

Discussion There are notable limitations with most publicly available multimodal datasets. To make comparisons easy, we outline most relevant works in Table 1. While the list shown there is non-exhaustive, these limitations can be grouped in terms of language coverage, modality composition, tasks supported i.e., single-purpose or multi-purpose tasks. To give more context to this categorization:

- In terms of languages, they cover only a handful of high-resourced languages like English.
- In terms of modality composition, the majority only contain image and text modalities, ignoring the audio component.
- With regards to tasks, the majority are meant for a single-purpose task such as image captioning.[3]

In contrast, our work presents a *multimodal* but also *multi-purpose* dataset for Bemba. Our aim is for BIG-C to be the first-of-its-kind dataset for an under-served language that can *simultaneously* serve as:

- a monolingual dataset for Bemba e.g., to be used for training language models on this under-served language;
- a parallel dataset to allow for building and evaluating machine translation solutions;
- an image captioning dataset with image descriptions in Bemba;
- an image-grounded dialogue dataset;
- a benchmark for any combination between the above modalities e.g., one could use our dataset to evaluate image-grounded dialogue translation systems.

Description	Count
Data	
# unique images	16,229
# hours transcribed and translated	187
# complete dialogues	16,697
# "incomplete" dialogues	2,314
# sentences/complete dialogue	5
# spoken utterances	92,117
# English translations	92,117
# Bemba tokens	870K
# English tokens	1.1M
Metadata	
# speakers	86
# transcribers	93
# translators	114
# validators	15

Table 2: BIG-C: Basic Dataset Statistics.

We achieve this through careful instructions and data collection practices, outlined in Section §4.

4 Dataset Description

Description The dataset consists of a parallel corpus of speech and transcriptions of image-grounded dialogues between Bemba speakers and their corresponding English translations. It contains 92,117 spoken utterances (complete and incomplete dialogues), amounting to 187 hours of speech data grounded on 16,229 unique images. There are 16,697 complete 5-turn unique dialogues grounded on 14,551 unique images. Of the total 16,697 complete dialogues, 2,146 are unique dialogues grounded on duplicated images, each recorded by unique pairs of speakers. A second set of dialogues is comprised of 2,314 incomplete dialogues missing one or more utterances as a result of the preprocessing step that involved removing all audio files that are silent and corrupted. The sum of utterances that make up the incomplete dialogues is 8,632 of the total 92,117 utterances. All audio files are encoded in Waveform Audio File format (WAVE) with a single track (mono) and sample rate of 16kHz. In Table 2, we provide basic dataset statistics.

Source of images We randomly selected images from the Flickr30K (Plummer et al., 2015) dataset, a publicly available multimodal dataset for vision and language that has become a standard benchmark for sentence-based image descriptions.

[3] An exception to this is the Bloom Library (Leong et al., 2022). But note that it lacks representation of any Zambian language among the covered languages.

Speakers To record conversations, we recruited 86 speakers of the Bemba language; 60% male and 40% female, based on their competency to speak, read and write the language. Based on the metadata information supplied by participants, we summarise the characteristics of our speakers as follows:

- **Age:** the majority of the speakers (98%) were youth whose age falls between 20 and 35 years old with the 2% being over 35 years old.
- **Education:** all speakers had some form of secondary education; 90% of the participant were either pursuing or recently graduated with a college/university degree; and the rest 8% had only completed high school.
- **Language(s):** all speakers were bilingual; with 90% indicating Bemba as their first language and Nyanja as the majority non-English second language.
- **Regions:** in terms of regional representations, over 90% of the speakers were drawn from Lusaka, Central, and Copperbelt regions; with small representations from Muchinga and Northen provinces. This in effect indicates that the dataset is composed of the current 'urban' Bemba variety.
- **Racial diversity:** the composition of our participants lacks racial diversity, as all speakers are identified as black.

Recording The speakers were randomly paired with gender-balancing in mind. Each pair was allocated 250 images to create 5 sentence-turn conversation per image for each recording session. There was no restriction to what each pair would converse about on an image. The participants were encouraged to be creative. However, the conversation starter (speaker 1) was instructed to first describe the image, so as to give context to the conversation (and essentially provide data for the image captioning component of our dataset). We provide the sample instructions that were given to the annotators in Appendix A. All recordings were conducted in minimally controlled conditions. The pairs recorded as per their comfort, we therefore expect that some spoken utterances have background noise. All participants used the LIG-AIKUMA (Gauthier et al., 2016) mobile application, using the 'elicitation by image' mode to record spoken utterances.

Transcribers To transcribe the audio data generated from the image-grounded conversations, we recruited 93 participants, who in their majority were students of the University of Zambia. All were competent Bemba speakers and writers. As shown in Table 2, 92,117 spoken utterances were transcribed representing 187 hours of Bemba speech data.

Translators To translate a subset of the transcriptions to English, we recruited 115 participants with experience in translating Bemba text to English or vice versa. Public education in Zambia is conducted in English, hence we are confident in a minimum translation quality.

Splitting We have split the dataset into training, validation and testing sets following the original splits in the Flickr30K (Plummer et al., 2015) dataset according to the images. See Table 3 for more details.

Data quality Several measures were set up during the data collection process to ensure quality submissions from project participants; speakers, transcribers and translators. First, at recruitment stage for audio recording, we considered only competent Bemba speakers with ability to speak, read and write in Bemba. All the speakers underwent a training exercise to make sure they understood and followed instructions of how to go about the task of creating and recording multi-turn conversations using the Lig-Aikuma (Gauthier et al., 2016) mobile application. For the transcriptions, we retained good number of the speakers - over 50% to also participate in transcribing the audio files at transcribing stage. In addition, we recruited validators, who together with the authors of this study checked and verified manually every submission made by the participants at every stage of the process. All audio files that were deemed to be of low quality i.e., silent, corrupted and inaudible due to background noise, were removed as part of data pre-processing at the quality assurance and validation stage.

Last, during the translation stage, besides the ability to speak, read and write, we recruited participant who had experience with translating Bemba text to English as translators. Most of the participants had prior experience as professional or volunteer translators.

Availability The dataset is made available to the research community licensed under the Creative Commons BY-NC-ND 4.0 license and can be ac-

				No. of speaker voices		
Split	**Images**	**utterances**	**hours**	**Male**	**Female**	**Unspecified**
Train	14,599	82,375	167	43,959	38,338	78
Valid	492	2,782	5	1,491	1,289	2
Test	501	2,779	5	1,457	1,318	4
Held	637	4,181	8	2,105	2,072	4
Total	16,229	92,117	185	49,012	43,017	88

Table 3: Summary details of the splits of the dataset.

cessed at our Github repository.[4] We do plan to keep a small held-out portion unpublished, to be used in future shared tasks or as part of leaderboards that require *hidden* test sets to ensure a fair measure of task progress.

5 Baseline Experiments

In this section, we detail some baseline experiments carried out to demonstrate the potential of the dataset. We provide unimodal baselines using the train-validation-test splits in Table 3 on the following tasks: ASR for Bemba, MT and ST of Bemba utterances to English text.

Data preprocessing For ASR and ST, similar to Wang et al. (2020a), all text i.e., transcriptions and translations, we lower the cases and remove punctuation except for apostrophes, and build 1K unigram character vocabularies with 100% coverage of all the characters using SentencePiece (Kudo and Richardson, 2018) without pre-tokenization. We extract 80-dimensional log-mel scale filterbank features from Bemba utterances using a 25ms window size and 10ms window shift using torchaudio.[5] The features are normalized to 0 mean and 1.0 standard deviation. All models are trained without an auxillary language model.

Model Architecture We use the small Transformer (Vaswani et al., 2017) base architecture with 71 M parameters, `s2t_transformer_s`, having 12-layers encoder, 6-layers decoder, and hidden dimension D=256 to train end-to-end (E2E) ASR and ST models using FAIRSEQ S2T Toolkit (Ott et al., 2019; Wang et al., 2020b). Models are trained on a single NVIDIA Tesla P100 GPU using the Google Colab+ platform.

5.1 Automatic Speech Recognition

For the ASR baseline model for Bemba, we trained the model for 500 epochs using the Adam optimiser (Kingma and Ba, 2015) with 10K warm up steps. The model is optimised to minimise the `label_smooth_cross_entropy` criterion function using the learning rate coefficient of `2e-3`. For decoding, we use the beam search algorithm with a beam size of 5. We use the average of the last 5 checkpoints for evaluation. In Table 4, we report the model performance on the Test set using word error rate (WER) metric.

5.2 Speech Translation

For speech to text translation of Bemba spoken utterances to English text, we use the same model architecture as ASR. The model is trained with same configuration as the ASR model except we use the learning rate coefficient of `3e-4`. Similarly, we use the beam search algorithm with beam size of 5 for decoding. We use the best checkpoint to evaluate the model on the test set. We report the detokenised case-sensitive BLEU (Papineni et al., 2002) using sacreBLEU (Post, 2018) in Table 4.

Evaluation We use beam search with a beam size of 5 for decoding. We use the average of the last 5 checkpoints to evaluate both ASR and the best checkpoint saved for ST model. We report the results in Table 4. For ST, we report detokenised case-sensitive BLEU (Papineni et al., 2002) using sacreBLEU (Post, 2018) and word error rate (WER) for ASR.

Results discussion For both ASR and ST, we consider the results obtained decent for the size of our dataset and the basic training configurations of our baseline models, which are without auxillary models, and mostly relied on default settings in the FAIRSEQ S2T implementation. We believe the results can be improved upon, and we leave

[4] `https://github.com/csikasote/bigc`

[5] https://github.com/pytorchaudio

Task	Metric: Value
Speech Recognition	WER (↓): 32.7
Speech Translation	BLEU (↑): 17.9

Table 4: Baseline results (test set) on speech-based tasks. For ST, we report detokenised case-sensitive BLEU (Papineni et al., 2002) using sacreBLEU (Post, 2018) and word error rate (WER) for ASR.

the full exploration of the best configurations to future work. We encourage the community to improve upon these baselines, for instance, by exploring cross-lingual transfer learning by leveraging large scale multilingual pretrained models like XLS-R (Babu et al., 2021) and Whisper (Radford et al., 2022).

5.3 Machine (Text) Translation

For Machine Translation we rely on the results of the WMT Shared Task on Large Scale Machine Translation Evaluation for African Languages (Adelani et al., 2022). In particular, we use the same system and approach as Alam and Anastasopoulos (2022), which ranked third in the Shared Task.[6] These models are based on the DeltaLM (Ma et al., 2021) pre-trained model, which is the adapted through fine-tuning on 24 African languages (note that Bemba is not included), as well as English and French. The adaptation happens using adapter units (Pfeiffer et al., 2020) organized in a hierarchy following language typology (Faisal and Anastasopoulos, 2022) so that similar languages share similar "family" adapters. We also compare against a baseline that finetunes the whole DeltaLM model on our training set.

Here, we only use our training data to fine-tune the adapter units for Bemba, and evaluate on both our test set as well as on the publicly available FLORES-200 devtest (NLLB Team et al., 2022). The results are presented in Table 5, where we report sentencepiece-BLEU (NLLB Team et al., 2022) with the FLORES-200 tokenizer. In general, translating into English seems to perform well, especially for the phylogeny-based model.

The difference between the performance in the two test sets can be explained by the difference of domains. All BIG-C training data are from dialogues, while the FLORES-200 evaluation dataset is comprised of translated Wikipedia articles. Of course, larger and more diverse data collection in the future should help mitigate these issues and allow us to build general translation systems capable of handling various domains adequately.

5.4 Other Tasks

The authors of this study unfortunately lack the financial and compute resources, as well as required expertise, to provide baseline results for additional multimodal tasks. Nevertheless, we devote this subsection to outlining some other potential downstream uses of BIG-C.

- **Image Captioning** The dataset could be used directly for image captioning in Bemba (or English), by pairing the images with the first utterance of the conversation, which will largely function as a caption by design.
- **Multimodal Language Modeling** Similarly, the corpus could be used for language and vision pre-training, and particularly to study multilingual approaches (in a field that has largely focused solely on English).
- **Multimodal Dialogue Modeling** Similar to other image-grounded tasks (see §3), one could use to BIG-C to study dialogue, with a focus on open-ended but still grounded conversation. One could also use our dialogues as (pre-)training data for chatbots in Bemba, which could then potentially be adapted to handle specific goals or domains with fewer in-domain data.
- **Multimodal Translation** In the experiments above we did not take advantage of the image when translating. One could explore whether multimodal machine translation approaches (Barrault et al., 2018, ; *inter alia*) could improve downstream performance in these resource-scarce settings.
- **Cross-Cultural NLP** A major limitation of our dataset (also discussed in the relevant Limitations section) is that most of the images that we use are not particularly relevant to the Zambian or sub-Saharan African context. We plan to mitigate this issue by collecting an addendum to BIG-C with images crowd-sourced *in Zambia*.
  Nevertheless, this limitation is simultaneously an opportunity to study cross-cultural understanding as well as the priors/assumptions/biases that speakers with a certain background exhibit. To highlight this potential, we show some additional

[6] We note that this is the best-performing system that is publicly available – to our knowledge, the first two performing systems were industry submissions without publicly released models or code.

	BIG-C		FLORES-200	
Model	**eng→bem**	**bem→eng**	**eng→bem**	**bem→eng**
DeltaLM FT	17.9	27.5	3.5	4.3
Phylogeny FT	16.5	**28.9**	**6.0**	**18.0**

Table 5: Baseline text translation results. The phylogeny-based model benefits from parameter sharing across all the other Bantu languages.

interesting examples from BIG-C in Figure 2. In the top-left example, the first speaker's utterances reveal several assumptions: that the musicians are "Indian" (likely correct, since this image is located in India); that they "are on a roof" (correct); that they "sing religious songs" (unsupported); or that "it's time to congregate and pray" (unsupported). In the example in the top-right, the first speakers assumes the image is "by the riverside", and not e.g., by the seaside or lakeside.[7]

6 Conclusion

In this paper, we presented a multimodal corpus comprised of multi-turn dialogues between speakers of the Zambian language, Bemba, grounded on images, transcribed and translated into English. It contains over 92,000 utterances/sentences, 180 hours of speech grounded over 16,000 images. The dataset aims to fill multiple roles: enable development of fundamental tools like speech recognition, machine translation and speech-to-text translation systems between Bemba and English; serve as a benchmark for academic and industry research; and to facilitate research in language grounding and multimodal model development towards building context-based dialogue agents, among other potential use cases. We have also provided baseline for ASR, MT and ST task.

In future work, we plan to conduct multimodal baseline experiments, as well as attempt to mitigate the image diversity limitation by collecting an addendum to BiG-C using images taken locally in Zambia. In addition, we plan to further expand to other Zambian languages such as Tonga, Tumbuka, Chewa, or Lozi, by translating the existing dataset (creating an n-way parallel corpus for Zambian languages) and by direct data collection. Further down the roan we plan to study the dialectal varieties of Bemba and the other languages, by collecting contrastive datasets from different regions of the country.

Limitations

We observe the following limitations with the dataset:

- **Language Diversity:** In terms of number of languages, the presented dataset only covers two languages; Bemba and English.
- **Image Diversity** All the images used in this dataset were obtained from Flickr30K image dataset. Therefore, in terms image composition, our dataset is limited to the image diversity in the Flickr30K dataset. It mostly lacks images that could be considered as "culturally relevant" ones for the Zambian or generally sub-Saharan African context. We plan to mitigate this in future work.

Ethics Statement

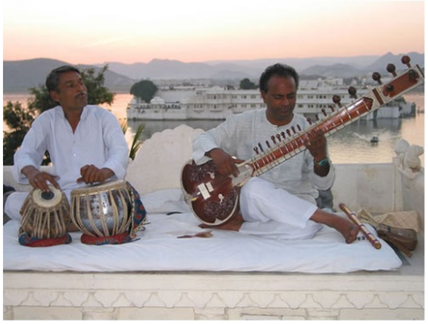
We make the following declarations for the ethics statement:

- **Research:** This work was carried out mostly in Zambia, and most authors are native speakers of Bemba who also worked as validators for the data collection process.
- **Participants:** All project participants; transcribers, translators and speakers/recorders were informed about the goals of the project and they signed consent forms to participate. All participants were monetarily compensated at around $20/h for all their work.
- **Personal Identifiable Information:** All information that can potentially be regarded as PII such as names of participants, IDs have been removed for anonymity and will not be released with the dataset.
- **Copyright:** There is no potential copyright matters associated with the data contained in this dataset. We are publicly releasing the dataset under the Creative Commons BY-NC-ND 4.0 license.

Acknowledgements

We would like to thank all the participants that were involved at different stages of the dataset creation

[7] Note that Zambia is a land-locked country.



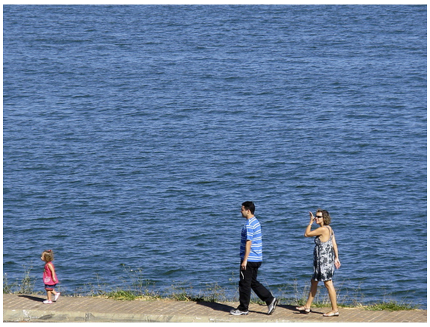
Bakemba babili bamwenye bali pa mutenge mupepi ne cishiba. Balelisha banjo, ingoma na masese
Two Indian musicians are on a roof top near a water body. They are playing a banjo, some drums and some beads that rattle.

Aba bakemba babili ba mwenye nibashitata abafwele ifyakufwala ifya buta napantu bekele balelisisha nipa nsalu yabuta.
These two Indian musicians are elderly men wearing white clothes and are seated on a white cloth

Nalimo ukwimba kwabo kwa kupepa. icimbo ca mapepo ntile
They seem to be singing religious songs. I am sure they are singing religious songs!

Emukwai. Ukwimba kwabo kulemoneka nakalimo tekwimbafye iyo, kwati nintambi. Emo basangila umutende nobutusho ngabaleimba kumipashi yabo.
That's right. Their singing doesn't seem to be more singing, it seems more like a religious practice. I am sure they find peace and rest as they sing to their gods

kwena pantu bali nipa muulu wa cikulwa nalimo beleishibisha abanabo ukutl ni nshita yakulongana kukupepa.
Surely, their being on top of that building seems to be a signal to the rest of their community that it's time to congregate and pray



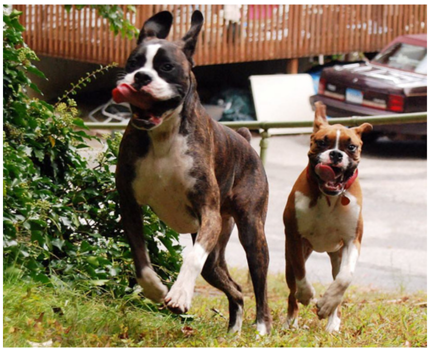
Ba nyina , ba wishi nomwana baleenda mumbali ya cimana nabatangisha nomwana.
The father, wife and child walking in front of them by the riverside.

Icimana cilemoneka icikulu saana bushe aba bafyashi tabalemona ati umwana kuti aponenamo fyaleta ubwafya?
This river is so huge and deep, are they not afraid of the child in front to slip off and fall?

Awe nifyofine, cikulu icimana ici icakweba ati ngaponenamo kuti bafilwa napakutampila ukumufwaya.
It is so big indeed, such that if the child fell in they would struggle so much.

Caliba icikankala saana abafyashi ukulolekesha pabana,pantu ngatabalelolekesha pa bana ngabali kuncende ngeshi kuti caleta ubwafya ubukalamba saana.
It is quite important for parents to ensure their children's safety, especially when outing to suchlike places because it would be a fatal encounter here.

Ee nifyo elo cipalile kwati umwana nasansamuka pakumutwala ku menshi, alemonekafye uwansansa
That is so true, and the very child is very excited to be brought to this place.



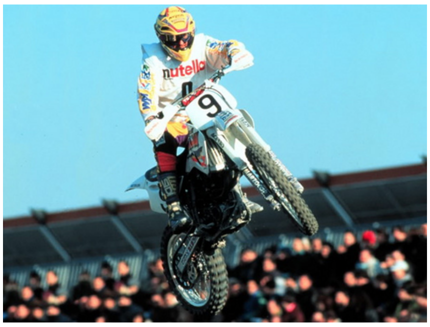
Imbwa shibili shileingila paka panga shilebutuka.
Two dogs are headed to a thicket.

boi imbwa ishi shilemoneka ishikali,nashifumya nendimi panse kwati shamona akakulya akanona.
Dear these are dogs that seem to be fierce, just their race is hunty, as if after some fatty food.

Shifwile shileyangalafye. Imbwa shalitemwa ukubutauka, kuti wasanga limbi pali abashipepeke.
These dogs must be just playing, as dogs naturally love running around.

boi ishi nimbwa shakweba ati ngawashimonafye ufwile watampako nolubilo, utunwa natukulisha.
No way my friend, these are dogs you run away from the moment you see them. Their mouths are too big.

ubwafya walifulisha umwenso, imbwa shalitemwa ukwangala nabantu,ngawabutuka ninshi wailetelelafye.
The problem is that you are full of cynophobia, dogs are friendly to humans and enjoy man's company.



Umulumendo aleyangala na Honda mumulu, muchibansa muli na ibumba lyabantu abaletamba uyu mulumendo.
A gentleman is on his motorbike spinning with a crowd of people around watching.

Abantu bonse mwibumba baletota, cilemoneka kwati nabasekalamo sana pafyo uyu muntu alepilibausha icela cakwe.
Everyone is excited and happy to see how he is drifting his machine.

Boi amangalo ya ifi yalaleta abantu abengi chapamo, balomfwa bwino ukutamba umuntu alecita ifintu ifyo abengi teti bacite.
My dear this event is such a big thing, many people come by to watch and enjoy how that one can do what exceptionally.

Nomba nangu bengomfwa bwino, umunabo ngaicena akacula eka nabalupwa bakwe.
However the crowd when you are hurt you are on your own with relatives only.

Boi umuntu pakucita ifi ninshi pali cimo, ubu ubwangalo bukulu saana,limbi balapela indalama ishingi saana kuli uyo uwacimfya.
Dear for one to participate in anything there must be a reason, this sport is well sponsored and the winner is awarded unreservedly.

Figure 2: Examples of the BIG-C dataset. The grounding image (top) and the ensuing Bemba dialog transcribed and translated in English.

process. We would also like to thank Desmond Elliott and Graham Neubig for insightful conversations and constructive feedback at earlier stages of our project. This project would not have been possible without generous funding by the Lacuna-Fund. Antonios Anastasopoulos is also supported by NSF-NEH grant BCS-2109578.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

David Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussÃ, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco GuzmÃ¡n, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022. Findings of the wmt'22 shared task on large-scale machine translation evaluation for african languages. In *Proceedings of the Seventh Conference on Machine Translation*, pages 773–800, Abu Dhabi. Association for Computational Linguistics.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. Vqa: Visual question answering: www.visualqa.org. *International Journal of Computer Vision*, 123.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2022. Language adapters for large-scale mt: The gmu system for the wmt 2022 large-scale machine translation evaluation for african languages shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 1015–1033, Abu Dhabi. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv*, abs/2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. volume 2020-December.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and DImosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. volume 2019-June.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. volume 2020-December.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. Callhome american english speech, ldc97s42.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 585–590, Lisbon, Portugal. Association for Computational Linguistics.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. Lig-aikuma: A mobile app to collect parallel speech for under-resourced language studies.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2.

David Harwath and James Glass. 2016. Deep multimodal semantic embeddings for speech and images.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. volume 2016-December.

Julia Hirschberg and Christopher D. Manning. 2015. Advances in natural language processing.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mazuba Kapambwe. 2018. An Introduction to Zambia's Bemba Tribe.

Mohammad Faiyaz Khan, S.M. Sadiq-Ur-Rahman Shifath, and Md Saiful Islam. 2022. BAN-cap: A multipurpose English-Bangla image descriptions dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6855–6865, Marseille, France. European Language Resources Association.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks.

In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. volume 32.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. arXiv:2106.13736.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hyeonseob Nam, Jung Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. volume 2017-January.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu : a method for automatic evaluation of machine translation. *Computational Linguistics*.

Shantipriya Parida, Ondrej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation. *CoRR*, abs/1907.08948.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Matt Post. 2018. A call for clarity in reporting bleu scores. volume 1.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. 2021. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731.

Alec Radford, Jong Wook Kim, Tao Xu, and Ilya Sutskever Greg Brockman, Christine McLeavey. 2022. Robust speech recognition via large-scale weak supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. 2017. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Debra Spitulnik and Mubanga E Kashoki. 2001. *Facts About the World's Languages: An Encyclopedia of the Worlds's Major Languages, Past and Present*. H.W. Wilson, New York.

Vidali D Spitulnik and Mubanga E Kashoki. 2014. Bemba Morphology.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. volume 2017-December.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

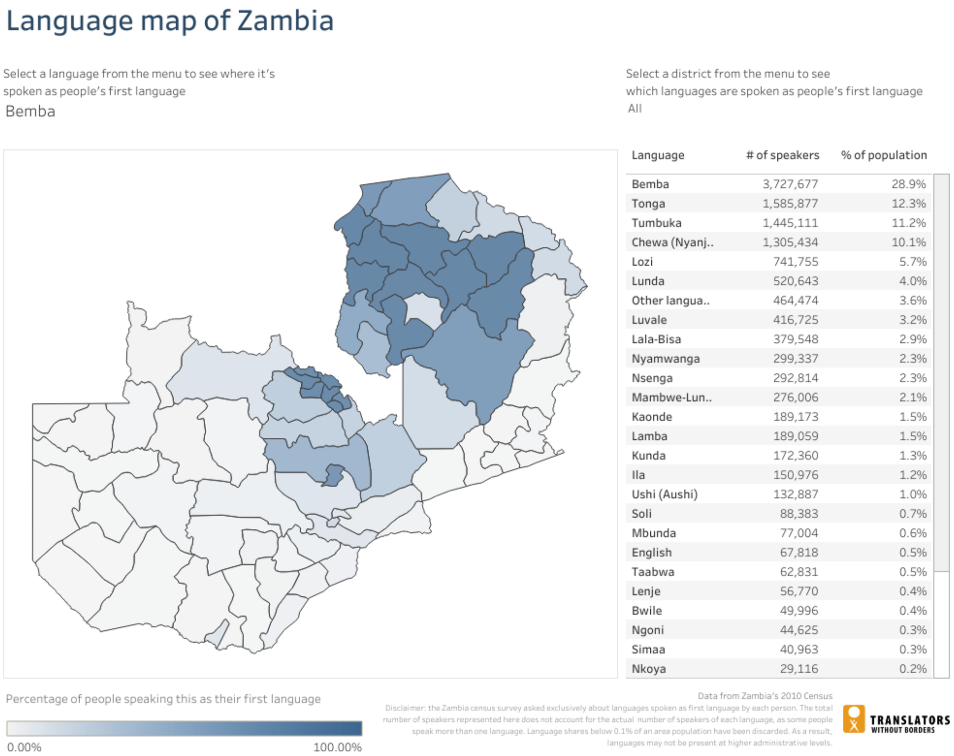
Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020b. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Željko Agic and Ivan Vulic. 2020. Jw300: A wide-coverage parallel corpus for low-resource languages.

A Language Map of Zambia



Language map of Zambia

Select a language from the menu to see where it's spoken as people's first language
Bemba

Select a district from the menu to see which languages are spoken as people's first language
All

Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvale	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	276,006	2.1%
Kaonde	189,173	1.5%
Lamba	189,059	1.5%
Kunda	172,360	1.3%
Ila	150,976	1.2%
Ushi (Aushi)	132,887	1.0%
Soli	88,383	0.7%
Mbunda	77,004	0.6%
English	67,818	0.5%
Taabwa	62,831	0.5%
Lenje	56,770	0.4%
Bwile	49,996	0.4%
Ngoni	44,625	0.3%
Simaa	40,963	0.3%
Nkoya	29,116	0.2%

Percentage of people speaking this as their first language
0.00%
100.00%
TRANSLATORS WITHOUT BORDERS

Data from Zambia's 2010 Census

Disclaimer: the Zambia census survey asked exclusively about languages spoken as first language by each person. The total number of speakers represented here does not account for the actual number of speakers of each language, as some people speak more than one language. Language shares below 0.1% of an area population have been discarded. As a result, languages may not be present at higher administrative levels.

Figure 3: Language Map of Zambia. Created by Translators without Borders. Retrieved from https://translatorswithoutborders.org/languages-of-zambia-interactive-en on January 2022.

A Participant Training Exercise

The following instructional steps depict the participants exercise/tutorial during a training exercise session before actual recording. The instructions were given to a pair of participant. The objective was to create a text conversations for 5 sample images in a specified image folder using Google Sheets. The recording session followed the same process, except with additional instructions involving the use of the LIG-Aikuma (Gauthier et al., 2016) app.

- **STEP 1**: Open the first image in your image folders. If you are P16, for example, Go to `P1_Session_01 > Image7501 > Speaker_01` [If you are `Speaker 1`] or `Speaker_02` [If you are `Speaker 2`]. Open any of the images in the folder.
- **STEP 2**: While you are able to view the image, open the spreadsheet. Now that you have both image and spreadsheet opened.
- **STEP 3**: `Speaker 1` should enter the image number (in this case, 7501) in cell A3.
- **STEP 4**: `Speaker 1` should describe what is in the image by a single sentence in cell B3. The description should be a single sentence giving a clear mental picture of what is in the image.
- **STEP 5** : `Speaker 2` should be able to respond to `Speaker 1` by entering their response in C3. The response can be a question, a statement or an addition to what `Speaker 1` said. As long as it's a sentence in Bemba. Remember this is a conversation and it should be able to naturally flow.
- **STEP 6**: `Speaker 1` should complete cell D3 with a sentence in response to what `Speaker 2` texted in cell C3.
- **STEP 7**: `Speaker 2` should put a response in cell E3 in response to what `Speaker 1` texted in cell D3.
- **STEP 8**: `Speaker 1` closes the conversation with a sentence, however it may be in cell F3.
- **STEP 9**: If you have successfully generated the conversation/dialogue in the spreadsheet for the first image, then go ahead and do so for the next 4 images.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After Section 6

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5

☑ B1. Did you cite the creators of artifacts you used?
5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
1,5,6

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
1,5,6

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
6

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Using default parameters and recipes

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5, No hyperparam search

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
In Bemba

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
4

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
6

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
4,6

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
4